New York State Examination in Grade 4
Elementary-Level Science

2012 Field Test Analysis, Equating Procedure, and Scaling of Operational Test Forms

Technical Report

Prepared for the New York State Education Department
by Pearson

May 2013

Copyright

Developed and published under contract with the New York State Education Department by Pearson. Copyright © 2012 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

List of Tables

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 2012. Specifically, contained within this report are procedures for, and results of, field test analysis, equating, and scaling of operational test forms that were conducted by Pearson.

Section II: Field Test Analysis

In May 2012, 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 the 70th percentile on the index with at least 100 students per square mile or enrollment greater than 2500
Rural	All districts at or above the 70th percentile with fewer than 50 students per square mile or enrollment of fewer 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-test and item response theory analyses were conducted using the data to evaluate the quality of the test items.

DATA CLEANUP

Field test forms contained multiple-choice and constructed-response item types. Response data were contained in one file that contained 15765 student records. After the exclusion rules[1] were applied, the resulting field test data file contained 15690 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 Item Response Theory (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 could 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 administration conditions might influence student responses, making a student's observed score higher or lower than the student's true score. The error is considered random. After repeated administrations, the mean of the error scores is virtually zero (0). 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,

[1] Exclusion rules define which test records are to be considered invalid. Such records include those without both an MC and a CR component, records with invalid or out-of-range form numbers, records without any responses, and duplicate records. These records were dropped and not included in any analyses.

reliability estimates, and various statistics related to rater agreement have been calculated and are summarized in the following section.

Item Difficulty

Item difficulty is an indication of students' achievement 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 determined by the proportion of students who answer an item correctly. For example, an item where 90% of the student responses to a multiple-choice item are correct is easier than a multiple-choice item having correct responses from only 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 (item score) 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 to constructed-response items. The possible range for the point-biserial correlation is −1.0 to 1.0, 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 from left to right identify the form number, the number of students who took each form, and the number of items on each field test form, respectively. The remaining columns are divided into two sections (i.e., item difficulty and point-biserial correlations). Recall that for constructed-response items, item means were divided by the maximum number of points possible in order to place them in the same metric as the multiple-choice items. Only 11 items had item difficulties greater than 0.90. With respect to the point-biserial correlations, none fell below 0.25.

Table 2. Classical Item Analysis

			Item Difficulty			Point-Biserial		
Form	N-Count	No. of Items	<0.50	0.50 to 0.90	>0.90	<0.25	0.25 to 0.50	>0.50
501	1,116	25	1	23	1	0	20	5
502	1,138	13	1	12	0	0	10	3
503	1,118	12	1	10	0	0	6	5
504	1,127	13	0	12	0	0	9	3
505	1,147	12	2	8	1	0	11	0
506	1,113	11	0	10	1	0	4	7
507	1,135	11	1	8	1	0	6	4
508	1,128	11	2	8	1	0	9	2
509	1,130	10	0	10	0	0	8	2
510	1,118	11	0	10	1	0	6	5
511	1,120	13	2	9	1	0	8	4
512	1,115	12	1	9	2	0	9	3
513	1,089	12	1	11	0	0	8	4
514	1,096	13	2	9	2	0	12	1

*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 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 items blank and “M0” through “M2” are the proportions of students who received scores of 0 through 2. “Mean” is the average of the scores received by the students. The final (right-most) column contains the point-biserial correlation for each item. There are some instances of items with 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: (1) test length and (2) homogeneity of the 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; Kuder and Richardson, 1937), which is the equivalent for dichotomous items.

Table 3 contains the internal consistency statistics for all of the field test forms. These statistics range from 0.52 to 0.83 and are based solely on the items in the individual field test forms. It is expected that these statistics would be greater when these items are used operationally since operational test forms contain more items than the field test forms.

