New York State Examination in Grade 4 Elementary-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

Item Difficulty *3*

Point-Biserial Correlation *3*

Test Reliability *4*

Scoring Reliability *5*

Inter-rater Agreement *6*

Constructed-Response Item Means and Standard Deviations *9*

Intra-class Correlation *9*

Weighted Kappa *9*

Item Response Theory (IRT) Statistics 10

Item Calibration *10*

Item Fit Evaluation *11*

Differential Item Functioning (DIF) Statistics 12

Section III: Equating Procedure 13

Section IV: Scaling of Operational Test Forms 15

References 17

Appendix A: Classical Item Analysis 18

Appendix B: Partial Credit Model Item Analysis 25

Appendix C: DIF Statistics 32

Appendix D: Operational Test Map 36

Appendix E: Scoring Table 39

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 8
Table 6. Partial Credit Model Item Analysis 12
Table 7. Initial Mean Abilities and Equating Constants 14
Table 8. Classical Item Analysis 19
Table 9. Partial Credit Model Item Analysis 26
Table 10. DIF Statistics 33
Table 11. Operational Test Map for June 2010 37
Table 12. Scoring Table for June 2010 40

Section I: Introduction

Purpose

The purpose of this report is to document the psychometric work on the New York State Examination in Grade 4 Elementary-Level Science in 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 4 Elementary-Level Science.

Section II: Field Test Analysis

In May 2010, field testing was conducted for the New York State Examination in Grade 4 Elementary-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 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 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 9,650 student records. After the exclusion rules were applied, the resulting field test data file contained 9,576 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 four, 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

			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
411	1,077	25	1	24	0	0	20	5
412	1,071	13	1	11	1	0	7	6
413	1,059	12	1	11	0	0	9	3
414	1,062	12	3	8	1	0	10	2
415	1,058	13	1	8	1	0	8	2
416	1,070	13	3	10	0	0	10	3
417	1,059	12	4	8	0	1	8	3
418	1,071	12	0	11	1	0	7	5
419	1,049	13	5	7	0	0	10	2

* 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 0.56 to 0.82 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
411	0.82	n/a
412	0.68	0.96
413	0.69	0.89
414	0.61	0.90
415	0.56	0.76
416	0.69	0.99
417	0.62	0.96
418	0.67	0.99
419	0.57	0.98

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 0.76 to 0.99, 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**	**-2**	**-1**	**0**	**1**	**2**
412	41	1	0.00	0.00	0.99	0.01	0.00
412	42	1	0.00	0.01	0.97	0.02	0.00
412	43	1	0.00	0.02	0.98	0.00	0.00
412	44	1	0.00	0.01	0.99	0.00	0.00
413	41	2	0.00	0.00	0.99	0.01	0.00
413	42	1	0.00	0.01	0.96	0.03	0.00
413	43	2	0.00	0.08	0.83	0.09	0.01
413	44	1	0.00	0.01	0.95	0.04	0.00
413	45	1	0.00	0.08	0.86	0.07	0.00
414	41	1	0.00	0.00	0.99	0.01	0.00
414	42	1	0.00	0.00	0.99	0.01	0.00
414	43	1	0.00	0.12	0.77	0.11	0.00
414	44	1	0.00	0.02	0.94	0.04	0.00
414	45	2	0.00	0.02	0.96	0.02	0.00
415	41	1	0.00	0.00	0.97	0.03	0.00
416	41	2	0.00	0.00	0.99	0.01	0.00
416	42	1	0.00	0.02	0.97	0.01	0.00
416	43	1	0.00	0.00	1.00	0.00	0.00
416	44	1	0.00	0.00	1.00	0.00	0.00
416	45	1	0.00	0.00	1.00	0.00	0.00
417	41	1	0.00	0.00	0.98	0.02	0.00
417	42	1	0.00	0.02	0.97	0.01	0.00
417	43	2	0.00	0.02	0.94	0.04	0.00
417	44	1	0.00	0.02	0.97	0.01	0.00
418	41	1	0.00	0.00	1.00	0.00	0.00
418	42	1	0.00	0.00	1.00	0.00	0.00
418	43	1	0.00	0.00	0.99	0.01	0.00
419	41	1	0.00	0.00	1.00	0.00	0.00
419	43	1	0.00	0.03	0.97	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 Corre-lation	Wt Kappa
412	41	1	104	99.0	1.0	100.0	0.6	0.6	0.50	0.50	0.98	0.98
412	42	1	103	97.1	2.9	100.0	0.8	0.8	0.39	0.40	0.91	0.91
412	43	1	102	98.0	2.0	100.0	0.7	0.7	0.47	0.46	0.96	0.95
412	44	1	104	99.0	1.0	100.0	0.6	0.6	0.49	0.49	0.98	0.98
413	41	2	101	99.0	1.0	100.0	1.3	1.3	0.63	0.63	0.99	0.98
413	42	1	104	96.2	3.8	100.0	0.6	0.6	0.49	0.49	0.92	0.92
413	43	2	103	82.5	16.5	99.0	1.0	1.0	0.77	0.79	0.83	0.78
413	44	1	104	95.2	4.8	100.0	0.8	0.7	0.42	0.44	0.87	0.87
413	45	1	104	85.6	14.4	100.0	0.6	0.6	0.49	0.49	0.70	0.70
414	41	1	95	98.9	1.1	100.0	0.5	0.5	0.50	0.50	0.98	0.98
414	42	1	94	98.9	1.1	100.0	0.2	0.1	0.37	0.36	0.96	0.96
414	43	1	93	77.4	22.6	100.0	0.5	0.5	0.50	0.50	0.55	0.55
414	44	1	93	93.5	6.5	100.0	0.5	0.4	0.50	0.50	0.87	0.87
414	45	2	92	95.7	4.3	100.0	0.9	0.9	0.81	0.81	0.97	0.95
415	41	1	95	96.8	3.2	100.0	0.7	0.7	0.47	0.48	0.93	0.93
416	41	2	97	99.0	1.0	100.0	1.1	1.1	0.64	0.64	0.99	0.98
416	42	1	104	97.1	2.9	100.0	0.6	0.6	0.48	0.48	0.94	0.94
416	43	1	105	100.0	0.0	100.0	0.6	0.6	0.50	0.50	1.00	1.00
416	44	1	105	100.0	0.0	100.0	0.6	0.6	0.49	0.49	1.00	1.00
416	45	1	105	100.0	0.0	100.0	0.9	0.9	0.34	0.34	1.00	1.00
417	41	1	102	98.0	2.0	100.0	0.7	0.7	0.46	0.47	0.96	0.95
417	42	1	98	96.9	3.1	100.0	0.4	0.4	0.48	0.49	0.93	0.93
417	43	2	100	94.0	6.0	100.0	1.0	1.0	0.93	0.94	0.97	0.94
417	44	1	101	97.0	3.0	100.0	0.8	0.8	0.43	0.42	0.92	0.92
418	41	1	99	100.0	0.0	100.0	0.8	0.8	0.42	0.42	1.00	1.00
418	42	1	98	100.0	0.0	100.0	0.8	0.8	0.38	0.38	1.00	1.00
418	43	1	96	99.0	1.0	100.0	0.8	0.8	0.40	0.41	0.97	0.97
419	41	1	104	100.0	0.0	100.0	0.5	0.5	0.50	0.50	1.00	1.00
419	43	1	103	97.1	2.9	100.0	0.6	0.6	0.50	0.49	0.94	0.94

