

**New York State Examination in Grade 8
Intermediate-Level Science**

**2010 Field Test Analysis,
Equating Procedure, and Scaling of
Operational Test Forms**

Technical Report



Prepared for the New York State Education Department
by Pearson

August 2011

Copyright

Developed and published under contract with the New York State Education Department by Pearson. Copyright © 2010 by the New York State Education Department.

Secure Materials.

All rights reserved. No part of this document may be reproduced or transmitted by any means. Use of these materials is expressly limited to the New York State Education Department.

Table of Contents

Table of Contents.....	i
List of Tables	ii
Section I: Introduction	1
Purpose	1
Section II: Field Test Analysis.....	1
Data Clean-up	2
Classical Analysis.....	2
<i>Item Difficulty</i>	3
<i>Point-Biserial Correlation</i>	3
<i>Test Reliability</i>	4
<i>Scoring Reliability</i>	5
<i>Inter-rater Agreement</i>	6
<i>Constructed-Response Item Means and Standard Deviations</i>	12
<i>Intra-class Correlation</i>	12
<i>Weighted Kappa</i>	13
Item Response Theory (IRT) Statistics	13
<i>Item Calibration</i>	14
<i>Item Fit Evaluation</i>	14
Differential Item Functioning (DIF) Statistics.....	15
Section III: Equating Procedure	17
Section IV: Scaling of Operational Test Forms	18
References	20
Appendix A: Classical Item Analysis.....	21
Appendix B: Partial Credit Model Item Analysis.....	32
Appendix C: DIF Statistics	43
Appendix D: Operational Test Map.....	49
Appendix E: Scoring Table	54

List of Tables

Table 1.	Need/Resource Capacity Category Definitions	1
Table 2.	Classical Item Analysis	4
Table 3.	Test and Scoring Reliability.....	5
Table 4.	Point Differences Between First and Second Reads	7
Table 5.	First and Second Read Descriptive Statistics and Agreement.....	10
Table 6.	Partial Credit Model Item Analysis	15
Table 7.	Initial Mean Abilities and Equating Constants	18
Table 8.	Classical Item Analysis	22
Table 9.	Partial Credit Model Item Analysis	33
Table 10.	DIF Statistics.....	44
Table 11.	Operational Test Map for May 2010.....	50
Table 12.	Scoring Table for May 2010	55

Section I: Introduction

Purpose

The purpose of this report is to document the psychometric work on the New York State Examination in Grade 8 Intermediate-Level Science for 2010. Specifically, contained within this report are procedures for and results of field test analysis, equating, and scaling of operational test forms. Because of a change in vendor mid-year, the field test equating was conducted by Pearson while the scaling was conducted by the previous vendor. Information on test development can be found in the test design and development report for the New York State Examination in Grade 8 Intermediate-Level Science.

Section II: Field Test Analysis

In May 2010, field testing was conducted for the New York State Examination in Grade 8 Intermediate-Level Science 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 70 th percentile on the index with at least 100 students per square mile or enrollment greater than 2500
Rural	All districts at or above the 70 th percentile with fewer than 50 students per square mile or enrollment of less than 2500
Average N/RC Districts	All districts between the 20 th and 70 th percentiles on the index
Low N/RC Districts	All districts below the 20 th percentile on the index
Charter Schools	Each charter school is a district

The data collected from field testing were scored by the New York State Education Department. Both classical and item response theory analyses were conducted using the data to evaluate the quality of the test items.

Data Clean-up

Field test forms contained multiple-choice and constructed-response item types. Response data were contained in one file that contained 8,745 student records. After the exclusion rules were applied, the resulting field test data file contained 8,717 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 may 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 student 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. 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. For all items except five, item difficulties were below 0.90. With respect to the point-biserial correlations, most of these correlations fell between 0.25 and 0.50.

Table 2. Classical Item Analysis

Form	N-Count	No of Items	Item Difficulty			Point-Biserial		
			<0.50	0.50 to 0.90	>0.90	<0.25	0.25 to 0.50	>0.50
401	965	23	5	18	0	0	17	6
402	975	23	2	18	1	2	13	6
403	941	21	3	15	0	0	12	6
404	976	30	6	24	0	2	24	4
405	988	21	3	16	0	0	14	5
406	967	24	5	18	1	1	21	2
407	966	21	3	17	1	2	17	2
408	971	23	5	17	0	1	16	5
409	968	23	8	13	2	0	20	3

* 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 'M6' are the proportions of students who received scores 0 through 6. '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 0.74 to 0.83 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
401	0.83	0.79
402	0.79	0.89
403	0.80	0.85
404	0.82	n/a
405	0.80	0.85
406	0.79	0.91
407	0.74	0.94
408	0.74	0.91
409	0.79	0.92

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 students' 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 0.79 to 0.94, 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 the differences between the first read and the second read were 0.

Table 4. Point Differences Between First and Second Reads

			Difference (First Read minus Second Read)						
Form	Item	Score Points	-3	-2	-1	0	1	2	3
401	61	1	0.00	0.00	0.11	0.80	0.10	0.00	0.00
401	62	1	0.00	0.00	0.04	0.88	0.08	0.00	0.00
401	63	1	0.00	0.00	0.02	0.98	0.00	0.00	0.00
401	64	1	0.00	0.00	0.01	0.98	0.01	0.00	0.00
401	65	1	0.00	0.00	0.05	0.86	0.09	0.00	0.00
401	66	1	0.00	0.00	0.04	0.90	0.06	0.00	0.00
401	67	1	0.00	0.00	0.06	0.90	0.04	0.00	0.00
401	68	1	0.00	0.00	0.02	0.98	0.00	0.00	0.00
401	69	1	0.00	0.00	0.02	0.94	0.04	0.00	0.00
401	70	1	0.00	0.00	0.10	0.86	0.04	0.00	0.00
402	61	1	0.00	0.00	0.00	1.00	0.00	0.00	0.00
402	62	1	0.00	0.00	0.05	0.86	0.09	0.00	0.00
402	63	1	0.00	0.00	0.00	0.99	0.01	0.00	0.00
402	64	1	0.00	0.00	0.06	0.93	0.01	0.00	0.00
402	65	1	0.00	0.00	0.05	0.93	0.02	0.00	0.00
402	66	1	0.00	0.00	0.06	0.91	0.02	0.00	0.00
402	67	1	0.00	0.00	0.07	0.89	0.03	0.00	0.00
402	68	1	0.00	0.00	0.00	0.99	0.01	0.00	0.00
403	61	1	0.00	0.00	0.10	0.86	0.04	0.00	0.00
403	62	1	0.00	0.00	0.01	0.98	0.01	0.00	0.00
403	63	1	0.00	0.00	0.05	0.89	0.05	0.00	0.00
403	64	1	0.00	0.00	0.05	0.90	0.04	0.00	0.00
403	65	1	0.00	0.00	0.01	0.98	0.01	0.00	0.00
403	66	2	0.00	0.00	0.01	0.95	0.04	0.00	0.00
403	67	1	0.00	0.00	0.01	0.99	0.00	0.00	0.00
403	68	1	0.00	0.00	0.06	0.93	0.01	0.00	0.00
403	69	1	0.00	0.00	0.00	0.98	0.02	0.00	0.00
403	70	1	0.00	0.00	0.05	0.88	0.06	0.00	0.00
405	61	1	0.00	0.00	0.00	0.98	0.02	0.00	0.00
405	62	1	0.00	0.00	0.00	0.99	0.01	0.00	0.00
405	63	1	0.00	0.00	0.03	0.96	0.01	0.00	0.00

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

			Difference (First Read minus Second Read)						
Form	Item	Score Points	-3	-2	-1	0	1	2	3
405	64	1	0.00	0.00	0.08	0.85	0.08	0.00	0.00
405	65	1	0.00	0.00	0.10	0.80	0.10	0.00	0.00
405	66	1	0.00	0.00	0.01	0.99	0.00	0.00	0.00
405	67	1	0.00	0.00	0.03	0.92	0.05	0.00	0.00
405	68	2	0.00	0.00	0.05	0.92	0.04	0.00	0.00
405	69	1	0.00	0.00	0.11	0.82	0.07	0.00	0.00
405	70	1	0.00	0.00	0.07	0.90	0.03	0.00	0.00
406	61	1	0.00	0.00	0.00	0.99	0.01	0.00	0.00
406	62	1	0.00	0.00	0.01	0.98	0.01	0.00	0.00
406	63	1	0.00	0.00	0.05	0.86	0.08	0.00	0.00
406	64	1	0.00	0.00	0.00	0.98	0.02	0.00	0.00
406	65	1	0.00	0.00	0.00	1.00	0.00	0.00	0.00
406	66	1	0.00	0.00	0.10	0.86	0.04	0.00	0.00
406	67	1	0.00	0.00	0.02	0.98	0.00	0.00	0.00
406	68	1	0.00	0.00	0.01	0.99	0.00	0.00	0.00
406	69	1	0.00	0.00	0.00	0.99	0.01	0.00	0.00
406	70	1	0.00	0.00	0.03	0.97	0.00	0.00	0.00
406	71	1	0.00	0.00	0.02	0.90	0.08	0.00	0.00
407	61	1	0.00	0.00	0.02	0.95	0.03	0.00	0.00
407	62	1	0.00	0.00	0.02	0.98	0.00	0.00	0.00
407	63	3	0.00	0.00	0.06	0.91	0.03	0.00	0.00
407	64	1	0.00	0.00	0.00	1.00	0.00	0.00	0.00
407	65	1	0.00	0.00	0.00	0.99	0.01	0.00	0.00
407	66	1	0.00	0.00	0.00	1.00	0.00	0.00	0.00
407	67	1	0.00	0.00	0.02	0.97	0.01	0.00	0.00
407	68	1	0.00	0.00	0.04	0.91	0.05	0.00	0.00
408	61	1	0.00	0.00	0.03	0.93	0.04	0.00	0.00
408	63	1	0.00	0.00	0.05	0.87	0.07	0.00	0.00
408	64	1	0.00	0.00	0.00	0.97	0.03	0.00	0.00
408	65	1	0.00	0.00	0.01	0.97	0.02	0.00	0.00
408	66	1	0.00	0.00	0.01	0.99	0.00	0.00	0.00