Table 3. Test and Scoring Reliability

Form Number	Test Reliability	Scoring Reliability
501	0.83	n/a
502	0.68	0.98
503	0.68	0.89
504	0.64	0.85
505	0.55	0.87
506	0.69	0.93
507	0.59	0.85
508	0.62	0.96
509	0.52	0.80
510	0.63	0.91
511	0.60	0.70
512	0.60	0.86
513	0.66	0.98
514	0.58	0.84

Scoring Reliability

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

As an overall measure of scoring reliability, the Pearson Correlation Coefficient between the first and second scores for each of the constructed-response items was computed. This statistic is often used as an overall indicator of scoring reliability and generally ranges from 0.00 to near 1.00. Table 3 contains the results from these analyses in the column headed “Scoring Reliability.” The correlations ranged from 0.70 to 0.98, 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 occurrences of these differences for each item. The proportion of perfect agreement between the first and second reads ranged from 0.84 to 1.00. Most were at 0.90 or above.

Table 4. Point Differences Between First and Second Reads

Form	Item	Score Points	Difference (First Read Minus Second Read)		
			–1	0	1
502	41	1	0.01	0.97	0.02
502	42	1	0.00	1.00	0.00
502	43	1	0.00	0.99	0.01
502	44	1	0.00	1.00	0.00
503	42	1	0.00	0.99	0.01
503	43	1	0.04	0.95	0.01
503	44	1	0.06	0.89	0.05
504	41	1	0.04	0.96	0.00
504	42	1	0.03	0.94	0.04
505	41	1	0.03	0.95	0.02
505	42	1	0.03	0.95	0.03
505	43	1	0.08	0.88	0.04
505	44	1	0.00	1.00	0.00
506	41	1	0.00	1.00	0.00
506	42	1	0.00	1.00	0.00
506	43	1	0.05	0.92	0.04
506	44	1	0.00	1.00	0.00
506	45	1	0.03	0.96	0.01
507	41	1	0.02	0.92	0.06

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

Form	Item	Score Points	Difference (First Read Minus Second Read)		
			−1	0	1
507	42	1	0.01	0.96	0.03
507	43	1	0.01	0.93	0.06
508	41	1	0.00	1.00	0.00
508	42	1	0.01	0.99	0.00
508	43	1	0.01	0.99	0.00
508	44	1	0.03	0.95	0.03
509	41	1	0.00	0.97	0.03
509	42	1	0.01	0.90	0.09
509	43	1	0.07	0.84	0.08
509	44	1	0.01	0.94	0.05
510	41	1	0.00	1.00	0.00
510	42	1	0.00	0.96	0.04
510	43	1	0.06	0.93	0.01
511	41	1	0.04	0.91	0.04
511	42	1	0.03	0.92	0.04
511	43	1	0.09	0.88	0.03
511	44	1	0.07	0.93	0.00
512	41	1	0.03	0.94	0.04
512	42	1	0.04	0.91	0.05
512	43	1	0.03	0.95	0.02
512	44	1	0.04	0.94	0.03
513	41	1	0.00	1.00	0.00
513	42	1	0.01	0.97	0.02
513	43	1	0.00	1.00	0.00
514	41	1	0.05	0.92	0.03
514	42	1	0.03	0.94	0.03
514	43	1	0.00	1.00	0.00
514	44	1	0.03	0.94	0.02
514	45	1	0.07	0.86	0.07

Table 5 contains additional summary information regarding the first and second reads. In the fifth column from the left, the percent of exact matches between the first and second scores is provided. “Adj.” is the percentage of differences with a magnitude of 1. “Total” is the sum of the two prior columns and contains values of 100%. Since all

items are worth one point, the sum of the exact and adjacent columns all sum to 100%. As previously discussed, most items had exact agreement rates between their first and second reads of 90% or greater.