* Adj. = difference of one Deviation

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 one item had intra-class 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. The Weighted Kappa statistics ranged from 0.55 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 one item 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**
411	1,077	25	0	24	1	0	25	0
412	1,071	13	1	12	0	0	13	0
413	1,059	12	0	12	0	0	12	0
414	1,062	12	0	11	1	0	12	0
415	1,058	13	0	9	1	0	10	0
416	1,070	13	0	12	1	0	13	0
417	1,059	12	0	11	1	0	11	1
418	1,071	12	1	11	0	0	12	0
419	1,049	13	0	12	0	0	12	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 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| ≥ 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 2010 field test administration for the New York State Examination in Grade 4 Elementary-Level Science consisted of 8 field test forms numbered 412–419 and one anchor form labeled 411. 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 four 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 411 was 1.42. 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 411). 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
412	0.96	0.42
413	0.89	0.49
414	0.82	0.55
415	1.21	0.19
416	0.60	0.76
417	0.62	0.74
418	1.20	0.20
419	0.58	0.77

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, content accuracy, 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 June 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 0.118 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 June 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 target was determined by the New York State Education Department to be May 2005. 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 2005 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), 2979 – 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_G4Sc_FT	411	MC	01	1	1077	0.82	0.00		0.05	0.09	0.84	0.03			0.84	0.35
2010_G4Sc_FT	411	MC	02	1	1077	0.82	0.00		0.09	0.85	0.02	0.04			0.85	0.42
2010_G4Sc_FT	411	MC	03	1	1077	0.82	0.00		0.14	0.61	0.08	0.17			0.61	0.47
2010_G4Sc_FT	411	MC	04	1	1077	0.82	0.00		0.23	0.58	0.09	0.09			0.58	0.39
2010_G4Sc_FT	411	MC	05	1	1077	0.82	0.00		0.02	0.60	0.28	0.09			0.60	0.34
2010_G4Sc_FT	411	MC	06	1	1077	0.82	0.00		0.03	0.03	0.84	0.10			0.84	0.46
2010_G4Sc_FT	411	MC	07	1	1077	0.82	0.00		0.39	0.13	0.39	0.08			0.39	0.30
2010_G4Sc_FT	411	MC	08	1	1077	0.82	0.00		0.28	0.69	0.02	0.01			0.69	0.32
2010_G4Sc_FT	411	MC	09	1	1077	0.82	0.00		0.04	0.85	0.02	0.08			0.85	0.35
2010_G4Sc_FT	411	MC	10	1	1077	0.82	0.00		0.88	0.04	0.03	0.04			0.88	0.46
2010_G4Sc_FT	411	MC	11	1	1077	0.82	0.00		0.21	0.04	0.68	0.06			0.68	0.56
2010_G4Sc_FT	411	MC	12	1	1077	0.82	0.00		0.81	0.12	0.05	0.02			0.81	0.38
2010_G4Sc_FT	411	MC	13	1	1077	0.82	0.01		0.07	0.21	0.09	0.62			0.62	0.45
2010_G4Sc_FT	411	MC	14	1	1077	0.82	0.01		0.57	0.12	0.21	0.09			0.57	0.52
2010_G4Sc_FT	411	MC	15	1	1077	0.82	0.00		0.86	0.03	0.05	0.05			0.86	0.48
2010_G4Sc_FT	411	MC	16	1	1077	0.82	0.01		0.08	0.05	0.09	0.78			0.78	0.53
2010_G4Sc_FT	411	MC	17	1	1077	0.82	0.01		0.04	0.02	0.90	0.03			0.90	0.49
2010_G4Sc_FT	411	MC	18	1	1077	0.82	0.01		0.80	0.06	0.04	0.09			0.80	0.49
2010_G4Sc_FT	411	MC	19	1	1077	0.82	0.01		0.11	0.75	0.05	0.08			0.75	0.51
2010_G4Sc_FT	411	MC	20	1	1077	0.82	0.01		0.01	0.09	0.78	0.10			0.78	0.40
2010_G4Sc_FT	411	MC	21	1	1077	0.82	0.01		0.14	0.03	0.06	0.76			0.76	0.44
2010_G4Sc_FT	411	MC	22	1	1077	0.82	0.01		0.03	0.89	0.02	0.05			0.89	0.48