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

			Difference (First Read minus Second Read)						
Form	Item	Score Points	-3	-2	-1	0	1	2	3
408	67	1	0.00	0.00	0.02	0.97	0.01	0.00	0.00
408	68	1	0.00	0.00	0.01	0.97	0.02	0.00	0.00
408	69	1	0.00	0.00	0.00	0.99	0.01	0.00	0.00
408	70	1	0.00	0.00	0.03	0.96	0.01	0.00	0.00
409	61	1	0.00	0.00	0.00	0.98	0.02	0.00	0.00
409	62	1	0.00	0.00	0.08	0.88	0.04	0.00	0.00
409	63	1	0.00	0.00	0.00	1.00	0.00	0.00	0.00
409	64	1	0.00	0.00	0.01	0.93	0.06	0.00	0.00
409	65	1	0.00	0.00	0.01	0.96	0.03	0.00	0.00
409	66	1	0.00	0.00	0.00	1.00	0.00	0.00	0.00
409	67	1	0.00	0.00	0.01	0.97	0.02	0.00	0.00
409	68	1	0.00	0.00	0.03	0.93	0.04	0.00	0.00
409	69	1	0.00	0.00	0.00	1.00	0.00	0.00	0.00
409	70	1	0.00	0.00	0.02	0.97	0.01	0.00	0.00

Table 5 contains additional summary information regarding the first and second reads. In the fourth column the percent of exact matches between the first and second scores is provided. 'Adj.' is the percentage of differences with a magnitude of one. 'Total' is the sum of the two prior columns and contains values between 99.0% 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	Intra-Class Correlation	Wt Kappa
401	61	1	104	79.8	20.2	100.0	0.7	0.7	0.45	0.45	0.49	0.49
401	62	1	103	88.3	11.7	100.0	0.7	0.6	0.47	0.48	0.75	0.74
401	63	1	104	98.1	1.9	100.0	0.8	0.8	0.39	0.37	0.93	0.93
401	64	1	103	98.1	1.9	100.0	0.9	0.9	0.34	0.34	0.92	0.92
401	65	1	103	86.4	13.6	100.0	0.7	0.7	0.47	0.48	0.70	0.69
401	66	1	97	89.7	10.3	100.0	0.6	0.6	0.49	0.49	0.78	0.78
401	67	1	102	90.2	9.8	100.0	0.6	0.6	0.50	0.49	0.80	0.80
401	68	1	96	97.9	2.1	100.0	0.7	0.7	0.47	0.46	0.95	0.95
401	69	1	99	93.9	6.1	100.0	0.6	0.6	0.49	0.49	0.87	0.87
401	70	1	101	86.1	13.9	100.0	0.4	0.4	0.49	0.50	0.72	0.72
402	61	1	96	100.0	0.0	100.0	0.8	0.8	0.38	0.38	1.00	1.00
402	62	1	94	86.2	13.8	100.0	0.6	0.6	0.49	0.49	0.71	0.71
402	63	1	97	99.0	1.0	100.0	0.8	0.8	0.41	0.42	0.97	0.97
402	64	1	96	92.7	7.3	100.0	0.5	0.6	0.50	0.50	0.86	0.85
402	65	1	96	92.7	7.3	100.0	0.6	0.7	0.48	0.47	0.84	0.84
402	66	1	93	91.4	8.6	100.0	0.6	0.6	0.50	0.49	0.82	0.82
402	67	1	94	89.4	10.6	100.0	0.7	0.7	0.46	0.44	0.74	0.74
402	68	1	97	99.0	1.0	100.0	0.8	0.8	0.42	0.43	0.97	0.97
403	61	1	92	85.9	14.1	100.0	0.7	0.7	0.47	0.44	0.66	0.66
403	62	1	91	97.8	2.2	100.0	0.5	0.5	0.50	0.50	0.96	0.96
403	63	1	94	89.4	10.6	100.0	0.7	0.7	0.45	0.45	0.74	0.74
403	64	1	94	90.4	9.6	100.0	0.6	0.7	0.48	0.48	0.79	0.79
403	65	1	94	97.9	2.1	100.0	0.8	0.8	0.40	0.40	0.93	0.93
403	66	2	92	94.6	5.4	100.0	1.1	1.1	0.65	0.65	0.94	0.92
403	67	1	93	98.9	1.1	100.0	0.8	0.8	0.43	0.42	0.97	0.97
403	68	1	94	92.6	7.4	100.0	0.8	0.9	0.38	0.32	0.72	0.70
403	69	1	94	97.9	2.1	100.0	0.9	0.8	0.36	0.38	0.92	0.92
403	70	1	93	88.2	11.8	100.0	0.3	0.3	0.45	0.44	0.70	0.70
405	61	1	135	97.8	2.2	100.0	0.6	0.5	0.50	0.50	0.96	0.95

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	Intra-Class Correlation	Wt Kappa
405	62	1	136	99.3	0.7	100.0	0.8	0.8	0.42	0.42	0.98	0.98
405	63	1	135	95.6	4.4	100.0	0.8	0.8	0.38	0.37	0.84	0.84
405	64	1	132	84.8	15.2	100.0	0.7	0.7	0.47	0.47	0.65	0.65
405	65	1	134	79.9	20.1	100.0	0.8	0.8	0.43	0.43	0.45	0.45
405	66	1	134	99.3	0.7	100.0	0.4	0.4	0.50	0.50	0.98	0.98
405	67	1	129	92.2	7.8	100.0	0.4	0.4	0.49	0.49	0.84	0.84
405	68	2	133	91.7	8.3	100.0	1.1	1.1	0.74	0.74	0.92	0.90
405	69	1	135	82.2	17.8	100.0	0.7	0.7	0.46	0.44	0.56	0.55
405	70	1	134	89.6	10.4	100.0	0.8	0.9	0.37	0.33	0.58	0.57
406	61	1	94	98.9	1.1	100.0	0.6	0.6	0.50	0.50	0.98	0.98
406	62	1	95	97.9	2.1	100.0	0.9	0.9	0.33	0.33	0.90	0.90
406	63	1	95	86.3	13.7	100.0	0.6	0.6	0.49	0.50	0.72	0.72
406	64	1	93	97.8	2.2	100.0	0.4	0.3	0.48	0.47	0.95	0.95
406	65	1	92	100.0	0.0	100.0	0.7	0.7	0.46	0.46	1.00	1.00
406	66	1	90	85.6	14.4	100.0	0.7	0.7	0.47	0.44	0.65	0.65
406	67	1	93	97.8	2.2	100.0	0.4	0.4	0.48	0.49	0.95	0.95
406	68	1	93	98.9	1.1	100.0	0.8	0.8	0.39	0.38	0.96	0.96
406	69	1	89	98.9	1.1	100.0	0.4	0.4	0.49	0.49	0.98	0.98
406	70	1	89	96.6	3.4	100.0	0.5	0.5	0.50	0.50	0.93	0.93
406	71	1	91	90.1	9.9	100.0	0.3	0.3	0.47	0.44	0.77	0.76
407	61	1	97	94.8	5.2	100.0	0.6	0.6	0.48	0.48	0.89	0.89
407	62	1	96	97.9	2.1	100.0	0.6	0.6	0.49	0.48	0.96	0.96
407	63	3	96	89.6	9.4	99.0	1.5	1.5	0.93	0.93	0.89	0.88
407	64	1	97	100.0	0.0	100.0	0.5	0.5	0.50	0.50	1.00	1.00
407	65	1	97	99.0	1.0	100.0	0.6	0.6	0.49	0.49	0.98	0.98
407	66	1	94	100.0	0.0	100.0	0.5	0.5	0.50	0.50	1.00	1.00
407	67	1	93	96.8	3.2	100.0	0.6	0.6	0.48	0.48	0.93	0.93
407	68	1	95	90.5	9.5	100.0	0.3	0.3	0.47	0.46	0.78	0.78
408	61	1	96	92.7	7.3	100.0	0.5	0.5	0.50	0.50	0.85	0.85
408	63	1	94	87.2	12.8	100.0	0.7	0.6	0.47	0.48	0.72	0.72

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	Intra-Class Correlation	Wt Kappa
408	64	1	97	96.9	3.1	100.0	0.6	0.6	0.49	0.49	0.94	0.93
408	65	1	98	96.9	3.1	100.0	0.9	0.9	0.32	0.33	0.85	0.85
408	66	1	98	99.0	1.0	100.0	0.9	0.9	0.29	0.28	0.94	0.94
408	67	1	97	96.9	3.1	100.0	0.3	0.3	0.45	0.46	0.92	0.92
408	68	1	98	96.9	3.1	100.0	0.7	0.7	0.45	0.45	0.92	0.92
408	69	1	97	99.0	1.0	100.0	0.9	0.9	0.22	0.24	0.91	0.90
408	70	1	97	95.9	4.1	100.0	0.8	0.8	0.40	0.38	0.87	0.86
409	61	1	98	98.0	2.0	100.0	0.5	0.5	0.50	0.50	0.96	0.96
409	62	1	97	87.6	12.4	100.0	0.3	0.4	0.48	0.49	0.74	0.73
409	63	1	98	100.0	0.0	100.0	0.9	0.9	0.33	0.33	1.00	1.00
409	64	1	98	92.9	7.1	100.0	0.8	0.7	0.41	0.44	0.81	0.80
409	65	1	98	95.9	4.1	100.0	0.6	0.6	0.50	0.50	0.92	0.92
409	66	1	97	100.0	0.0	100.0	0.4	0.4	0.49	0.49	1.00	1.00
409	67	1	94	96.8	3.2	100.0	0.9	0.8	0.36	0.37	0.88	0.88
409	68	1	98	92.9	7.1	100.0	0.2	0.2	0.41	0.41	0.78	0.78
409	69	1	98	100.0	0.0	100.0	0.4	0.4	0.49	0.49	1.00	1.00
409	70	1	98	96.9	3.1	100.0	0.3	0.4	0.48	0.48	0.93	0.93

* Adj. = difference of one

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.

Intra-class Correlation

The intra-class 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

reads. Correlations greater than 0.60 are considered very strong because they explain more than one-third of the variance in scores. All but four items had intra-class correlations greater than or equal to 0.65 (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. The Weighted Kappa statistics ranged from 0.45 to 1.0, 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 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 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 they have 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 provide an advantage over the use of classical item means because they can be compared across test forms. Different samples of students 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 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 may 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. 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. Nearly all of the items fell within the moderate -2.0 to +2.0 difficulty range and only two items had an INFIT statistic outside the typical range.

Table 6. Partial Credit Model Item Analysis

Form	N-Count	No of Items	Rasch			INFIT		
			<-2.0	-2.0 to 2.0	>2.0	<-0.70	-0.70 to 1.30	>1.30
401	965	23	0	23	0	0	23	0
402	975	23	1	20	0	0	20	1
403	941	21	0	17	1	0	18	0
404	976	30	0	30	0	0	30	0
405	988	21	0	19	0	0	19	0
406	967	24	1	23	0	0	24	0
407	966	21	1	20	0	0	20	1
408	971	23	0	22	0	0	22	0
409	968	23	1	21	1	0	23	0

* A complete listing of all Partial Credit Model item analysis results is presented in Appendix B.