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
502	41	1	111	97.3	2.7	100.0	0.5	0.5	0.50	0.50	0.95	0.95
502	42	1	109	100.0	0.0	100.0	0.8	0.8	0.38	0.38	1.00	1.00
502	43	1	107	99.1	0.9	100.0	0.6	0.6	0.49	0.49	0.98	0.98
502	44	1	106	100.0	0.0	100.0	0.9	0.9	0.35	0.35	1.00	1.00
503	42	1	103	99.0	1.0	100.0	0.7	0.7	0.44	0.44	0.97	0.97
503	43	1	102	95.1	4.9	100.0	0.4	0.4	0.48	0.49	0.90	0.89
503	44	1	104	89.4	10.6	100.0	0.5	0.5	0.50	0.50	0.79	0.79
504	41	1	109	96.3	3.7	100.0	0.9	0.9	0.35	0.30	0.84	0.83
504	42	1	110	93.6	6.4	100.0	0.7	0.7	0.45	0.46	0.84	0.84
505	41	1	111	95.5	4.5	100.0	0.4	0.4	0.49	0.49	0.91	0.91
505	42	1	111	94.6	5.4	100.0	0.7	0.7	0.46	0.46	0.87	0.87
505	43	1	108	88.0	12.0	100.0	0.8	0.8	0.42	0.38	0.63	0.62
505	44	1	111	100.0	0.0	100.0	0.9	0.9	0.34	0.34	1.00	1.00
506	41	1	107	100.0	0.0	100.0	1.0	1.0	0.21	0.21	1.00	1.00
506	42	1	106	100.0	0.0	100.0	0.8	0.8	0.43	0.43	1.00	1.00
506	43	1	107	91.6	8.4	100.0	0.7	0.7	0.47	0.47	0.81	0.81
506	44	1	102	100.0	0.0	100.0	0.7	0.7	0.47	0.47	1.00	1.00
506	45	1	102	96.1	3.9	100.0	0.8	0.8	0.41	0.40	0.88	0.88
507	41	1	125	92.0	8.0	100.0	0.5	0.5	0.50	0.50	0.84	0.84
507	42	1	125	96.0	4.0	100.0	0.8	0.8	0.41	0.42	0.89	0.88
507	43	1	124	92.7	7.3	100.0	0.8	0.8	0.38	0.42	0.78	0.77
508	41	1	114	100.0	0.0	100.0	0.7	0.7	0.45	0.45	1.00	1.00
508	42	1	114	99.1	0.9	100.0	0.8	0.8	0.43	0.43	0.98	0.98
508	43	1	111	99.1	0.9	100.0	0.3	0.3	0.46	0.47	0.98	0.98
508	44	1	111	94.6	5.4	100.0	0.7	0.7	0.45	0.45	0.86	0.86
509	41	1	107	97.2	2.8	100.0	0.8	0.8	0.38	0.40	0.91	0.91
509	42	1	107	89.7	10.3	100.0	0.6	0.5	0.49	0.50	0.80	0.79
509	43	1	107	84.1	15.9	100.0	0.6	0.6	0.48	0.49	0.66	0.66
509	44	1	107	94.4	5.6	100.0	0.8	0.8	0.39	0.42	0.83	0.83
510	41	1	106	100.0	0.0	100.0	0.7	0.7	0.45	0.45	1.00	1.00
510	42	1	104	96.2	3.8	100.0	0.5	0.5	0.50	0.50	0.93	0.92
510	43	1	106	93.4	6.6	100.0	0.9	0.9	0.33	0.27	0.66	0.63

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 Corre-lation	Wt Kappa
511	41	1	117	91.5	8.5	100.0	0.8	0.8	0.37	0.37	0.69	0.69
511	42	1	117	92.3	7.7	100.0	0.9	0.9	0.34	0.35	0.67	0.67
511	43	1	117	88.0	12.0	100.0	0.7	0.7	0.47	0.45	0.73	0.72
511	44	1	117	93.2	6.8	100.0	0.9	0.9	0.35	0.27	0.70	0.66
512	41	1	111	93.7	6.3	100.0	0.6	0.6	0.50	0.50	0.87	0.87
512	42	1	110	90.9	9.1	100.0	0.6	0.6	0.49	0.50	0.81	0.81
512	43	1	110	95.5	4.5	100.0	0.7	0.7	0.46	0.46	0.89	0.89
512	44	1	109	93.6	6.4	100.0	0.7	0.7	0.48	0.47	0.86	0.86
513	41	1	109	100.0	0.0	100.0	0.3	0.3	0.47	0.47	1.00	1.00
513	42	1	111	97.3	2.7	100.0	0.7	0.7	0.44	0.45	0.93	0.93
513	43	1	109	100.0	0.0	100.0	0.7	0.7	0.47	0.47	1.00	1.00
514	41	1	122	91.8	8.2	100.0	0.7	0.8	0.44	0.43	0.78	0.78
514	42	1	120	94.2	5.8	100.0	0.9	0.9	0.30	0.31	0.69	0.69
514	43	1	122	100.0	0.0	100.0	0.8	0.8	0.40	0.40	1.00	1.00
514	44	1	122	94.3	5.7	100.0	0.6	0.6	0.50	0.50	0.88	0.88
514	45	1	121	86.0	14.0	100.0	0.4	0.4	0.49	0.49	0.71	0.71