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_G4Sc_FT	411	MC	23	1	1077	0.82	0.01		0.07	0.05	0.04	0.82			0.82	0.53
2010_G4Sc_FT	411	MC	24	1	1077	0.82	0.02		0.15	0.59	0.10	0.13			0.59	0.47
2010_G4Sc_FT	411	MC	25	1	1077	0.82	0.04		0.03	0.01	0.10	0.81			0.81	0.48
2010_G4Sc_FT	412	MC	01	1	1071	0.68	0.01		0.07	0.08	0.72	0.12			0.72	0.55
2010_G4Sc_FT	412	MC	02	1	1071	0.68	0.00		0.13	0.78	0.07	0.02			0.78	0.53
2010_G4Sc_FT	412	MC	03	1	1071	0.68	0.01		0.01	0.03	0.67	0.28			0.67	0.43
2010_G4Sc_FT	412	MC	04	1	1071	0.68	0.01		0.68	0.04	0.17	0.10			0.68	0.52
2010_G4Sc_FT	412	MC	05	1	1071	0.68	0.00		0.04	0.21	0.30	0.44			0.44	0.49
2010_G4Sc_FT	412	MC	06	1	1071	0.68	0.01		0.33	0.61	0.04	0.02			0.61	0.53
2010_G4Sc_FT	412	MC	07	1	1071	0.68	0.01		0.01	0.01	0.01	0.96			0.96	0.28
2010_G4Sc_FT	412	MC	08	1	1071	0.68	0.01		0.04	0.08	0.82	0.03			0.82	0.53
2010_G4Sc_FT	412	MC	09	1	1071	0.68	0.02		0.06	0.23	0.08	0.61			0.61	0.37
2010_G4Sc_FT	412	CR	41	1	1071	0.68	0.01	0.45	0.54						0.54	0.37
2010_G4Sc_FT	412	CR	42	1	1071	0.68	0.01	0.20	0.79						0.79	0.37
2010_G4Sc_FT	412	CR	43	1	1071	0.68	0.02	0.39	0.60						0.60	0.59
2010_G4Sc_FT	412	CR	44	1	1071	0.68	0.01	0.38	0.61						0.61	0.39
2010_G4Sc_FT	413	MC	01	1	1059	0.69	0.00		0.45	0.07	0.11	0.37			0.45	0.32
2010_G4Sc_FT	413	MC	02	1	1059	0.69	0.00		0.15	0.10	0.03	0.72			0.72	0.47
2010_G4Sc_FT	413	MC	03	1	1059	0.69	0.00		0.02	0.06	0.04	0.87			0.87	0.34
2010_G4Sc_FT	413	MC	04	1	1059	0.69	0.00		0.06	0.07	0.15	0.71			0.71	0.49
2010_G4Sc_FT	413	MC	05	1	1059	0.69	0.00		0.14	0.08	0.07	0.70			0.70	0.45
2010_G4Sc_FT	413	MC	06	1	1059	0.69	0.01		0.03	0.68	0.07	0.22			0.68	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_G4Sc_FT	413	MC	07	1	1059	0.69	0.02		0.03	0.10	0.82	0.03			0.82	0.46
2010_G4Sc_FT	413	CR	41	2	1059	0.69	0.02	0.12	0.46	0.40					1.25	0.64
2010_G4Sc_FT	413	CR	42	1	1059	0.69	0.01	0.36	0.63						0.63	0.51
2010_G4Sc_FT	413	CR	43	2	1059	0.69	0.02	0.30	0.31	0.37					1.04	0.67
2010_G4Sc_FT	413	CR	44	1	1059	0.69	0.02	0.21	0.77						0.77	0.50
2010_G4Sc_FT	413	CR	45	1	1059	0.69	0.02	0.42	0.56						0.56	0.41
2010_G4Sc_FT	414	MC	01	1	1062	0.61	0.00		0.13	0.67	0.08	0.11			0.67	0.34
2010_G4Sc_FT	414	MC	02	1	1062	0.61	0.00		0.03	0.87	0.07	0.03			0.87	0.45
2010_G4Sc_FT	414	MC	03	1	1062	0.61	0.00		0.00	0.02	0.93	0.04			0.93	0.30
2010_G4Sc_FT	414	MC	04	1	1062	0.61	0.00		0.88	0.07	0.01	0.03			0.88	0.34
2010_G4Sc_FT	414	MC	05	1	1062	0.61	0.00		0.10	0.76	0.08	0.05			0.76	0.44
2010_G4Sc_FT	414	MC	06	1	1062	0.61	0.01		0.67	0.08	0.04	0.20			0.67	0.45
2010_G4Sc_FT	414	MC	07	1	1062	0.61	0.02		0.67	0.00	0.13	0.18			0.67	0.46
2010_G4Sc_FT	414	CR	41	1	1062	0.61	0.02	0.41	0.57						0.57	0.57
2010_G4Sc_FT	414	CR	42	1	1062	0.61	0.01	0.82	0.16						0.16	0.42
2010_G4Sc_FT	414	CR	43	1	1062	0.61	0.02	0.39	0.59						0.59	0.35
2010_G4Sc_FT	414	CR	44	1	1062	0.61	0.02	0.51	0.47						0.47	0.36
2010_G4Sc_FT	414	CR	45	2	1062	0.61	0.04	0.32	0.37	0.27					0.91	0.66
2010_G4Sc_FT	415	MC	01	1	1058	0.56	0.00		0.79	0.14	0.05	0.02			0.79	0.45
2010_G4Sc_FT	415	MC	02	1	1058	0.56	0.00		0.16	0.07	0.72	0.06			0.72	0.42
2010_G4Sc_FT	415	MC	03	1	1058	0.56	0.00		0.01	0.03	0.01	0.94			0.94	0.37
2010_G4Sc_FT	415	MC	04	1	1058	0.56	0.00		0.04	0.83	0.02	0.11			0.83	0.44