** 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.

All of the individual IRT item statistics are provided in Appendix B. The column titled 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 $\{|MHD| \geq 1.0 \text{ and } < 1.5\}$ **or** MHD not significantly different from 0 and $|MHD| \geq 1.0$ are classified as B.
- $|MHD| \geq 1.5$ 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 $|ES| \leq 0.17$, classified as AA.
- Probability is > 0.05 and if $0.17 < |ES| \leq 0.25$, classified as BB.
- Probability is > 0.05 and if $|ES| > 0.25$, 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 2010 field test administration for the New York State Examination in Grade 8 Intermediate-Level Science consisted of 8 field test forms numbered 401–403 and 405–409 and one anchor form labeled 404. The field test forms contained multiple-choice and constructed-response items. All students participating in the field test were administered one of the 9 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 2009 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 non-extreme students. This average ability estimate was used to equate the remaining field test forms as well as updating the item parameters for the anchor form.

As part of the anchor item equating, an item-stability check was performed. After fixing all of the items to their 2009 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 two items having a displacement value with a magnitude greater than 0.30. This indicates stability in the items used on the anchor form.

The equated mean ability estimate for form 404 was 0.81. This value served as the target mean ability for the remainder of 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 404). 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
401	0.48	0.31
402	0.94	-0.13
403	0.78	0.03
405	0.79	0.02
406	0.84	-0.03
407	0.91	-0.10
408	0.81	-0.01
409	0.78	0.03

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 May 2010 operational test map with content information regarding each item included on the form.

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 0.068 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 table used for the May 2010 operational test form. Four steps are taken in order to produce this table and resulting conversion chart.

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 May 2006. 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 May 2006 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 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, 213–20.
- Cronbach, L.J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- Dorans, N.J. and Holland, P.W. 1992. DIF Detection and Description: Mantel–Haenszel and Standardization. In *Differential Item Functioning: Theory and Practice*, edited by P. W. Holland and H. Wainer, 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
2010_G8Sc_FT	401	MC	01	1	965	0.83	0.00		0.52	0.14	0.27	0.07			0.52	0.48
2010_G8Sc_FT	401	MC	02	1	965	0.83	0.00		0.03	0.41	0.03	0.53			0.41	0.45
2010_G8Sc_FT	401	MC	03	1	965	0.83	0.00		0.65	0.03	0.25	0.06			0.65	0.53
2010_G8Sc_FT	401	MC	04	1	965	0.83	0.00		0.04	0.19	0.26	0.51			0.51	0.54
2010_G8Sc_FT	401	MC	05	1	965	0.83	0.00		0.36	0.44	0.12	0.07			0.44	0.33
2010_G8Sc_FT	401	MC	06	1	965	0.83	0.00		0.06	0.54	0.28	0.13			0.54	0.56
2010_G8Sc_FT	401	MC	07	1	965	0.83	0.00		0.04	0.15	0.79	0.02			0.79	0.37
2010_G8Sc_FT	401	MC	08	1	965	0.83	0.00		0.16	0.62	0.10	0.12			0.62	0.46
2010_G8Sc_FT	401	MC	09	1	965	0.83	0.00		0.03	0.02	0.87	0.07			0.87	0.29
2010_G8Sc_FT	401	MC	10	1	965	0.83	0.00		0.23	0.07	0.05	0.64			0.64	0.49
2010_G8Sc_FT	401	MC	11	1	965	0.83	0.00		0.15	0.36	0.35	0.14			0.36	0.37
2010_G8Sc_FT	401	MC	12	1	965	0.83	0.01		0.18	0.33	0.43	0.06			0.43	0.40
2010_G8Sc_FT	401	MC	13	1	965	0.83	0.02		0.16	0.65	0.08	0.09			0.65	0.44
2010_G8Sc_FT	401	CR	61	1	965	0.83	0.01	0.22	0.77						0.77	0.32
2010_G8Sc_FT	401	CR	62	1	965	0.83	0.04	0.36	0.60						0.60	0.47
2010_G8Sc_FT	401	CR	63	1	965	0.83	0.02	0.22	0.77						0.77	0.42
2010_G8Sc_FT	401	CR	64	1	965	0.83	0.02	0.16	0.82						0.82	0.47
2010_G8Sc_FT	401	CR	65	1	965	0.83	0.09	0.37	0.54						0.54	0.50
2010_G8Sc_FT	401	CR	66	1	965	0.83	0.11	0.38	0.51						0.51	0.48
2010_G8Sc_FT	401	CR	67	1	965	0.83	0.08	0.38	0.54						0.54	0.57
2010_G8Sc_FT	401	CR	68	1	965	0.83	0.12	0.29	0.59						0.59	0.55
2010_G8Sc_FT	401	CR	69	1	965	0.83	0.09	0.37	0.54						0.54	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
2010_G8Sc_FT	401	CR	70	1	965	0.83	0.09	0.50	0.41						0.41	0.35
2010_G8Sc_FT	402	MC	01	1	975	0.79	0.01		0.11	0.13	0.63	0.12			0.63	0.43
2010_G8Sc_FT	402	MC	02	1	975	0.79	0.00		0.02	0.06	0.11	0.81			0.81	0.45
2010_G8Sc_FT	402	MC	03	1	975	0.79	0.00		0.05	0.01	0.01	0.93			0.93	0.44
2010_G8Sc_FT	402	MC	04	1	975	0.79	0.00		0.63	0.26	0.03	0.07			0.63	0.58
2010_G8Sc_FT	402	MC	05	1	975	0.79	0.00		0.08	0.03	0.03	0.87			0.87	0.53
2010_G8Sc_FT	402	MC	06	1	975	0.79	0.00		0.18	0.32	0.21	0.28			0.28	0.37
2010_G8Sc_FT	402	MC	07	1	975	0.79	0.00		0.06	0.12	0.75	0.07			0.75	0.45
2010_G8Sc_FT	402	MC	08	1	975	0.79	0.00		0.29	0.57	0.10	0.04			0.57	0.41
2010_G8Sc_FT	402	MC	09	1	975	0.79	0.00		0.69	0.10	0.15	0.06			0.69	0.40
2010_G8Sc_FT	402	MC	10	1	975	0.79	0.00		0.39	0.27	0.24	0.09			0.39	0.23
2010_G8Sc_FT	402	MC	11	1	975	0.79	0.01		0.64	0.04	0.05	0.26			0.64	0.29
2010_G8Sc_FT	402	MC	12	1	975	0.79	0.01		0.05	0.13	0.64	0.17			0.64	0.21
2010_G8Sc_FT	402	MC	13	1	975	0.79	0.01		0.14	0.03	0.76	0.06			0.76	0.47
2010_G8Sc_FT	402	CR	61	1	975	0.79	0.00	0.13	0.87						0.87	0.46
2010_G8Sc_FT	402	CR	62	1	975	0.79	0.03	0.39	0.58						0.58	0.47
2010_G8Sc_FT	402	CR	63	1	975	0.79	0.01	0.24	0.75						0.75	0.43
2010_G8Sc_FT	402	CR	64	1	975	0.79	0.02	0.33	0.65						0.65	0.57
2010_G8Sc_FT	402	CR	65	1	975	0.79	0.03	0.36	0.61						0.61	0.48
2010_G8Sc_FT	402	CR	66	1	975	0.79	0.05	0.41	0.54						0.54	0.60
2010_G8Sc_FT	402	CR	67	1	975	0.79	0.04	0.24	0.72						0.72	0.53
2010_G8Sc_FT	402	CR	68	1	975	0.79	0.03	0.20	0.77						0.77	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
2010_G8Sc_FT	402	CR	69	1	975											
2010_G8Sc_FT	402	CR	70	1	975											
2010_G8Sc_FT	403	MC	01	1	941	0.80	0.00		0.51	0.07	0.09	0.32			0.51	0.49
2010_G8Sc_FT	403	MC	02	1	941	0.80	0.00		0.06	0.06	0.03	0.85			0.85	0.48
2010_G8Sc_FT	403	MC	03	1	941	0.80	0.00		0.82	0.06	0.10	0.01			0.82	0.41
2010_G8Sc_FT	403	MC	04	1												
2010_G8Sc_FT	403	MC	05	1	941	0.80	0.00		0.08	0.81	0.05	0.06			0.81	0.51
2010_G8Sc_FT	403	MC	06	1	941	0.80	0.00		0.39	0.40	0.17	0.03			0.39	0.43
2010_G8Sc_FT	403	MC	07	1	941	0.80	0.00		0.52	0.33	0.08	0.06			0.52	0.43
2010_G8Sc_FT	403	MC	08	1	941	0.80	0.00		0.26	0.55	0.18	0.01			0.55	0.45
2010_G8Sc_FT	403	MC	09	1												
2010_G8Sc_FT	403	MC	10	1	941	0.80	0.01		0.10	0.84	0.05	0.01			0.84	0.43
2010_G8Sc_FT	403	CR	61	1	941	0.80	0.03	0.30	0.67						0.67	0.41
2010_G8Sc_FT	403	CR	62	1	941	0.80	0.06	0.52	0.42						0.42	0.43
2010_G8Sc_FT	403	CR	63	1	941	0.80	0.03	0.34	0.63						0.63	0.54
2010_G8Sc_FT	403	CR	64	1	941	0.80	0.02	0.35	0.63						0.63	0.53
2010_G8Sc_FT	403	CR	65	1	941	0.80	0.03	0.28	0.69						0.69	0.47
2010_G8Sc_FT	403	CR	66	2	941	0.80	0.06	0.15	0.54	0.25					1.05	0.62
2010_G8Sc_FT	403	CR	67	1	941	0.80	0.03	0.24	0.73						0.73	0.53
2010_G8Sc_FT	403	CR	68	1	941	0.80	0.03	0.15	0.82						0.82	0.40
2010_G8Sc_FT	403	CR	69	1	941	0.80	0.07	0.16	0.77						0.77	0.51
2010_G8Sc_FT	403	CR	70	1	941	0.80	0.12	0.60	0.28						0.28	0.50