*Adj. = Difference of 1

Constructed-Response Item Means and Standard Deviations

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

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

Intraclass Correlation

The intraclass correlation was computed for each item. This correlation is an estimate of the reliability of scoring based on an average of the first and second read scores. Correlations greater than 0.60 are considered very strong because they explain more than one-third of the variance in the scores. All items had intraclass correlations greater than to 0.60 (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 0.62 to 1.00, which in all cases indicates good-to-excellent reproducibility.

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

ITEM RESPONSE THEORY (IRT) STATISTICS

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

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

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

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

Item Calibration

As discussed above, the use of Rasch item difficulty statistics provides an advantage over the use of classical item means because they can be compared across test forms. Students from different samples responded to the various test forms. Although the samples were selected to be similar with respect to student ability, there were 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 2.00 is considered very difficult and should be examined carefully. If the item is measuring an important concept that students are having difficulty with, then the item can be useful. However, if the item is measuring a trivial concept or is written in a confusing manner, then it might not be appropriate to use on an operational test form. Likewise, any item with a Rasch difficulty less than −2.00 is considered very easy and usually provides little information regarding student achievement. The vast majority of test items should range between −2.00 and 2.00. This range represents approximately two standard deviations around the average difficulty of 0.00 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 0.70 to 1.30.

Table 6 contains a summary of the Partial Credit Model item analysis for each of the field test forms. The leftmost column lists the form numbers. The next two columns list the number of students who participated and the number of items on each field test

form, respectively. The remaining columns are divided into two sections. The first section pertains to the Rasch-item difficulty values, while the second pertains to the INFIT statistics. Nearly all of the items fell within the moderate −2.00 to 2.00 difficulty range, and only one item had an INFIT statistic outside the good-fit 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
501	1,116	25	0	24	1	0	25	0
502	1,138	13	0	13	0	0	13	0
503	1,118	12	0	10	1	0	11	0
504	1,127	13	0	12	0	0	12	0
505	1,147	12	0	11	0	0	11	0
506	1,113	11	0	11	0	0	11	0
507	1,135	11	0	10	0	0	10	0
508	1,128	11	0	10	1	0	11	0
509	1,130	10	0	10	0	0	10	0
510	1,118	11	0	11	0	0	10	1
511	1,120	13	0	12	0	0	12	0
512	1,115	12	0	11	1	0	12	0
513	1,089	12	0	11	1	0	12	0
514	1,096	13	0	12	1	0	13	0

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

All of the individual IRT item statistics are provided in Appendix B. The column entitled “RID” contains the Rasch-item difficulty statistics. S1–S6 contain the step values for the constructed-response items. Finally, the INFIT column 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 as to 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: (A) negligible DIF, (B) moderate DIF, and (C)large DIF. An item was

flagged if it exhibited a B or C category of DIF, using the following rules derived from the National Assessment of Educational Progress (NAEP) guidelines (Allen, Carlson, & Zalanak, 1999):

- MHDs not significantly different from 0.00 (based on alpha = 0.05) **or** |MHD| < 1.00 are classified as A.
- MHDs significantly different from 0.00 and {|MHD| ≥ 1.00 and < 1.50} **or** MHDs not significantly different from 0.00 and |MHD| ≥ 1.00 are classified as B.
- |MHD| ≥ 1.50 and significantly different from 0.00 are classified as C.

For the constructed-response items, the effect size (ES) of the standardized mean difference (SMD) was used to flag the 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 that 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: (AA) negligible DIF, (BB) moderate DIF, and (CC) large DIF. Only categories BB and CC were flagged in the results.

- Probability is > 0.05 **or** if |ES| is ≤ 0.17 is classified as AA.
- Probability is > 0.05 and if 0.17 < |ES| ≤ 0.25 is classified as BB.
- Probability is > 0.05 and if |ES| is > 0.25 is 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 they 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 2012 field test administration for the New York State Examination in Grade 4 Elementary-Level Science consisted of 13 field test forms numbered 502–514 and one anchor form labeled 501. The field test forms contained multiple-choice and constructed-response items. Each student participating in the field test was administered one of the 14 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.