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_G4Sc_FT	415	MC	05	1	1058	0.56	0.00		0.03	0.02	0.85	0.11			0.85	0.41
2010_G4Sc_FT	415	MC	06	1	1058	0.56	0.00		0.21	0.20	0.34	0.24			0.34	0.43
2010_G4Sc_FT	415	MC	07	1	1058	0.56	0.01		0.53	0.19	0.11	0.16			0.53	0.53
2010_G4Sc_FT	415	MC	08	1	1058	0.56	0.01		0.07	0.06	0.74	0.12			0.74	0.49
2010_G4Sc_FT	415	MC	09	1	1058	0.56	0.01		0.86	0.03	0.08	0.03			0.86	0.44
2010_G4Sc_FT	415	CR	41	1	1058	0.56	0.01	0.30	0.70						0.70	0.51
2010_G4Sc_FT	415	CR	42	1	1058											
2010_G4Sc_FT	415	CR	43	1	1058											
2010_G4Sc_FT	415	CR	44	1	1058											
2010_G4Sc_FT	416	MC	01	1	1070	0.69	0.00		0.03	0.04	0.38	0.55			0.38	0.50
2010_G4Sc_FT	416	MC	02	1	1070	0.69	0.00		0.02	0.06	0.10	0.82			0.82	0.41
2010_G4Sc_FT	416	MC	03	1	1070	0.69	0.00		0.13	0.04	0.78	0.05			0.78	0.42
2010_G4Sc_FT	416	MC	04	1	1070	0.69	0.00		0.53	0.10	0.25	0.12			0.53	0.45
2010_G4Sc_FT	416	MC	05	1	1070	0.69	0.00		0.10	0.04	0.84	0.02			0.84	0.35
2010_G4Sc_FT	416	MC	06	1	1070	0.69	0.00		0.11	0.13	0.54	0.22			0.54	0.44
2010_G4Sc_FT	416	MC	07	1	1070	0.69	0.01		0.11	0.06	0.74	0.09			0.74	0.42
2010_G4Sc_FT	416	MC	08	1	1070	0.69	0.02		0.47	0.33	0.06	0.13			0.33	0.39
2010_G4Sc_FT	416	CR	41	2	1070	0.69	0.10	0.20	0.51	0.19					0.89	0.58
2010_G4Sc_FT	416	CR	42	1	1070	0.69	0.03	0.33	0.64						0.64	0.50
2010_G4Sc_FT	416	CR	43	1	1070	0.69	0.03	0.45	0.51						0.51	0.57
2010_G4Sc_FT	416	CR	44	1	1070	0.69	0.03	0.40	0.56						0.56	0.51
2010_G4Sc_FT	416	CR	45	1	1070	0.69	0.04	0.13	0.83						0.83	0.43

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_G4Sc_FT	417	MC	01	1	1059	0.62	0.00		0.69	0.24	0.07	0.00			0.69	0.42
2010_G4Sc_FT	417	MC	02	1	1059	0.62	0.00		0.76	0.04	0.04	0.17			0.76	0.33
2010_G4Sc_FT	417	MC	03	1	1059	0.62	0.00		0.27	0.14	0.13	0.45			0.45	0.19
2010_G4Sc_FT	417	MC	04	1	1059	0.62	0.00		0.17	0.58	0.09	0.16			0.58	0.49
2010_G4Sc_FT	417	MC	05	1	1059	0.62	0.00		0.12	0.13	0.11	0.64			0.64	0.51
2010_G4Sc_FT	417	MC	06	1	1059	0.62	0.01		0.32	0.49	0.12	0.06			0.49	0.41
2010_G4Sc_FT	417	MC	07	1	1059	0.62	0.00		0.04	0.78	0.08	0.09			0.78	0.34
2010_G4Sc_FT	417	MC	08	1	1059	0.62	0.02		0.02	0.03	0.86	0.08			0.86	0.47
2010_G4Sc_FT	417	CR	41	1	1059	0.62	0.01	0.29	0.70						0.70	0.36
2010_G4Sc_FT	417	CR	42	1	1059	0.62	0.03	0.64	0.34						0.34	0.64
2010_G4Sc_FT	417	CR	43	2	1059	0.62	0.02	0.46	0.18	0.34					0.86	0.70
2010_G4Sc_FT	417	CR	44	1	1059	0.62	0.02	0.24	0.74						0.74	0.37
2010_G4Sc_FT	418	MC	01	1	1071	0.67	0.00		0.04	0.51	0.08	0.37			0.51	0.30
2010_G4Sc_FT	418	MC	02	1	1071	0.67	0.00		0.81	0.09	0.06	0.03			0.81	0.45
2010_G4Sc_FT	418	MC	03	1	1071	0.67	0.00		0.16	0.79	0.04	0.01			0.79	0.42
2010_G4Sc_FT	418	MC	04	1	1071	0.67	0.00		0.01	0.96	0.01	0.01			0.96	0.32
2010_G4Sc_FT	418	MC	05	1	1071	0.67	0.00		0.27	0.04	0.02	0.67			0.67	0.49
2010_G4Sc_FT	418	MC	06	1	1071	0.67	0.00		0.12	0.81	0.04	0.03			0.81	0.52
2010_G4Sc_FT	418	MC	07	1	1071	0.67	0.00		0.68	0.07	0.08	0.17			0.68	0.52
2010_G4Sc_FT	418	MC	08	1	1071	0.67	0.00		0.02	0.07	0.75	0.16			0.75	0.56
2010_G4Sc_FT	418	MC	09	1	1071	0.67	0.03		0.19	0.05	0.11	0.62			0.62	0.58
2010_G4Sc_FT	418	CR	41	1	1071	0.67	0.01	0.25	0.74						0.74	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
2010_G4Sc_FT	418	CR	42	1	1071	0.67	0.02	0.25	0.73						0.73	0.49
2010_G4Sc_FT	418	CR	43	1	1071	0.67	0.02	0.17	0.81						0.81	0.51
2010_G4Sc_FT	419	MC	01	1	1049	0.57	0.00		0.09	0.11	0.29	0.51			0.51	0.36
2010_G4Sc_FT	419	MC	02	1	1049	0.57	0.00		0.02	0.04	0.06	0.88			0.88	0.39
2010_G4Sc_FT	419	MC	03	1	1049	0.57	0.00		0.70	0.06	0.12	0.11			0.70	0.45
2010_G4Sc_FT	419	MC	04	1	1049	0.57	0.00		0.66	0.16	0.15	0.03			0.66	0.33
2010_G4Sc_FT	419	MC	05	1	1049	0.57	0.00		0.72	0.13	0.05	0.10			0.72	0.37
2010_G4Sc_FT	419	MC	06	1	1049	0.57	0.00		0.05	0.48	0.02	0.45			0.45	0.40
2010_G4Sc_FT	419	MC	07	1	1049	0.57	0.00		0.11	0.12	0.69	0.08			0.69	0.48
2010_G4Sc_FT	419	MC	08	1	1049	0.57	0.00		0.19	0.28	0.04	0.49			0.49	0.38
2010_G4Sc_FT	419	MC	09	1	1049	0.57	0.00		0.12	0.02	0.82	0.02			0.82	0.36
2010_G4Sc_FT	419	MC	10	1	1049	0.57	0.02		0.42	0.48	0.02	0.05			0.48	0.39
2010_G4Sc_FT	419	CR	41	1	1049	0.57	0.01	0.54	0.46						0.46	0.53
2010_G4Sc_FT	419	CR	42	1												
2010_G4Sc_FT	419	CR	43	1	1049	0.57	0.03	0.48	0.49						0.49	0.57