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
2010_G8Sc_FT	403	CR	71	1												
2010_G8Sc_FT	404	MC	01	1	976	0.82	0.00		0.02	0.11	0.17	0.70			0.70	0.31
2010_G8Sc_FT	404	MC	02	1	976	0.82	0.00		0.04	0.11	0.65	0.20			0.65	0.46
2010_G8Sc_FT	404	MC	03	1	976	0.82	0.00		0.07	0.06	0.23	0.65			0.65	0.42
2010_G8Sc_FT	404	MC	04	1	976	0.82	0.00		0.05	0.75	0.17	0.03			0.75	0.30
2010_G8Sc_FT	404	MC	05	1	976	0.82	0.00		0.06	0.21	0.21	0.52			0.52	0.31
2010_G8Sc_FT	404	MC	06	1	976	0.82	0.00		0.15	0.29	0.51	0.04			0.51	0.38
2010_G8Sc_FT	404	MC	07	1	976	0.82	0.00		0.29	0.15	0.40	0.16			0.40	0.20
2010_G8Sc_FT	404	MC	08	1	976	0.82	0.00		0.38	0.14	0.41	0.06			0.41	0.32
2010_G8Sc_FT	404	MC	09	1	976	0.82	0.00		0.07	0.85	0.05	0.03			0.85	0.46
2010_G8Sc_FT	404	MC	10	1	976	0.82	0.00		0.13	0.03	0.04	0.80			0.80	0.38
2010_G8Sc_FT	404	MC	11	1	976	0.82	0.00		0.07	0.75	0.13	0.04			0.75	0.45
2010_G8Sc_FT	404	MC	12	1	976	0.82	0.00		0.05	0.04	0.23	0.68			0.68	0.49
2010_G8Sc_FT	404	MC	13	1	976	0.82	0.00		0.12	0.72	0.13	0.03			0.72	0.31
2010_G8Sc_FT	404	MC	14	1	976	0.82	0.01		0.74	0.16	0.03	0.06			0.74	0.48
2010_G8Sc_FT	404	MC	15	1	976	0.82	0.01		0.05	0.77	0.05	0.12			0.77	0.42
2010_G8Sc_FT	404	MC	16	1	976	0.82	0.01		0.14	0.64	0.03	0.19			0.64	0.49
2010_G8Sc_FT	404	MC	17	1	976	0.82	0.01		0.07	0.08	0.04	0.80			0.80	0.49
2010_G8Sc_FT	404	MC	18	1	976	0.82	0.01		0.49	0.31	0.15	0.04			0.49	0.36
2010_G8Sc_FT	404	MC	19	1	976	0.82	0.01		0.03	0.13	0.52	0.31			0.52	0.39
2010_G8Sc_FT	404	MC	20	1	976	0.82	0.01		0.72	0.05	0.14	0.08			0.72	0.48
2010_G8Sc_FT	404	MC	21	1	976	0.82	0.01		0.10	0.10	0.72	0.06			0.72	0.56

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
2010_G8Sc_FT	404	MC	22	1	976	0.82	0.01		0.49	0.21	0.20	0.09			0.49	0.43
2010_G8Sc_FT	404	MC	23	1	976	0.82	0.02		0.11	0.60	0.09	0.18			0.60	0.46
2010_G8Sc_FT	404	MC	24	1	976	0.82	0.02		0.88	0.03	0.04	0.03			0.88	0.51
2010_G8Sc_FT	404	MC	25	1	976	0.82	0.02		0.04	0.04	0.08	0.83			0.83	0.49
2010_G8Sc_FT	404	MC	26	1	976	0.82	0.02		0.25	0.19	0.14	0.40			0.40	0.28
2010_G8Sc_FT	404	MC	27	1	976	0.82	0.02		0.10	0.37	0.33	0.19			0.37	0.22
2010_G8Sc_FT	404	MC	28	1	976	0.82	0.02		0.18	0.06	0.08	0.66			0.66	0.52
2010_G8Sc_FT	404	MC	29	1	976	0.82	0.02		0.10	0.79	0.04	0.04			0.79	0.52
2010_G8Sc_FT	404	MC	30	1	976	0.82	0.03		0.15	0.07	0.08	0.67			0.67	0.44
2010_G8Sc_FT	405	MC	01	1	988	0.80	0.00		0.06	0.05	0.87	0.01			0.87	0.41
2010_G8Sc_FT	405	MC	02	1	988	0.80	0.00		0.10	0.22	0.13	0.55			0.55	0.47
2010_G8Sc_FT	405	MC	03	1	988	0.80	0.00		0.03	0.03	0.85	0.09			0.85	0.41
2010_G8Sc_FT	405	MC	04	1												
2010_G8Sc_FT	405	MC	05	1	988	0.80	0.00		0.22	0.45	0.16	0.16			0.45	0.38
2010_G8Sc_FT	405	MC	06	1	988	0.80	0.00		0.10	0.18	0.14	0.58			0.58	0.43
2010_G8Sc_FT	405	MC	07	1	988	0.80	0.00		0.11	0.06	0.14	0.69			0.69	0.50
2010_G8Sc_FT	405	MC	08	1	988	0.80	0.00		0.12	0.18	0.62	0.07			0.62	0.40
2010_G8Sc_FT	405	MC	09	1	988	0.80	0.00		0.87	0.04	0.04	0.05			0.87	0.41
2010_G8Sc_FT	405	MC	10	1	988	0.80	0.01		0.05	0.24	0.01	0.69			0.69	0.36
2010_G8Sc_FT	405	CR	61	1	988	0.80	0.01	0.45	0.54						0.54	0.53
2010_G8Sc_FT	405	CR	62	1	988	0.80	0.02	0.20	0.78						0.78	0.53
2010_G8Sc_FT	405	CR	63	1	988	0.80	0.03	0.21	0.76						0.76	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
2010_G8Sc_FT	405	CR	64	1	988	0.80	0.05	0.34	0.61						0.61	0.57
2010_G8Sc_FT	405	CR	65	1	988	0.80	0.05	0.24	0.70						0.70	0.49
2010_G8Sc_FT	405	CR	66	1	988	0.80	0.05	0.59	0.37						0.37	0.46
2010_G8Sc_FT	405	CR	67	1	988	0.80	0.08	0.59	0.33						0.33	0.41
2010_G8Sc_FT	405	CR	68	2	988	0.80	0.05	0.22	0.43	0.30					1.02	0.63
2010_G8Sc_FT	405	CR	69	1	988	0.80	0.07	0.26	0.68						0.68	0.50
2010_G8Sc_FT	405	CR	70	1	988	0.80	0.08	0.16	0.76						0.76	0.53
2010_G8Sc_FT	405	CR	71	1												
2010_G8Sc_FT	406	MC	01	1	967	0.79	0.00		0.20	0.77	0.02	0.01			0.77	0.42
2010_G8Sc_FT	406	MC	02	1	967	0.79	0.01		0.04	0.05	0.86	0.04			0.86	0.39
2010_G8Sc_FT	406	MC	03	1	967	0.79	0.00		0.02	0.01	0.10	0.87			0.87	0.27
2010_G8Sc_FT	406	MC	04	1	967	0.79	0.00		0.94	0.02	0.02	0.02			0.94	0.35
2010_G8Sc_FT	406	MC	05	1	967	0.79	0.00		0.75	0.12	0.10	0.03			0.75	0.42
2010_G8Sc_FT	406	MC	06	1	967	0.79	0.00		0.42	0.13	0.39	0.05			0.42	0.19
2010_G8Sc_FT	406	MC	07	1	967	0.79	0.00		0.05	0.90	0.03	0.02			0.90	0.40
2010_G8Sc_FT	406	MC	08	1	967	0.79	0.00		0.17	0.07	0.72	0.04			0.72	0.45
2010_G8Sc_FT	406	MC	09	1	967	0.79	0.00		0.13	0.17	0.08	0.62			0.62	0.42
2010_G8Sc_FT	406	MC	10	1	967	0.79	0.00		0.13	0.10	0.70	0.07			0.70	0.54
2010_G8Sc_FT	406	MC	11	1	967	0.79	0.00		0.15	0.10	0.68	0.07			0.68	0.42
2010_G8Sc_FT	406	MC	12	1	967	0.79	0.00		0.57	0.08	0.32	0.03			0.57	0.37
2010_G8Sc_FT	406	MC	13	1	967	0.79	0.02		0.20	0.09	0.64	0.06			0.64	0.29
2010_G8Sc_FT	406	CR	61	1	967	0.79	0.01	0.41	0.57						0.57	0.39

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
2010_G8Sc_FT	406	CR	62	1	967	0.79	0.01	0.11	0.88						0.88	0.34
2010_G8Sc_FT	406	CR	63	1	967	0.79	0.03	0.38	0.59						0.59	0.46
2010_G8Sc_FT	406	CR	64	1	967	0.79	0.04	0.56	0.40						0.40	0.50
2010_G8Sc_FT	406	CR	65	1	967	0.79	0.03	0.34	0.63						0.63	0.46
2010_G8Sc_FT	406	CR	66	1	967	0.79	0.08	0.30	0.62						0.62	0.50
2010_G8Sc_FT	406	CR	67	1	967	0.79	0.06	0.59	0.35						0.35	0.38
2010_G8Sc_FT	406	CR	68	1	967	0.79	0.04	0.26	0.71						0.71	0.55
2010_G8Sc_FT	406	CR	69	1	967	0.79	0.08	0.47	0.45						0.45	0.50
2010_G8Sc_FT	406	CR	70	1	967	0.79	0.11	0.37	0.53						0.53	0.45
2010_G8Sc_FT	406	CR	71	1	967	0.79	0.11	0.60	0.28						0.28	0.42
2010_G8Sc_FT	407	MC	01	1	966	0.74	0.00		0.02	0.02	0.01	0.95			0.95	0.38
2010_G8Sc_FT	407	MC	02	1	966	0.74	0.00		0.82	0.09	0.05	0.05			0.82	0.36
2010_G8Sc_FT	407	MC	03	1	966	0.74	0.00		0.16	0.20	0.13	0.51			0.51	0.47
2010_G8Sc_FT	407	MC	04	1	966	0.74	0.00		0.02	0.06	0.88	0.04			0.88	0.49
2010_G8Sc_FT	407	MC	05	1	966	0.74	0.00		0.09	0.85	0.02	0.04			0.85	0.45
2010_G8Sc_FT	407	MC	06	1	966	0.74	0.00		0.05	0.83	0.04	0.07			0.83	0.45
2010_G8Sc_FT	407	MC	07	1	966	0.74	0.00		0.34	0.12	0.48	0.06			0.34	0.37
2010_G8Sc_FT	407	MC	08	1	966	0.74	0.00		0.07	0.05	0.01	0.87			0.87	0.36
2010_G8Sc_FT	407	MC	09	1	966	0.74	0.00		0.04	0.12	0.09	0.75			0.75	0.47
2010_G8Sc_FT	407	MC	10	1	966	0.74	0.00		0.12	0.04	0.80	0.04			0.80	0.44
2010_G8Sc_FT	407	MC	11	1	966	0.74	0.00		0.31	0.64	0.03	0.01			0.64	0.28
2010_G8Sc_FT	407	MC	12	1	966	0.74	0.00		0.18	0.05	0.66	0.10			0.66	0.23