EQUIVALENT-GROUP EQUATING DESIGN

The anchor form was equated to the item bank using a common-item equating design. The anchor item difficulty parameters were fixed to their 2011 item bank values. This places the item difficulty estimates and the ability estimates of the students taking the anchor form onto the item bank scale. After the anchor form was placed onto the bank scale, the average of the two mean ability estimates for the two forms was computed using ability estimates of nonextreme students. This average ability estimate was used to equate the remaining field test forms, as well as to update the item parameters for the anchor form.

As part of the anchor item equating, an item-stability check was performed. After fixing all of the items to their 2011 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 five 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 501 was 1.49. 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 501). 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
502	1.11	0.35
503	0.82	0.61
504	1.09	0.36
505	1.02	0.43
506	1.35	0.13
507	1.06	0.38
508	1.03	0.42
509	0.79	0.64
510	1.37	0.11
511	1.05	0.40
512	1.16	0.31
513	1.09	0.37
514	1.27	0.20

The equated item parameters for the field test items can now be compared across test forms, given that 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 2012 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

2012 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 (0) 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 (0) is assigned a scale score of zero. The maximum raw score is assigned a scale score of 100. With respect to the 65 and 85 scale scores, the raw scores with scale scores of 65 or 85, after rounding, are assigned those values.

References

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

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

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

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

Kuder, G. F. & Richardson, M. W. (1937) The theory of the estimation of test reliability. *Psychometrika*, *2*, 151–160.

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

Appendix A: Classical Item Analysis

Test	Form	Type	Item,	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	MEAN	Point-Biserial
2012_G4Sc_FT	501	MC	01	1	1,116	0.83	0.00		0.07	0.08	0.82	0.04			0.82	0.37
2012_G4Sc_FT	501	MC	02	1	1,116	0.83	0.00		0.09	0.86	0.02	0.03			0.86	0.38
2012_G4Sc_FT	501	MC	03	1	1,116	0.83	0.00		0.12	0.61	0.08	0.18			0.61	0.41
2012_G4Sc_FT	501	MC	04	1	1,116	0.83	0.00		0.25	0.61	0.07	0.07			0.61	0.43
2012_G4Sc_FT	501	MC	05	1	1,116	0.83	0.00		0.03	0.55	0.32	0.11			0.55	0.36
2012_G4Sc_FT	501	MC	06	1	1,116	0.83	0.01		0.01	0.02	0.84	0.11			0.84	0.45
2012_G4Sc_FT	501	MC	07	1	1,116	0.83	0.00		0.38	0.13	0.40	0.09			0.40	0.35
2012_G4Sc_FT	501	MC	08	1	1,116	0.83	0.00		0.28	0.68	0.02	0.02			0.68	0.35
2012_G4Sc_FT	501	MC	09	1	1,116	0.83	0.01		0.05	0.86	0.02	0.06			0.86	0.37
2012_G4Sc_FT	501	MC	10	1	1,116	0.83	0.01		0.87	0.04	0.04	0.04			0.87	0.50
2012_G4Sc_FT	501	MC	11	1	1,116	0.83	0.01		0.16	0.05	0.73	0.05			0.73	0.53
2012_G4Sc_FT	501	MC	12	1	1,116	0.83	0.02		0.83	0.08	0.04	0.03			0.83	0.43
2012_G4Sc_FT	501	MC	13	1	1,116	0.83	0.02		0.06	0.20	0.07	0.66			0.66	0.43
2012_G4Sc_FT	501	MC	14	1	1,116	0.83	0.02		0.63	0.10	0.19	0.07			0.63	0.52
2012_G4Sc_FT	501	MC	15	1	1,116	0.83	0.02		0.87	0.03	0.03	0.05			0.87	0.49
2012_G4Sc_FT	501	MC	16	1	1,116	0.83	0.02		0.06	0.04	0.06	0.82			0.82	0.53
2012_G4Sc_FT	501	MC	17	1	1,116	0.83	0.02		0.04	0.02	0.88	0.03			0.88	0.52
2012_G4Sc_FT	501	MC	18	1	1,116	0.83	0.02		0.80	0.05	0.05	0.09			0.80	0.47
2012_G4Sc_FT	501	MC	19	1	1,116	0.83	0.02		0.11	0.75	0.04	0.07			0.75	0.47
2012_G4Sc_FT	501	MC	20	1	1,116	0.83	0.02		0.01	0.10	0.76	0.12			0.76	0.43
2012_G4Sc_FT	501	MC	21	1	1,116	0.83	0.02		0.11	0.02	0.08	0.77			0.77	0.42
2012_G4Sc_FT	501	MC	22	1	1,116	0.83	0.02		0.01	0.91	0.02	0.03			0.91	0.44
2012_G4Sc_FT	501	MC	23	1	1,116	0.83	0.02		0.07	0.06	0.03	0.82			0.82	0.56