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_G4Sc_FT	411	MC	01	1	1077	-0.6300							1.11
2010_G4Sc_FT	411	MC	02	1	1077	-0.7400							1.02
2010_G4Sc_FT	411	MC	03	1	1077	0.9800							1.00
2010_G4Sc_FT	411	MC	04	1	1077	1.3200							1.11
2010_G4Sc_FT	411	MC	05	1	1077	1.0300							1.18
2010_G4Sc_FT	411	MC	06	1	1077	-0.7700							1.06
2010_G4Sc_FT	411	MC	07	1	1077	2.0885							1.22
2010_G4Sc_FT	411	MC	08	1	1077	0.5600							1.18
2010_G4Sc_FT	411	MC	09	1	1077	-0.4100							0.95
2010_G4Sc_FT	411	MC	10	1	1077	-0.9100							0.89
2010_G4Sc_FT	411	MC	11	1	1077	0.4100							0.93
2010_G4Sc_FT	411	MC	12	1	1077	-0.1000							0.99
2010_G4Sc_FT	411	MC	13	1	1077	0.8600							1.05
2010_G4Sc_FT	411	MC	14	1	1077	1.1400							0.93
2010_G4Sc_FT	411	MC	15	1	1077	-0.9000							0.99
2010_G4Sc_FT	411	MC	16	1	1077	-0.0558							0.90
2010_G4Sc_FT	411	MC	17	1	1077	-1.0600							0.79
2010_G4Sc_FT	411	MC	18	1	1077	-0.2000							0.93
2010_G4Sc_FT	411	MC	19	1	1077	0.0934							0.93
2010_G4Sc_FT	411	MC	20	1	1077	-0.2500							1.11
2010_G4Sc_FT	411	MC	21	1	1077	-0.1100							1.08

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

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2010_G4Sc_FT	411	MC	22	1	1077	-0.9300							0.83
2010_G4Sc_FT	411	MC	23	1	1077	-0.5600							0.96
2010_G4Sc_FT	411	MC	24	1	1077	0.9300							1.03
2010_G4Sc_FT	411	MC	25	1	1077	-0.3263							0.94
2010_G4Sc_FT	412	MC	01	1	1071	0.3034							0.90
2010_G4Sc_FT	412	MC	02	1	1071	-0.0545							0.89
2010_G4Sc_FT	412	MC	03	1	1071	0.5983							1.06
2010_G4Sc_FT	412	MC	04	1	1071	0.5320							0.95
2010_G4Sc_FT	412	MC	05	1	1071	1.7514							0.97
2010_G4Sc_FT	412	MC	06	1	1071	0.9454							0.95
2010_G4Sc_FT	412	MC	07	1	1071	-2.1710							0.95
2010_G4Sc_FT	412	MC	08	1	1071	-0.3579							0.86
2010_G4Sc_FT	412	MC	09	1	1071	0.9216							1.15
2010_G4Sc_FT	412	CR	41	1	1071	1.2818							1.16
2010_G4Sc_FT	412	CR	42	1	1071	-0.1248							1.07
2010_G4Sc_FT	412	CR	43	1	1071	0.9928							0.86
2010_G4Sc_FT	412	CR	44	1	1071	0.9026							1.12
2010_G4Sc_FT	413	MC	01	1	1059	1.6524							1.19
2010_G4Sc_FT	413	MC	02	1	1059	0.3002							0.98
2010_G4Sc_FT	413	MC	03	1	1059	-0.8268							1.02
2010_G4Sc_FT	413	MC	04	1	1059	0.3377							0.97
2010_G4Sc_FT	413	MC	05	1	1059	0.4063							1.02
2010_G4Sc_FT	413	MC	06	1	1059	0.5142							1.10