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
2010_G8Sc_FT	407	MC	13	1	966	0.74	0.02		0.61	0.10	0.16	0.11			0.61	0.43
2010_G8Sc_FT	407	CR	61	1	966	0.74	0.03	0.35	0.62						0.62	0.38
2010_G8Sc_FT	407	CR	62	1	966	0.74	0.02	0.28	0.70						0.70	0.34
2010_G8Sc_FT	407	CR	63	3	966	0.74	0.04	0.14	0.26	0.45	0.10				1.47	0.69
2010_G8Sc_FT	407	CR	64	1	966	0.74	0.01	0.45	0.54						0.54	0.47
2010_G8Sc_FT	407	CR	65	1	966	0.74	0.01	0.44	0.55						0.55	0.07
2010_G8Sc_FT	407	CR	66	1	966	0.74	0.06	0.44	0.50						0.50	0.48
2010_G8Sc_FT	407	CR	67	1	966	0.74	0.07	0.42	0.51						0.51	0.51
2010_G8Sc_FT	407	CR	68	1	966	0.74	0.05	0.64	0.30						0.30	0.43
2010_G8Sc_FT	408	MC	01	1	971	0.74	0.00		0.21	0.15	0.56	0.07			0.56	0.45
2010_G8Sc_FT	408	MC	02	1	971	0.74	0.00		0.29	0.40	0.28	0.03			0.40	0.28
2010_G8Sc_FT	408	MC	03	1	971	0.74	0.00		0.05	0.64	0.06	0.24			0.64	0.38
2010_G8Sc_FT	408	MC	04	1	971	0.74	0.00		0.01	0.83	0.12	0.04			0.83	0.28
2010_G8Sc_FT	408	MC	05	1	971	0.74	0.00		0.84	0.09	0.05	0.01			0.84	0.39
2010_G8Sc_FT	408	MC	06	1	971	0.74	0.00		0.86	0.04	0.06	0.03			0.86	0.43
2010_G8Sc_FT	408	MC	07	1	971	0.74	0.00		0.52	0.12	0.08	0.27			0.52	0.52
2010_G8Sc_FT	408	MC	08	1	971	0.74	0.00		0.41	0.23	0.27	0.09			0.41	0.18
2010_G8Sc_FT	408	MC	09	1	971	0.74	0.00		0.06	0.89	0.03	0.02			0.89	0.40
2010_G8Sc_FT	408	MC	10	1	971	0.74	0.00		0.08	0.20	0.15	0.58			0.58	0.40
2010_G8Sc_FT	408	MC	11	1	971	0.74	0.00		0.02	0.02	0.74	0.21			0.74	0.39
2010_G8Sc_FT	408	MC	12	1	971	0.74	0.00		0.16	0.05	0.09	0.69			0.69	0.31
2010_G8Sc_FT	408	MC	13	1	971	0.74	0.01		0.38	0.10	0.13	0.38			0.38	0.25

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
2010_G8Sc_FT	408	CR	61	1	971	0.74	0.05	0.51	0.44						0.44	0.53
2010_G8Sc_FT	408	CR	62	1												
2010_G8Sc_FT	408	CR	63	1	971	0.74	0.06	0.44	0.50						0.50	0.53
2010_G8Sc_FT	408	CR	64	1	971	0.74	0.05	0.40	0.54						0.54	0.59
2010_G8Sc_FT	408	CR	65	1	971	0.74	0.02	0.18	0.80						0.80	0.51
2010_G8Sc_FT	408	CR	66	1	971	0.74	0.01	0.11	0.88						0.88	0.33
2010_G8Sc_FT	408	CR	67	1	971	0.74	0.01	0.70	0.29						0.29	0.36
2010_G8Sc_FT	408	CR	68	1	971	0.74	0.02	0.31	0.67						0.67	0.44
2010_G8Sc_FT	408	CR	69	1	971	0.74	0.03	0.09	0.88						0.88	0.38
2010_G8Sc_FT	408	CR	70	1	971	0.74	0.04	0.24	0.72						0.72	0.43
2010_G8Sc_FT	409	MC	01	1	968	0.79	0.00		0.91	0.03	0.03	0.04			0.91	0.45
2010_G8Sc_FT	409	MC	02	1	968	0.79	0.00		0.10	0.82	0.06	0.02			0.82	0.49
2010_G8Sc_FT	409	MC	03	1	968	0.79	0.00		0.04	0.05	0.82	0.09			0.82	0.41
2010_G8Sc_FT	409	MC	04	1	968	0.79	0.00		0.17	0.76	0.04	0.04			0.76	0.33
2010_G8Sc_FT	409	MC	05	1	968	0.79	0.00		0.03	0.47	0.35	0.16			0.47	0.45
2010_G8Sc_FT	409	MC	06	1	968	0.79	0.00		0.67	0.07	0.16	0.09			0.67	0.43
2010_G8Sc_FT	409	MC	07	1	968	0.79	0.00		0.96	0.01	0.02	0.01			0.96	0.29
2010_G8Sc_FT	409	MC	08	1	968	0.79	0.00		0.16	0.17	0.57	0.10			0.57	0.46
2010_G8Sc_FT	409	MC	09	1	968	0.79	0.00		0.34	0.05	0.42	0.18			0.42	0.43
2010_G8Sc_FT	409	MC	10	1	968	0.79	0.00		0.70	0.07	0.17	0.05			0.70	0.42
2010_G8Sc_FT	409	MC	11	1	968	0.79	0.00		0.11	0.04	0.51	0.33			0.51	0.26
2010_G8Sc_FT	409	MC	12	1	968	0.79	0.00		0.13	0.72	0.11	0.04			0.72	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
2010_G8Sc_FT	409	MC	13	1	968	0.79	0.01		0.03	0.02	0.11	0.83			0.83	0.38
2010_G8Sc_FT	409	CR	61	1	968	0.79	0.02	0.52	0.46						0.46	0.45
2010_G8Sc_FT	409	CR	62	1	968	0.79	0.02	0.55	0.43						0.43	0.52
2010_G8Sc_FT	409	CR	63	1	968	0.79	0.00	0.11	0.88						0.88	0.34
2010_G8Sc_FT	409	CR	64	1	968	0.79	0.02	0.18	0.80						0.80	0.48
2010_G8Sc_FT	409	CR	65	1	968	0.79	0.07	0.40	0.53						0.53	0.58
2010_G8Sc_FT	409	CR	66	1	968	0.79	0.05	0.55	0.40						0.40	0.51
2010_G8Sc_FT	409	CR	67	1	968	0.79	0.04	0.17	0.79						0.79	0.50
2010_G8Sc_FT	409	CR	68	1	968	0.79	0.04	0.75	0.21						0.21	0.31
2010_G8Sc_FT	409	CR	69	1	968	0.79	0.04	0.60	0.36						0.36	0.45
2010_G8Sc_FT	409	CR	70	1	968	0.79	0.03	0.59	0.38						0.38	0.39

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
2010_G8Sc_FT	401	MC	01	1	965	0.7193							0.99
2010_G8Sc_FT	401	MC	02	1	965	1.2740							1.00
2010_G8Sc_FT	401	MC	03	1	965	0.0217							0.92
2010_G8Sc_FT	401	MC	04	1	965	0.7603							0.92
2010_G8Sc_FT	401	MC	05	1	965	1.0745							1.17
2010_G8Sc_FT	401	MC	06	1	965	0.6165							0.89
2010_G8Sc_FT	401	MC	07	1	965	-0.8434							1.06
2010_G8Sc_FT	401	MC	08	1	965	0.1699							1.01
2010_G8Sc_FT	401	MC	09	1	965	-1.4993							1.06
2010_G8Sc_FT	401	MC	10	1	965	0.0716							0.97
2010_G8Sc_FT	401	MC	11	1	965	1.5181							1.10
2010_G8Sc_FT	401	MC	12	1	965	1.1370							1.09
2010_G8Sc_FT	401	MC	13	1	965	0.0495							1.02
2010_G8Sc_FT	401	CR	61	1	965	-0.6643							1.10
2010_G8Sc_FT	401	CR	62	1	965	0.3091							1.00
2010_G8Sc_FT	401	CR	63	1	965	-0.6505							1.00
2010_G8Sc_FT	401	CR	64	1	965	-1.0140							0.90
2010_G8Sc_FT	401	CR	65	1	965	0.5856							0.97
2010_G8Sc_FT	401	CR	66	1	965	0.7398							1.00
2010_G8Sc_FT	401	CR	67	1	965	0.5959							0.88
2010_G8Sc_FT	401	CR	68	1	965	0.3618							0.90
2010_G8Sc_FT	401	CR	69	1	965	0.5804							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
2010_G8Sc_FT	401	CR	70	1	965	1.2634							1.16
2010_G8Sc_FT	402	MC	01	1	975	0.1683							1.04
2010_G8Sc_FT	402	MC	02	1	975	-0.8781							0.97
2010_G8Sc_FT	402	MC	03	1	975	-2.1177							0.89
2010_G8Sc_FT	402	MC	04	1	975	0.1522							0.87
2010_G8Sc_FT	402	MC	05	1	975	-1.4362							0.82
2010_G8Sc_FT	402	MC	06	1	975	1.9961							1.01
2010_G8Sc_FT	402	MC	07	1	975	-0.4754							0.99
2010_G8Sc_FT	402	MC	08	1	975	0.4871							1.07
2010_G8Sc_FT	402	MC	09	1	975	-0.1305							1.07
2010_G8Sc_FT	402	MC	10	1	975	1.3701							1.26
2010_G8Sc_FT	402	MC	11	1	975	0.1415							1.21
2010_G8Sc_FT	402	MC	12	1	975	0.1199							1.31
2010_G8Sc_FT	402	MC	13	1	975	-0.5867							0.96
2010_G8Sc_FT	402	CR	61	1	975	-1.4159							0.90
2010_G8Sc_FT	402	CR	62	1	975	0.4151							0.99
2010_G8Sc_FT	402	CR	63	1	975	-0.5207							1.02
2010_G8Sc_FT	402	CR	64	1	975	0.0655							0.87
2010_G8Sc_FT	402	CR	65	1	975	0.2798							0.99
2010_G8Sc_FT	402	CR	66	1	975	0.6347							0.83
2010_G8Sc_FT	402	CR	67	1	975	-0.3073							0.90
2010_G8Sc_FT	402	CR	68	1	975	-0.6408							0.90
2010_G8Sc_FT	402	CR	69	1	975								