Test	Form	Type	Item,	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	MEAN	Point-Biserial
2012_G4Sc_FT	501	MC	24	1	1,116	0.83	0.03		0.14	0.63	0.08	0.12			0.63	0.44
2012_G4Sc_FT	501	MC	25	1	1,116	0.83	0.04		0.03	0.02	0.09	0.81			0.81	0.49
2012_G4Sc_FT	502	MC	01	1	1,138	0.68	0.00		0.04	0.77	0.06	0.13			0.77	0.46
2012_G4Sc_FT	502	MC	02	1	1,138	0.68	0.00		0.08	0.72	0.16	0.04			0.72	0.53
2012_G4Sc_FT	502	MC	03	1	1,138	0.68	0.00		0.10	0.05	0.77	0.09			0.77	0.39
2012_G4Sc_FT	502	MC	04	1	1,138	0.68	0.00		0.59	0.20	0.08	0.12			0.59	0.47
2012_G4Sc_FT	502	MC	05	1	1,138	0.68	0.00		0.02	0.04	0.08	0.86			0.86	0.37
2012_G4Sc_FT	502	MC	06	1	1,138	0.68	0.00		0.03	0.04	0.83	0.10			0.83	0.47
2012_G4Sc_FT	502	MC	07	1	1,138	0.68	0.00		0.04	0.15	0.09	0.71			0.71	0.38
2012_G4Sc_FT	502	MC	08	1	1,138	0.68	0.00		0.21	0.15	0.04	0.59			0.59	0.56
2012_G4Sc_FT	502	MC	09	1	1,138	0.68	0.03		0.72	0.14	0.09	0.03			0.72	0.44
2012_G4Sc_FT	502	CR	41	1	1,138	0.68	0.01	0.50	0.49						0.49	0.43
2012_G4Sc_FT	502	CR	42	1	1,138	0.68	0.04	0.12	0.85						0.85	0.45
2012_G4Sc_FT	502	CR	43	1	1,138	0.68	0.03	0.35	0.62						0.62	0.54
2012_G4Sc_FT	502	CR	44	1	1,138	0.68	0.03	0.10	0.86						0.86	0.43
2012_G4Sc_FT	503	MC	01	1	1,118	0.68	0.00		0.29	0.06	0.61	0.04			0.61	0.51
2012_G4Sc_FT	503	MC	02	1	1,118	0.68	0.00		0.09	0.78	0.07	0.06			0.78	0.46
2012_G4Sc_FT	503	MC	03	1	1,118	0.68	0.00		0.58	0.08	0.10	0.25			0.58	0.41
2012_G4Sc_FT	503	MC	04	1	1,118	0.68	0.00		0.10	0.66	0.03	0.21			0.66	0.52
2012_G4Sc_FT	503	MC	05	1	1,118	0.68	0.00		0.09	0.09	0.73	0.09			0.73	0.50
2012_G4Sc_FT	503	MC	06	1	1,118	0.68	0.00		0.04	0.02	0.08	0.85			0.85	0.45
2012_G4Sc_FT	503	MC	07	1	1,118	0.68	0.01		0.07	0.27	0.02	0.64			0.64	0.50
2012_G4Sc_FT	503	MC	08	1	1,118	0.68	0.02		0.78	0.06	0.02	0.12			0.78	0.51
2012_G4Sc_FT	503	CR	41													