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

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2010_G4Sc_FT	413	MC	07	1	1059	-0.3843							0.94
2010_G4Sc_FT	413	CR	41	2	1059	0.6911	-1.1644	1.1644					0.89
2010_G4Sc_FT	413	CR	42	1	1059	0.7551							0.95
2010_G4Sc_FT	413	CR	43	2	1059	1.3094	-0.3303	0.3303					0.89
2010_G4Sc_FT	413	CR	44	1	1059	-0.0323							0.92
2010_G4Sc_FT	413	CR	45	1	1059	1.1006							1.09
2010_G4Sc_FT	414	MC	01	1	1062	0.5202							1.12
2010_G4Sc_FT	414	MC	02	1	1062	-0.7964							0.88
2010_G4Sc_FT	414	MC	03	1	1062	-1.5418							0.96
2010_G4Sc_FT	414	MC	04	1	1062	-0.8872							0.98
2010_G4Sc_FT	414	MC	05	1	1062	0.0077							0.97
2010_G4Sc_FT	414	MC	06	1	1062	0.5349							1.00
2010_G4Sc_FT	414	MC	07	1	1062	0.5349							0.99
2010_G4Sc_FT	414	CR	41	1	1062	0.9971							0.87
2010_G4Sc_FT	414	CR	42	1	1062	3.2953							0.96
2010_G4Sc_FT	414	CR	43	1	1062	0.9199							1.14
2010_G4Sc_FT	414	CR	44	1	1062	1.4673							1.15
2010_G4Sc_FT	414	CR	45	2	1062	1.5653	-0.6505	0.6505					0.88
2010_G4Sc_FT	415	MC	01	1	1058	-0.0709							0.99
2010_G4Sc_FT	415	MC	02	1	1058	0.4077							1.08
2010_G4Sc_FT	415	MC	03	1	1058	-1.5601							0.94
2010_G4Sc_FT	415	MC	04	1	1058	-0.3412							0.98
2010_G4Sc_FT	415	MC	05	1	1058	-0.4959							1.00

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

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2010_G4Sc_FT	415	MC	06	1	1058	2.3789							1.10
2010_G4Sc_FT	415	MC	07	1	1058	1.4017							0.96
2010_G4Sc_FT	415	MC	08	1	1058	0.2400							0.97
2010_G4Sc_FT	415	MC	09	1	1058	-0.5534							0.95
2010_G4Sc_FT	415	CR	41	1	1058	0.5254							0.96
2010_G4Sc_FT	415	CR	42	1	1058								
2010_G4Sc_FT	415	CR	43	1	1058								
2010_G4Sc_FT	415	CR	44	1	1058								
2010_G4Sc_FT	416	MC	01	1	1070	1.9384							0.94
2010_G4Sc_FT	416	MC	02	1	1070	-0.4064							0.99
2010_G4Sc_FT	416	MC	03	1	1070	-0.1093							1.02
2010_G4Sc_FT	416	MC	04	1	1070	1.2316							1.05
2010_G4Sc_FT	416	MC	05	1	1070	-0.5460							1.04
2010_G4Sc_FT	416	MC	06	1	1070	1.1908							1.06
2010_G4Sc_FT	416	MC	07	1	1070	0.1320							1.04
2010_G4Sc_FT	416	MC	08	1	1070	2.2238							1.07
2010_G4Sc_FT	416	CR	41	2	1070	1.7006	-1.3011	1.3011					1.02
2010_G4Sc_FT	416	CR	42	1	1070	0.6750							0.96
2010_G4Sc_FT	416	CR	43	1	1070	1.2949							0.89
2010_G4Sc_FT	416	CR	44	1	1070	1.0543							0.96
2010_G4Sc_FT	416	CR	45	1	1070	-0.5234							0.95
2010_G4Sc_FT	417	MC	01	1	1059	0.4542							1.01
2010_G4Sc_FT	417	MC	02	1	1059	0.0497							1.05

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

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2010_G4Sc_FT	417	MC	03	1	1059	1.5888							1.31
2010_G4Sc_FT	417	MC	04	1	1059	0.9878							0.96
2010_G4Sc_FT	417	MC	05	1	1059	0.7047							0.92
2010_G4Sc_FT	417	MC	06	1	1059	1.4117							1.06
2010_G4Sc_FT	417	MC	07	1	1059	-0.0561							1.06
2010_G4Sc_FT	417	MC	08	1	1059	-0.6477							0.87
2010_G4Sc_FT	417	CR	41	1	1059	0.3735							1.06
2010_G4Sc_FT	417	CR	42	1	1059	2.1673							0.76
2010_G4Sc_FT	417	CR	43	2	1059	1.6160	0.4196	-0.4196					0.78
2010_G4Sc_FT	417	CR	44	1	1059	0.1948							1.04
2010_G4Sc_FT	418	MC	01	1	1071	1.5598							1.30
2010_G4Sc_FT	418	MC	02	1	1071	-0.1793							0.99
2010_G4Sc_FT	418	MC	03	1	1071	-0.0282							1.05
2010_G4Sc_FT	418	MC	04	1	1071	-2.0848							0.90
2010_G4Sc_FT	418	MC	05	1	1071	0.7346							1.02
2010_G4Sc_FT	418	MC	06	1	1071	-0.1370							0.92
2010_G4Sc_FT	418	MC	07	1	1071	0.7086							0.97
2010_G4Sc_FT	418	MC	08	1	1071	0.2953							0.90
2010_G4Sc_FT	418	MC	09	1	1071	1.0184							0.91
2010_G4Sc_FT	418	CR	41	1	1071	0.3187							1.03
2010_G4Sc_FT	418	CR	42	1	1071	0.3763							1.00
2010_G4Sc_FT	418	CR	43	1	1071	-0.1510							0.92
2010_G4Sc_FT	419	MC	01	1	1049	1.3177							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
2010_G4Sc_FT	419	MC	02	1	1049	-0.7951							0.92
2010_G4Sc_FT	419	MC	03	1	1049	0.4145							0.96
2010_G4Sc_FT	419	MC	04	1	1049	0.6044							1.10
2010_G4Sc_FT	419	MC	05	1	1049	0.2815							1.03
2010_G4Sc_FT	419	MC	06	1	1049	1.6038							1.06
2010_G4Sc_FT	419	MC	07	1	1049	0.4344							0.93
2010_G4Sc_FT	419	MC	08	1	1049	1.4052							1.07
2010_G4Sc_FT	419	MC	09	1	1049	-0.3572							0.97
2010_G4Sc_FT	419	MC	10	1	1049	1.4315							1.06
2010_G4Sc_FT	419	CR	41	1	1049	1.5549							0.90
2010_G4Sc_FT	419	CR	42	1									
2010_G4Sc_FT	419	CR	43	1	1049	1.3833							0.86