Table 9. Partial Credit Model Item Analysis (continued)

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2010_G8Sc_FT	402	CR	70	1	975								
2010_G8Sc_FT	403	MC	01	1	941	0.7742							0.99
2010_G8Sc_FT	403	MC	02	1	941	-1.2989							0.92
2010_G8Sc_FT	403	MC	03	1	941	-1.1066							1.04
2010_G8Sc_FT	403	MC	04	1									
2010_G8Sc_FT	403	MC	05	1	941	-0.9633							0.91
2010_G8Sc_FT	403	MC	06	1	941	1.3745							1.06
2010_G8Sc_FT	403	MC	07	1	941	0.6880							1.10
2010_G8Sc_FT	403	MC	08	1	941	0.5582							1.05
2010_G8Sc_FT	403	MC	09	1									
2010_G8Sc_FT	403	MC	10	1	941	-1.2234							0.98
2010_G8Sc_FT	403	CR	61	1	941	-0.0920							1.11
2010_G8Sc_FT	403	CR	62	1	941	1.2408							1.07
2010_G8Sc_FT	403	CR	63	1	941	0.1191							0.95
2010_G8Sc_FT	403	CR	64	1	941	0.1363							0.95
2010_G8Sc_FT	403	CR	65	1	941	-0.1770							1.03
2010_G8Sc_FT	403	CR	66	2	941	0.6621	-1.4327	1.4327					0.94
2010_G8Sc_FT	403	CR	67	1	941	-0.4579							0.93
2010_G8Sc_FT	403	CR	68	1	941	-1.0379							1.04
2010_G8Sc_FT	403	CR	69	1	941	-0.7163							0.94
2010_G8Sc_FT	403	CR	70	1	941	2.0060							0.92
2010_G8Sc_FT	403	CR	71	1									
2010_G8Sc_FT	404	MC	01	1	976	-0.4000							1.19

Table 9. Partial Credit Model Item Analysis (continued)

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2010_G8Sc_FT	404	MC	02	1	976	-0.0700							0.99
2010_G8Sc_FT	404	MC	03	1	976	0.0784							1.00
2010_G8Sc_FT	404	MC	04	1	976	-0.4700							1.07
2010_G8Sc_FT	404	MC	05	1	976	0.7100							1.12
2010_G8Sc_FT	404	MC	06	1	976	0.6600							1.04
2010_G8Sc_FT	404	MC	07	1	976	1.1600							1.21
2010_G8Sc_FT	404	MC	08	1	976	1.1100							1.05
2010_G8Sc_FT	404	MC	09	1	976	-1.1300							0.87
2010_G8Sc_FT	404	MC	10	1	976	-0.5300							0.89
2010_G8Sc_FT	404	MC	11	1	976	-0.3000							0.88
2010_G8Sc_FT	404	MC	12	1	976	-0.2900							1.00
2010_G8Sc_FT	404	MC	13	1	976	-0.2200							1.07
2010_G8Sc_FT	404	MC	14	1	976	-0.3000							0.87
2010_G8Sc_FT	404	MC	15	1	976	-0.7400							1.01
2010_G8Sc_FT	404	MC	16	1	976	0.0600							0.94
2010_G8Sc_FT	404	MC	17	1	976	-1.1200							1.06
2010_G8Sc_FT	404	MC	18	1	976	0.8400							1.06
2010_G8Sc_FT	404	MC	19	1	976	0.6500							1.03
2010_G8Sc_FT	404	MC	20	1	976	-0.3000							0.92
2010_G8Sc_FT	404	MC	21	1	976	-0.1700							0.80
2010_G8Sc_FT	404	MC	22	1	976	0.7800							0.98
2010_G8Sc_FT	404	MC	23	1	976	0.1700							0.99
2010_G8Sc_FT	404	MC	24	1	976	-1.4812							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
2010_G8Sc_FT	404	MC	25	1	976	-1.0200							0.87
2010_G8Sc_FT	404	MC	26	1	976	1.1500							1.11
2010_G8Sc_FT	404	MC	27	1	976	1.5100							1.19
2010_G8Sc_FT	404	MC	28	1	976	0.0000							0.90
2010_G8Sc_FT	404	MC	29	1	976	-0.5600							0.79
2010_G8Sc_FT	404	MC	30	1	976	0.2300							0.93
2010_G8Sc_FT	405	MC	01	1	988	-1.4665							0.99
2010_G8Sc_FT	405	MC	02	1	988	0.5939							1.01
2010_G8Sc_FT	405	MC	03	1	988	-1.2812							0.98
2010_G8Sc_FT	405	MC	04	1									
2010_G8Sc_FT	405	MC	05	1	988	1.0876							1.10
2010_G8Sc_FT	405	MC	06	1	988	0.4206							1.07
2010_G8Sc_FT	405	MC	07	1	988	-0.1519							0.96
2010_G8Sc_FT	405	MC	08	1	988	0.2066							1.11
2010_G8Sc_FT	405	MC	09	1	988	-1.5187							0.96
2010_G8Sc_FT	405	MC	10	1	988	-0.1808							1.15
2010_G8Sc_FT	405	CR	61	1	988	0.6343							0.94
2010_G8Sc_FT	405	CR	62	1	988	-0.7278							0.89
2010_G8Sc_FT	405	CR	63	1	988	-0.6173							1.05
2010_G8Sc_FT	405	CR	64	1	988	0.2805							0.88
2010_G8Sc_FT	405	CR	65	1	988	-0.2277							0.97
2010_G8Sc_FT	405	CR	66	1	988	1.5094							0.99
2010_G8Sc_FT	405	CR	67	1	988	1.6797							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
2010_G8Sc_FT	405	CR	68	2	988	0.7618	-0.9055	0.9055					0.96
2010_G8Sc_FT	405	CR	69	1	988	-0.0833							0.96
2010_G8Sc_FT	405	CR	70	1	988	-0.5971							0.90
2010_G8Sc_FT	405	CR	71	1									
2010_G8Sc_FT	406	MC	01	1	967	-0.6233							0.98
2010_G8Sc_FT	406	MC	02	1	967	-1.3223							0.94
2010_G8Sc_FT	406	MC	03	1	967	-1.3812							1.07
2010_G8Sc_FT	406	MC	04	1	967	-2.4008							0.90
2010_G8Sc_FT	406	MC	05	1	967	-0.5112							0.99
2010_G8Sc_FT	406	MC	06	1	967	1.1681							1.24
2010_G8Sc_FT	406	MC	07	1	967	-1.7009							0.91
2010_G8Sc_FT	406	MC	08	1	967	-0.3065							0.97
2010_G8Sc_FT	406	MC	09	1	967	0.2359							1.03
2010_G8Sc_FT	406	MC	10	1	967	-0.2064							0.88
2010_G8Sc_FT	406	MC	11	1	967	-0.0869							1.02
2010_G8Sc_FT	406	MC	12	1	967	0.4861							1.08
2010_G8Sc_FT	406	MC	13	1	967	0.1202							1.17
2010_G8Sc_FT	406	CR	61	1	967	0.4559							1.06
2010_G8Sc_FT	406	CR	62	1	967	-1.5606							0.98
2010_G8Sc_FT	406	CR	63	1	967	0.3544							0.98
2010_G8Sc_FT	406	CR	64	1	967	1.2852							0.91
2010_G8Sc_FT	406	CR	65	1	967	0.1836							0.97
2010_G8Sc_FT	406	CR	66	1	967	0.2098							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
2010_G8Sc_FT	406	CR	67	1	967	1.5210							1.05
2010_G8Sc_FT	406	CR	68	1	967	-0.2414							0.87
2010_G8Sc_FT	406	CR	69	1	967	1.0276							0.92
2010_G8Sc_FT	406	CR	70	1	967	0.6756							0.99
2010_G8Sc_FT	406	CR	71	1	967	1.9109							0.98
2010_G8Sc_FT	407	MC	01	1	966	-2.5857							0.84
2010_G8Sc_FT	407	MC	02	1	966	-0.9495							1.00
2010_G8Sc_FT	407	MC	03	1	966	0.7058							0.96
2010_G8Sc_FT	407	MC	04	1	966	-1.5016							0.84
2010_G8Sc_FT	407	MC	05	1	966	-1.2121							0.90
2010_G8Sc_FT	407	MC	06	1	966	-1.0843							0.91
2010_G8Sc_FT	407	MC	07	1	966	1.5704							1.01
2010_G8Sc_FT	407	MC	08	1	966	-1.3504							0.96
2010_G8Sc_FT	407	MC	09	1	966	-0.4781							0.92
2010_G8Sc_FT	407	MC	10	1	966	-0.8380							0.94
2010_G8Sc_FT	407	MC	11	1	966	0.0751							1.15
2010_G8Sc_FT	407	MC	12	1	966	-0.0203							1.20
2010_G8Sc_FT	407	MC	13	1	966	0.2554							1.00
2010_G8Sc_FT	407	CR	61	1	966	0.1891							1.05
2010_G8Sc_FT	407	CR	62	1	966	-0.2070							1.07
2010_G8Sc_FT	407	CR	63	3	966	0.9876	-1.2461	-0.7932	2.0393				0.85
2010_G8Sc_FT	407	CR	64	1	966	0.5933							0.96
2010_G8Sc_FT	407	CR	65	1	966	0.5246							1.39

Table 9. Partial Credit Model Item Analysis (continued)