Test	Form	Type	Item,	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	MEAN	Point-Biserial
2012_G4Sc_FT	503	CR	42	1	1,118	0.68	0.01	0.18	0.81						0.81	0.50
2012_G4Sc_FT	503	CR	43	1	1,118	0.68	0.02	0.62	0.36						0.36	0.54
2012_G4Sc_FT	503	CR	44	1	1,118	0.68	0.02	0.48	0.50						0.50	0.52
2012_G4Sc_FT	504	MC	01	1	1,127	0.64	0.00		0.85	0.02	0.10	0.03			0.85	0.40
2012_G4Sc_FT	504	MC	02	1	1,127	0.64	0.00		0.73	0.17	0.04	0.05			0.73	0.36
2012_G4Sc_FT	504	MC	03	1	1,127	0.64	0.00		0.02	0.07	0.09	0.81			0.81	0.40
2012_G4Sc_FT	504	MC	04	1	1,127	0.64	0.00		0.33	0.62	0.04	0.01			0.62	0.53
2012_G4Sc_FT	504	MC	05	1	1,127	0.64	0.00		0.83	0.11	0.04	0.01			0.83	0.51
2012_G4Sc_FT	504	MC	06	1	1,127	0.64	0.00		0.20	0.07	0.05	0.67			0.67	0.49
2012_G4Sc_FT	504	MC	07	1	1,127	0.64	0.01		0.21	0.70	0.04	0.03			0.70	0.59
2012_G4Sc_FT	504	MC	08	1	1,127	0.64	0.01		0.04	0.06	0.17	0.72			0.72	0.50
2012_G4Sc_FT	504	MC	09	1	1,127	0.64	0.02		0.02	0.03	0.08	0.86			0.86	0.38
2012_G4Sc_FT	504	MC	10	1	1,127	0.64	0.03		0.06	0.28	0.05	0.59			0.59	0.42
2012_G4Sc_FT	504	CR	41	1	1,127	0.64	0.00	0.21	0.79						0.79	0.37
2012_G4Sc_FT	504	CR	42	1	1,127	0.64	0.00	0.30	0.70						0.70	0.43
2012_G4Sc_FT	504	CR	43													
2012_G4Sc_FT	505	MC	01	1	1,147	0.55	0.00		0.25	0.06	0.08	0.61			0.61	0.47
2012_G4Sc_FT	505	MC	02	1	1,147	0.55	0.00		0.16	0.22	0.42	0.20			0.42	0.40
2012_G4Sc_FT	505	MC	03	1	1,147	0.55	0.00		0.04	0.03	0.05	0.87			0.87	0.44
2012_G4Sc_FT	505	MC	04	1	1,147	0.55	0.00		0.18	0.05	0.06	0.71			0.71	0.38
2012_G4Sc_FT	505	MC	05													
2012_G4Sc_FT	505	MC	06	1	1,147	0.55	0.00		0.03	0.88	0.06	0.02			0.88	0.38
2012_G4Sc_FT	505	MC	07	1	1,147	0.55	0.00		0.91	0.03	0.03	0.03			0.91	0.47
2012_G4Sc_FT	505	MC	08	1	1,147	0.55	0.02		0.37	0.08	0.51	0.02			0.51	0.41