Appendix C: DIF Statistics

Table 10. DIF Statistics

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
411	1	MC	1.02	5.48	0.14	B	F
411	2	MC	-0.67	2.06	-0.08		
411	3	MC	0.21	0.37	0.04		
411	4	MC	0.08	0.06	0.02		
411	5	MC	0.08	0.06	0.02		
411	6	MC	-0.04	0.01	-0.01		
411	7	MC	-0.45	1.88	-0.08		
411	8	MC	0.05	0.02	0.01		
411	9	MC	-1.02	5.31	-0.12	B	M
411	10	MC	-1.24	5.54	-0.13	B	M
411	11	MC	0.09	0.05	0.00		
411	12	MC	-0.50	1.53	-0.06		
411	13	MC	0.49	2.08	0.07		
411	14	MC	0.30	0.67	0.04		
411	15	MC	0.97	3.63	0.10		
411	16	MC	-0.94	4.72	-0.12		
411	17	MC	0.88	2.22	0.09		
411	18	MC	0.50	1.44	0.06		
411	19	MC	0.45	1.30	0.06		
411	20	MC	-0.18	0.20	-0.02		
411	21	MC	-1.00	7.07	-0.14	B	M
411	22	MC	-0.19	0.12	-0.02		

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
411	23	MC	1.03	5.13	0.11	B	F
411	24	MC	-0.16	0.23	-0.03		
411	25	MC	0.46	1.10	0.06		
412	1	MC	-0.26	0.42	-0.04		
412	2	MC	0.87	4.19	0.11		
412	3	MC	0.23	0.46	0.04		
412	4	MC	-0.21	0.33	-0.03		
412	5	MC	-0.18	0.25	-0.02		
412	6	MC	0.98	7.58	0.15		
412	7	MC	1.09	1.71	0.09		
412	8	MC	0.31	0.46	0.02		
412	9	MC	0.35	1.15	0.06		
412	41	OE		2.56	-0.08		
412	42	OE		6.35	-0.15		
412	43	OE		2.26	0.07		
412	44	OE		8.26	-0.17		
413	1	MC	-0.60	3.54	-0.11		
413	2	MC	-0.28	0.54	-0.04		
413	3	MC	0.53	1.08	0.04		
413	4	MC	-0.05	0.02	-0.01		
413	5	MC	0.12	0.10	0.03		
413	6	MC	-0.10	0.09	-0.01		

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
413	7	MC	-0.38	0.73	-0.05		
413	41	OE		0.59	0.03		
413	42	OE		1.83	0.07		
413	43	OE		0.10	0.01		
413	44	OE		2.27	-0.07		
413	45	OE		1.58	0.08		
414	1	MC	-0.06	0.03	-0.01		
414	2	MC	0.48	0.90	0.05		
414	3	MC	-0.13	0.05	-0.01		
414	4	MC	-0.06	0.01	0.00		
414	5	MC	0.48	1.53	0.07		
414	6	MC	0.22	0.41	0.04		
414	7	MC	-0.62	3.13	-0.10		
414	41	OE		0.09	0.02		
414	42	OE		1.30	0.06		
414	43	OE		0.14	0.02		
414	44	OE		2.14	0.09		
414	45	OE		6.09	-0.12		
415	1	MC	0.54	1.81	0.08		
415	2	MC	-0.27	0.60	-0.04		
415	3	MC	0.90	1.77	0.09		
415	4	MC	-0.58	1.74	-0.08		
415	5	MC	0.03	0.01	0.01		

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
415	6	MC	-0.04	0.01	0.00		
415	7	MC	0.02	0.00	0.01		
415	8	MC	0.04	0.01	0.01		
415	9	MC	-1.24	6.99	-0.17	B	M
415	41	OE		0.00	-0.01		
415	42	OE					
415	43	OE					
415	44	OE					
416	1	MC	0.47	1.71	0.06		
416	2	MC	-0.33	0.66	-0.03		
416	3	MC	0.88	5.02	0.14		
416	4	MC	-0.44	1.75	-0.08		
416	5	MC	-0.97	4.85	-0.15		
416	6	MC	0.13	0.15	0.02		
416	7	MC	-0.13	0.13	-0.01		
416	8	MC	-0.43	1.49	-0.06		
416	41	OE		0.04	-0.02		
416	42	OE		0.06	-0.01		
416	43	OE		0.18	-0.02		
416	44	OE		5.09	0.13		
416	45	OE		0.18	0.01		
417	1	MC	-1.39	15.04	-0.23	B	M
417	2	MC	0.05	0.02	0.01		