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2010_G8Sc_FT	407	CR	66	1	966	0.7693							0.95
2010_G8Sc_FT	407	CR	67	1	966	0.7156							0.91
2010_G8Sc_FT	407	CR	68	1	966	1.7435							0.96
2010_G8Sc_FT	408	MC	01	1	971	0.4838							0.98
2010_G8Sc_FT	408	MC	02	1	971	1.2349							1.15
2010_G8Sc_FT	408	MC	03	1	971	0.0853							1.05
2010_G8Sc_FT	408	MC	04	1	971	-1.0169							1.08
2010_G8Sc_FT	408	MC	05	1	971	-1.1405							0.98
2010_G8Sc_FT	408	MC	06	1	971	-1.3202							0.92
2010_G8Sc_FT	408	MC	07	1	971	0.6642							0.90
2010_G8Sc_FT	408	MC	08	1	971	1.1851							1.24
2010_G8Sc_FT	408	MC	09	1	971	-1.5657							0.94
2010_G8Sc_FT	408	MC	10	1	971	0.4000							1.04
2010_G8Sc_FT	408	MC	11	1	971	-0.4183							1.02
2010_G8Sc_FT	408	MC	12	1	971	-0.1405							1.11
2010_G8Sc_FT	408	MC	13	1	971	1.3712							1.16
2010_G8Sc_FT	408	CR	61	1	971	1.0571							0.87
2010_G8Sc_FT	408	CR	62	1									
2010_G8Sc_FT	408	CR	63	1	971	0.7707							0.89
2010_G8Sc_FT	408	CR	64	1	971	0.5670							0.83
2010_G8Sc_FT	408	CR	65	1	971	-0.8056							0.88
2010_G8Sc_FT	408	CR	66	1	971	-1.5220							0.97
2010_G8Sc_FT	408	CR	67	1	971	1.8347							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
2010_G8Sc_FT	408	CR	68	1	971	-0.0529							0.97
2010_G8Sc_FT	408	CR	69	1	971	-1.4587							0.93
2010_G8Sc_FT	408	CR	70	1	971	-0.3287							0.96
2010_G8Sc_FT	409	MC	01	1	968	-1.8725							0.86
2010_G8Sc_FT	409	MC	02	1	968	-1.0176							0.88
2010_G8Sc_FT	409	MC	03	1	968	-0.9855							0.98
2010_G8Sc_FT	409	MC	04	1	968	-0.5733							1.11
2010_G8Sc_FT	409	MC	05	1	968	0.9594							1.00
2010_G8Sc_FT	409	MC	06	1	968	-0.0805							1.03
2010_G8Sc_FT	409	MC	07	1	968	-2.7678							0.94
2010_G8Sc_FT	409	MC	08	1	968	0.4742							1.00
2010_G8Sc_FT	409	MC	09	1	968	1.1837							1.01
2010_G8Sc_FT	409	MC	10	1	968	-0.2304							1.04
2010_G8Sc_FT	409	MC	11	1	968	0.7377							1.25
2010_G8Sc_FT	409	MC	12	1	968	-0.3689							1.04
2010_G8Sc_FT	409	MC	13	1	968	-1.1086							1.02
2010_G8Sc_FT	409	CR	61	1	968	1.0100							1.01
2010_G8Sc_FT	409	CR	62	1	968	1.1580							0.92
2010_G8Sc_FT	409	CR	63	1	968	-1.5804							0.99
2010_G8Sc_FT	409	CR	64	1	968	-0.8545							0.91
2010_G8Sc_FT	409	CR	65	1	968	0.6570							0.85
2010_G8Sc_FT	409	CR	66	1	968	1.3084							0.92

Table 9. Partial Credit Model Item Analysis (continued)

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2010_G8Sc_FT	409	CR	67	1	968	-0.7809							0.89
2010_G8Sc_FT	409	CR	68	1	968	2.3925							1.07
2010_G8Sc_FT	409	CR	69	1	968	1.5167							1.00
2010_G8Sc_FT	409	CR	70	1	968	1.4142							1.07

Appendix C: DIF Statistics

Table 10. DIF Statistics

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
401	1	MC	-0.25	0.49	-0.05		
401	2	MC	-0.20	0.32	-0.03		
401	3	MC	0.26	0.45	0.04		
401	4	MC	0.40	1.12	0.05		
401	5	MC	0.38	1.27	0.08		
401	6	MC	0.54	1.95	0.08		
401	7	MC	-0.02	0.00	0.01		
401	8	MC	-0.11	0.09	-0.02		
401	9	MC	0.17	0.11	0.01		
401	10	MC	0.20	0.28	0.02		
401	11	MC	-0.87	5.92	-0.15		
401	12	MC	0.00	0.00	0.00		
401	13	MC	-0.22	0.34	-0.04		
401	61	OE		1.26	0.09		
401	62	OE		0.18	0.03		
401	63	OE		0.00	0.01		
401	64	OE		0.78	-0.05		
401	65	OE		0.21	-0.02		
401	66	OE		5.01	0.13		
401	67	OE		7.81	-0.16		
401	68	OE		0.10	-0.01		
401	69	OE		0.02	-0.01		

Table 10. DIF Statistics (continued)

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
401	70	OE		0.02	0.01		
402	1	MC	0.26	0.52	0.05		
402	2	MC	-0.25	0.34	-0.05		
402	3	MC	-0.48	0.42	-0.03		
402	4	MC	0.31	0.61	0.06		
402	5	MC	0.60	1.05	0.05		
402	6	MC	-0.30	0.60	-0.06		
402	7	MC	0.09	0.05	0.01		
402	8	MC	0.20	0.36	0.05		
402	9	MC	0.46	1.60	0.09		
402	10	MC	-0.18	0.28	-0.02		
402	11	MC	0.41	1.45	0.08		
402	12	MC	-0.26	0.60	-0.06		
402	13	MC	0.38	0.80	0.05		
402	61	OE		0.02	0.00		
402	62	OE		0.49	0.05		
402	63	OE		4.75	0.13		
402	64	OE		0.82	-0.05		
402	65	OE		8.03	0.18	BB	F
402	66	OE		10.81	-0.18	BB	M
402	67	OE		3.63	-0.11		
402	68	OE		0.11	-0.02		

Table 10. DIF Statistics (continued)

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
402	69	OE					
402	70	OE					
403	1	MC	-0.95	6.82	-0.17		
403	2	MC	-0.68	1.63	-0.07		
403	3	MC	0.34	0.54	0.06		
403	4	MC					
403	5	MC	-1.15	5.70	-0.17	B	M
403	6	MC	-1.20	10.94	-0.21	B	M
403	7	MC	-0.63	3.15	-0.10		
403	8	MC	-0.36	0.99	-0.05		
403	9	MC					
403	10	MC	-0.89	3.20	-0.11		
403	61	OE		2.52	0.10		
403	62	OE		1.20	0.07		
403	63	OE		2.19	0.10		
403	64	OE		0.97	-0.07		
403	65	OE		1.03	0.05		
403	66	OE		0.13	-0.02		
403	67	OE		2.22	0.09		
403	68	OE		12.16	0.23	BB	F
403	69	OE		10.73	0.19	BB	F
403	70	OE		5.05	0.11		
403	71	OE					

Table 10. DIF Statistics (continued)

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
404	1	MC	0.03	0.01	-0.01		
404	2	MC	-0.21	0.33	-0.01		
404	3	MC	0.70	3.62	0.11		
404	4	MC	0.62	2.62	0.12		
404	5	MC	-1.42	17.76	-0.27	B	M
404	6	MC	-0.38	1.29	-0.06		
404	7	MC	0.68	4.27	0.12		
404	8	MC	-0.41	1.42	-0.07		
404	9	MC	-0.01	0.00	0.00		
404	10	MC	0.40	0.83	0.05		
404	11	MC	0.02	0.00	0.00		
404	12	MC	-1.27	10.39	-0.20	B	M
404	13	MC	-0.11	0.09	-0.02		
404	14	MC	-0.29	0.49	-0.07		
404	15	MC	0.68	2.73	0.09		
404	16	MC	-0.38	1.04	-0.06		
404	17	MC	-0.40	0.79	-0.07		
404	18	MC	-0.01	0.00	-0.01		
404	19	MC	0.42	1.53	0.08		
404	20	MC	-0.01	0.00	0.00		
404	21	MC	-0.79	3.39	-0.10		
404	22	MC	0.78	4.94	0.14		
404	23	MC	0.09	0.07	0.03		

Table 10. DIF Statistics (continued)

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
404	24	MC	0.52	0.83	0.07		
404	25	MC	0.35	0.56	0.06		
404	26	MC	-0.63	3.50	-0.12		
404	27	MC	0.71	4.26	0.14		
404	28	MC	-0.69	3.39	-0.12		
404	29	MC	0.65	2.11	0.08		
404	30	MC	0.86	5.30	0.16		
405	1	MC	1.15	4.88	0.13	B	F
405	2	MC	0.48	1.80	0.07		
405	3	MC	0.52	1.14	0.06		
405	4	MC					
405	5	MC	-0.42	1.60	-0.09		
405	6	MC	-0.60	2.92	-0.09		
405	7	MC	0.97	6.36	0.15		
405	8	MC	0.23	0.42	0.04		
405	9	MC	0.33	0.38	0.03		
405	10	MC	-0.78	4.53	-0.13		
405	61	OE		0.08	-0.01		
405	62	OE		1.47	0.05		
405	63	OE		0.12	-0.02		
405	64	OE		0.20	0.03		
405	65	OE		0.02	0.00		
405	66	OE		0.09	-0.01		

Table 10. DIF Statistics (continued)

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
405	67	OE		0.00	0.00		
405	68	OE		2.10	-0.07		
405	69	OE		2.29	0.09		
405	70	OE		0.64	0.05		
405	71	OE					
406	1	MC	-0.81	3.99	-0.14		
406	2	MC	-0.01	0.00	-0.01		
406	3	MC	0.80	2.68	0.11		
406	4	MC	0.76	1.02	0.07		
406	5	MC	-0.55	1.91	-0.06		
406	6	MC	-0.22	0.47	-0.06		
406	7	MC	-0.10	0.03	-0.01		
406	8	MC	-1.31	11.12	-0.19	B	M
406	9	MC	-0.61	2.96	-0.12		
406	10	MC	0.17	0.17	0.02		
406	11	MC	-0.64	3.02	-0.11		
406	12	MC	-1.02	9.26	-0.20	B	M
406	13	MC	-1.02	8.93	-0.19	B	M
406	61	OE		0.22	0.02		
406	62	OE		3.00	0.11		
406	63	OE		0.17	0.03		
406	64	OE		7.51	0.15		
406	65	OE		4.38	0.14		

Table 10. DIF Statistics (continued)

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
406	66	OE		8.26	0.18	BB	F
406	67	OE		3.81	0.11		
406	68	OE		2.63	-0.10		
406	69	OE		0.45	0.02		
406	70	OE		5.77	0.17		
406	71	OE		2.62	0.11		
407	1	MC	0.39	0.17	0.02		
407	2	MC	0.16	0.12	0.02		
407	3	MC	-0.36	1.05	-0.07		
407	4	MC	2.08	11.03	0.18	C	F
407	5	MC	0.52	1.06	0.06		
407	6	MC	-0.80	2.79	-0.10		
407	7	MC	0.56	2.30	0.09		
407	8	MC	-1.05	4.29	-0.12	B	M
407	9	MC	-0.22	0.28	-0.02		
407	10	MC	0.05	0.01	0.02		
407	11	MC	0.32	0.90	0.06		
407	12	MC	0.96	8.06	0.19		
407	13	MC	-0.69	3.90	-0.12		
407	61	OE		3.04	0.11		
407	62	OE		2.04	-0.08		
407	63	OE		0.18	0.01		
407	64	OE		0.89	-0.05		

Table 10. DIF Statistics (continued)