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
417	3	MC	-0.04	0.02	-0.01		
417	4	MC	0.43	1.56	0.06		
417	5	MC	-0.06	0.03	-0.01		
417	6	MC	0.10	0.09	0.01		
417	7	MC	0.22	0.34	0.04		
417	8	MC	-0.53	1.12	-0.06		
417	41	OE		8.69	0.16		
417	42	OE		0.06	0.02		
417	43	OE		1.17	-0.05		
417	44	OE		1.17	0.07		
418	1	MC	-0.83	6.80	-0.15		
418	2	MC	-0.05	0.02	-0.01		
418	3	MC	-0.10	0.07	-0.02		
418	4	MC	0.05	0.00	0.01		
418	5	MC	-0.29	0.64	-0.05		
418	6	MC	0.17	0.16	0.02		
418	7	MC	-0.20	0.31	-0.03		
418	8	MC	-0.21	0.27	-0.03		

*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
418	9	MC	0.57	2.35	0.08		
418	41	OE		0.18	0.02		
418	42	OE		9.17	0.17		
418	43	OE		0.00	0.01		
419	1	MC	-0.18	0.34	-0.04		
419	2	MC	1.15	5.19	0.14	B	F
419	3	MC	0.48	1.77	0.07		
419	4	MC	-0.61	3.30	-0.11		
419	5	MC	0.66	3.39	0.10		
419	6	MC	-0.59	3.21	-0.11		
419	7	MC	0.50	1.78	0.08		
419	8	MC	0.10	0.10	0.02		
419	9	MC	0.62	2.10	0.09		
419	10	MC	-0.12	0.15	-0.02		
419	41	OE		0.00	0.00		
419	42	OE					
419	43	OE		10.02	-0.17		

Appendix D: Operational Test Map

Table 11. Operational Test Map for June 2010

Position	Item Type	Max Points	Weight	Strand	Mean	Point-Biserial	Rasch	S1	S2	S3	S4
1	MC	1	1	4	0.90	0.31	-1.33				
2	MC	1	1	4	0.89	0.38	-1.21				
3	MC	1	1	4	0.89	0.34	-1.17				
4	MC	1	1	4	0.87	0.29	-1.00				
5	MC	1	1	4	0.51	0.50	1.18				
6	MC	1	1	4	0.76	0.43	-0.05				
7	MC	1	1	4	0.75	0.47	-0.10				
8	MC	1	1	4	0.73	0.47	0.05				
9	MC	1	1	4	0.66	0.40	0.42				
10	MC	1	1	4	0.67	0.37	0.33				
11	MC	1	1	4	0.59	0.52	0.84				
12	MC	1	1	1	0.72	0.44	0.11				
13	MC	1	1	4	0.72	0.45	0.12				
14	MC	1	1	4	0.82	0.30	-0.60				
15	MC	1	1	4	0.40	0.32	1.66				
16	MC	1	1	4	0.95	0.24	-2.06				
17	MC	1	1	4	0.68	0.46	0.30				
18	MC	1	1	4	0.87	0.38	-1.02				
19	MC	1	1	4	0.86	0.40	-0.91				
20	MC	1	1	4	0.85	0.39	-0.77				
21	MC	1	1	4	0.68	0.45	0.28				

Table 11. Operational Test Map for June 2010 *(continued)*

Position	Item Type	Max Points	Weight	Strand	Mean	Point-Biserial	Rasch	S1	S2	S3	S4
22	MC	1	1	4	0.73	0.34	0.00				
23	MC	1	1	4	0.66	0.41	0.50				
24	MC	1	1	4	0.65	0.48	0.51				
25	MC	1	1	4	0.64	0.52	0.60				
26	MC	1	1	4	0.56	0.42	0.95				
27	MC	1	1	4	0.51	0.41	1.15				
28	MC	1	1	4	0.61	0.42	0.72				
29	MC	1	1	4	0.45	0.39	1.41				
30	MC	1	1	1	0.67	0.46	0.37				
31	CR	1	1	1	0.75	0.43	-0.12				
32	CR	1	1	4	0.82	0.40	-0.51				
33	CR	1	1	4	0.79	0.35	-0.35				
34	CR	1	1	4	0.60	0.60	0.74				
35	CR	1	1	4	0.34	0.58	2.08				
36	CR	1	1	4	0.44	0.57	1.53				
37	CR	1	1	4	0.70	0.41	0.27				
38	CR	1	1	4	0.76	0.47	-0.10				
39	CR	2	1	4	0.93	0.62		-0.32	0.32		
40	CR	1	1	4	0.95	0.28	-1.94				
41	CR	1	1	6	0.82	0.42	-0.56				
42	CR	1	1	6	0.64	0.48	0.53				
43	CR	1	1	4	0.63	0.53	0.66				
44	CR	1	1	4	0.30	0.51	2.44				

Appendix E: Scoring Table

Table 12. Scoring Table for June 2010

Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score
0	-4.801	0.000		16	-0.510	39.418		32	1.278	78.905
1	-4.056	1.855		17	-0.395	42.164		33	1.407	80.900
2	-3.311	3.967		18	-0.282	44.900		34	1.541	82.795
3	-2.854	6.067		19	-0.171	47.637		35	1.683	84.640
4	-2.515	8.266		20	-0.061	50.344		36	1.834	86.409
5	-2.241	10.560		21	0.047	52.993		37	1.997	88.124
6	-2.008	12.922		22	0.155	55.644		38	2.176	89.784
7	-1.803	15.386		23	0.262	58.225		39	2.374	91.376
8	-1.619	17.894		24	0.369	60.752		40	2.601	92.925
9	-1.451	20.461		25	0.477	63.234		41	2.869	94.421
10	-1.295	23.084		26	0.585	65.652		42	3.201	95.867
11	-1.148	25.760		27	0.695	68.023		43	3.652	97.275
12	-1.010	28.432		28	0.806	70.328		44	4.391	98.654
13	-0.878	31.160		29	0.920	72.580		45	5.130	99.948
14	-0.751	33.910		30	1.036	74.757		32	1.278	78.905
15	-0.629	36.658		31	1.155	76.863		33	1.407	80.900