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
407	65	OE		5.23	0.14		
407	66	OE		0.00	0.00		
407	67	OE		2.85	-0.10		
407	68	OE		13.63	-0.21	BB	M
408	1	MC	0.06	0.03	0.01		
408	2	MC	0.62	3.49	0.13		
408	3	MC	0.08	0.06	0.02		
408	4	MC	0.56	1.65	0.10		
408	5	MC	-0.10	0.04	-0.02		
408	6	MC	1.16	4.89	0.14	B	F
408	7	MC	0.03	0.01	0.00		
408	8	MC	-0.06	0.03	-0.02		
408	9	MC	-0.82	2.15	-0.09		
408	10	MC	0.26	0.60	0.03		
408	11	MC	0.60	2.38	0.11		
408	12	MC	0.59	2.68	0.12		
408	13	MC	-1.37	16.43	-0.25	B	M
408	61	OE		5.52	-0.12		
408	62	OE					
408	63	OE		5.43	0.14		
408	64	OE		10.97	-0.18	BB	M
408	65	OE		3.73	-0.13		
408	66	OE		0.91	-0.06		

Table 10. DIF Statistics (continued)

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
408	67	OE		0.47	-0.04		
408	68	OE		1.83	-0.10		
408	69	OE		0.19	0.00		
408	70	OE		1.69	0.08		
409	1	MC	0.54	0.67	0.05		
409	2	MC	0.22	0.21	0.03		
409	3	MC	1.17	6.73	0.16	B	F
409	4	MC	-0.42	1.20	-0.08		
409	5	MC	-0.74	4.38	-0.11		
409	6	MC	0.08	0.05	0.02		
409	7	MC	0.17	0.04	0.00		
409	8	MC	0.10	0.08	0.03		
409	9	MC	-0.30	0.75	-0.04		
409	10	MC	-0.12	0.10	-0.01		

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

Table 10. DIF Statistics (continued)

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
409	11	MC	-0.18	0.30	-0.04		
409	12	MC	-0.33	0.73	-0.04		
409	13	MC	-1.11	5.77	-0.14	B	M
409	61	OE		4.13	-0.13		
409	62	OE		8.80	0.17		
409	63	OE		4.00	-0.13		
409	64	OE		2.11	0.07		
409	65	OE		2.20	0.08		
409	66	OE		0.08	-0.02		
409	67	OE		2.83	0.11		
409	68	OE		0.27	-0.04		
409	69	OE		1.00	-0.07		
409	70	OE		6.24	0.13		

Appendix D: Operational Test Map

Table 11. Operational Test Map for May 2010

Position	Item Type	Max Points	Weight	Strand	Mean	Point-Biserial	Rasch	S1	S2	S3	S4
1	MC	1	1	1	0.78	0.33	-0.78				
2	MC	1	1	1	0.84	0.37	-1.19				
3	MC	1	1	1	0.82	0.44	-0.96				
4	MC	1	1	4	0.78	0.43	-0.74				
5	MC	1	1	6	0.64	0.35	0.07				
6	MC	1	1	4	0.60	0.41	0.30				
7	MC	1	1	4	0.69	0.45	-0.28				
8	MC	1	1	4	0.58	0.47	0.35				
9	MC	1	1	4	0.65	0.30	0.00				
10	MC	1	1	4	0.79	0.50	-0.74				
11	MC	1	1	4	0.66	0.52	0.01				
12	MC	1	1	4	0.60	0.47	0.29				
13	MC	1	1	4	0.80	0.25	-0.82				
14	MC	1	1	4	0.77	0.47	-0.62				
15	MC	1	1	4	0.80	0.42	-0.89				
16	MC	1	1	4	0.80	0.41	-0.87				
17	MC	1	1	4	0.64	0.47	0.06				
18	MC	1	1	4	0.59	0.44	0.32				
19	MC	1	1	4	0.68	0.34	-0.19				
20	MC	1	1	4	0.62	0.53	0.13				
21	MC	1	1	4	0.68	0.42	-0.17				

Table 11. Operational Test Map for May 2010 (continued)

Position	Item Type	Max Points	Weight	Strand	Mean	Point-Biserial	Rasch	S1	S2	S3	S4
22	MC	1	1	4	0.75	0.49	-0.54				
23	MC	1	1	4	0.47	0.43	0.90				
24	MC	1	1	4	0.86	0.39	-1.35				
25	MC	1	1	4	0.84	0.35	-1.24				
26	MC	1	1	4	0.79	0.45	-0.79				
27	MC	1	1	4	0.69	0.40	-0.25				
28	MC	1	1	4	0.59	0.18	0.21				
29	MC	1	1	4	0.53	0.49	0.57				
30	MC	1	1	4	0.68	0.31	-0.18				
31	MC	1	1	4	0.62	0.41	0.08				
32	MC	1	1	4	0.50	0.39	0.72				
33	MC	1	1	4	0.56	0.35	0.47				
34	MC	1	1	4	0.71	0.48	-0.24				
35	MC	1	1	4	0.77	0.45	-0.71				
36	MC	1	1	4	0.56	0.32	0.44				
37	MC	1	1	4	0.43	0.42	1.09				
38	MC	1	1	4	0.49	0.43	0.79				
39	MC	1	1	4	0.63	0.45	0.12				
40	MC	1	1	4	0.83	0.43	-1.13				
41	MC	1	1	4	0.63	0.51	0.08				
42	MC	1	1	4	0.75	0.38	-0.53				
43	MC	1	1	4	0.66	0.35	-0.02				

Table 11. Operational Test Map for May 2010 (continued)

Position	Item Type	Max Points	Weight	Strand	Mean	Point-Biserial	Rasch	S1	S2	S3	S4
44	MC	1	1	4	0.75	0.48	-0.60				
45	MC	1	1	4	0.45	0.41	0.98				
46	CR	2	1	1	0.54	0.49	1.76	0.40	-0.40		
47	CR	1	1	1	0.51	0.45	0.76				
48	CR	1	1	1	0.48	0.47	0.84				
49	CR	2	1	1	1.25	0.60	0.27	0.40	-0.40		
50	CR	1	1	1	0.47	0.49	0.89				
51	CR	1	1	1	0.51	0.45	0.71				
52	CR	1	1	1	0.61	0.50	0.19				
53	CR	1	1	4	0.62	0.56	0.19				
54	CR	1	1	6	0.67	0.48	-0.05				
55	CR	2	1	1	0.96	0.46	0.85	-0.98	0.98		
56	CR	2	1	4	0.91	0.54	1.00	-1.21	1.21		
57	CR	1	1	4	0.80	0.45	-0.92				
58	CR	1	1	4	0.49	0.55	0.79				
59	CR	1	1	4	0.45	0.57	0.98				
60	CR	1	1	4	0.78	0.40	-0.77				
61	CR	1	1	4	0.42	0.58	1.06				
62	CR	1	1	4	0.35	0.46	1.40				
63	CR	1	1	4	0.93	0.32	-2.15				
64	CR	1	1	4	0.72	0.49	-0.39				
65	CR	1	1	4	0.74	0.48	-0.45				

Table 11. Operational Test Map for May 2010 (continued)

Position	Item Type	Max Points	Weight	Strand	Mean	Point-Biserial	Rasch	S1	S2	S3	S4
66	CR	1	1	4	0.41	0.45	1.16				
67	CR	1	1	4	0.41	0.53	1.20				
68	CR	1	1	4	0.71	0.37	-0.24				
69	CR	1	1	4	0.30	0.37	1.75				
70	CR	1	1	4	0.94	0.26	-2.28				
71	CR	1	1	4	0.60	0.39	0.27				
72	CR	1	1	4	0.65	0.54	0.07				
73	CR	1	1	4	0.64	0.39	0.13				
74	CR	1	1	4	0.80	0.30	-0.84				
75	CR	1	1	4	0.54	0.43	0.60				
76	CR	1	1	4	0.56	0.48	0.46				
77	CR	1	1	4	0.40	0.53	1.23				
78	CR	1	1	4	0.41	0.51	1.20				
79	CR	1	1	4	0.54	0.41	0.46				
80	CR	1	1	4	0.79	0.38	-0.87				
81	CR	1	1	4	0.47	0.49	0.86				

Appendix E: Scoring Table

Table 12. Scoring Table for May 2010

Raw Score	Ability	Scale Score	Raw Score	Ability	Scale Score	Raw Score	Ability	Scale Score	Raw Score	Ability	Scale Score
0	-5.410	0.500	22	-1.096	32.943	44	0.190	59.421	66	1.517	79.917
1	-4.691	2.343	23	-1.028	34.540	45	0.244	60.368	67	1.593	81.595
2	-3.972	4.274	24	-0.961	35.665	46	0.298	61.316	68	1.671	82.720
3	-3.541	5.447	25	-0.895	37.215	47	0.353	62.274	69	1.753	83.644
4	-3.228	7.193	26	-0.831	38.688	48	0.407	63.201	70	1.838	84.563
5	-2.981	8.618	27	-0.769	39.706	49	0.462	64.135	71	1.927	85.484
6	-2.774	10.505	28	-0.707	40.746	50	0.518	65.085	72	2.021	86.409
7	-2.596	11.925	29	-0.647	41.750	51	0.573	66.009	73	2.121	87.338
8	-2.438	13.365	30	-0.587	43.559	52	0.630	66.951	74	2.227	88.266
9	-2.297	14.977	31	-0.529	44.776	53	0.686	68.036	75	2.342	89.207
10	-2.168	16.303	32	-0.471	45.776	54	0.744	69.467	76	2.466	90.146
11	-2.049	17.944	33	-0.414	46.842	55	0.802	70.738	77	2.602	91.088
12	-1.938	19.509	34	-0.357	48.402	56	0.861	71.670	78	2.753	92.035
13	-1.835	20.732	35	-0.301	49.739	57	0.921	72.598	79	2.925	92.989
14	-1.737	22.442	36	-0.245	50.730	58	0.981	73.507	80	3.125	93.954
15	-1.645	23.888	37	-0.190	51.703	59	1.043	74.431	81	3.365	94.924
16	-1.557	24.993	38	-0.135	52.678	60	1.106	75.361	82	3.670	95.911
17	-1.473	26.357	39	-0.081	53.642	61	1.170	76.282	83	4.093	96.944
18	-1.392	28.057	40	-0.027	54.624	62	1.236	77.205	84	4.803	98.855
19	-1.314	29.274	41	0.028	56.143	63	1.303	78.117	85	5.513	99.938
20	-1.239	30.733	42	0.082	57.526	64	1.373	79.000			
21	-1.167	32.023	43	0.136	58.474	65	1.444	79.035			