Test	Form	Type	Item,	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	MEAN	Point-Biserial
2012_G4Sc_FT	505	CR	41	1	1,147	0.55	0.00	0.60	0.40						0.40	0.44
2012_G4Sc_FT	505	CR	42	1	1,147	0.55	0.00	0.38	0.62						0.62	0.49
2012_G4Sc_FT	505	CR	43	1	1,147	0.55	0.06	0.20	0.74						0.74	0.44
2012_G4Sc_FT	505	CR	44	1	1,147	0.55	0.03	0.15	0.81						0.81	0.46
2012_G4Sc_FT	506	MC	01	1	1,113	0.69	0.00		0.09	0.85	0.03	0.02			0.85	0.57
2012_G4Sc_FT	506	MC	02	1	1,113	0.69	0.00		0.03	0.04	0.03	0.90			0.90	0.56
2012_G4Sc_FT	506	MC	03	1	1,113	0.69	0.00		0.76	0.06	0.10	0.07			0.76	0.54
2012_G4Sc_FT	506	MC	04	1	1,113	0.69	0.01		0.05	0.24	0.63	0.08			0.63	0.43
2012_G4Sc_FT	506	MC	05	1	1,113	0.69	0.01		0.01	0.89	0.01	0.08			0.89	0.35
2012_G4Sc_FT	506	MC	06	1	1,113	0.69	0.02		0.79	0.06	0.11	0.02			0.79	0.53
2012_G4Sc_FT	506	CR	41	1	1,113	0.69	0.00	0.06	0.94						0.94	0.39
2012_G4Sc_FT	506	CR	42	1	1,113	0.69	0.01	0.27	0.71						0.71	0.56
2012_G4Sc_FT	506	CR	43	1	1,113	0.69	0.00	0.38	0.62						0.62	0.58
2012_G4Sc_FT	506	CR	44	1	1,113	0.69	0.02	0.32	0.66						0.66	0.51
2012_G4Sc_FT	506	CR	45	1	1,113	0.69	0.03	0.19	0.78						0.78	0.42
2012_G4Sc_FT	507	MC	01	1	1,135	0.59	0.00		0.19	0.07	0.10	0.64			0.64	0.44
2012_G4Sc_FT	507	MC	02	1	1,135	0.59	0.00		0.06	0.07	0.85	0.02			0.85	0.51
2012_G4Sc_FT	507	MC	03	1	1,135	0.59	0.00		0.01	0.10	0.04	0.84			0.84	0.46
2012_G4Sc_FT	507	MC	04	1	1,135	0.59	0.00		0.57	0.33	0.03	0.06			0.57	0.52
2012_G4Sc_FT	507	MC	05	1	1,135	0.59	0.00		0.13	0.01	0.80	0.07			0.80	0.37
2012_G4Sc_FT	507	MC	06	1	1,135	0.59	0.00		0.02	0.04	0.03	0.91			0.91	0.47
2012_G4Sc_FT	507	MC	07	1	1,135	0.59	0.01		0.58	0.04	0.06	0.31			0.58	0.51
2012_G4Sc_FT	507	CR	41	1	1,135	0.59	0.00	0.56	0.44						0.44	0.44
2012_G4Sc_FT	507	CR	42	1	1,135	0.59	0.01	0.27	0.72						0.72	0.56

Test	Form	Type	Item,	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	MEAN	Point-Biserial
2012_G4Sc_FT	507	CR	43	1	1,135	0.59	0.01	0.20	0.79						0.79	0.41
2012_G4Sc_FT	507	CR	44													
2012_G4Sc_FT	508	MC	01	1	1,128	0.62	0.00		0.06	0.08	0.05	0.81			0.81	0.40
2012_G4Sc_FT	508	MC	02	1	1,128	0.62	0.00		0.44	0.52	0.04	0.00			0.52	0.46
2012_G4Sc_FT	508	MC	03	1	1,128	0.62	0.00		0.01	0.31	0.49	0.18			0.49	0.49
2012_G4Sc_FT	508	MC	04	1	1,128	0.62	0.00		0.03	0.86	0.03	0.08			0.86	0.27
2012_G4Sc_FT	508	MC	05	1	1,128	0.62	0.00		0.02	0.01	0.03	0.94			0.94	0.36
2012_G4Sc_FT	508	MC	06	1	1,128	0.62	0.01		0.08	0.06	0.75	0.10			0.75	0.45
2012_G4Sc_FT	508	MC	07	1	1,128	0.62	0.02		0.08	0.68	0.02	0.20			0.68	0.55
2012_G4Sc_FT	508	CR	41	1	1,128	0.62	0.01	0.22	0.78						0.78	0.50
2012_G4Sc_FT	508	CR	42	1	1,128	0.62	0.01	0.20	0.79						0.79	0.50
2012_G4Sc_FT	508	CR	43	1	1,128	0.62	0.03	0.67	0.30						0.30	0.53
2012_G4Sc_FT	508	CR	44	1	1,128	0.62	0.02	0.27	0.71						0.71	0.47
2012_G4Sc_FT	509	MC	01	1	1,130	0.52	0.00		0.23	0.65	0.08	0.04			0.65	0.47
2012_G4Sc_FT	509	MC	02	1	1,130	0.52	0.00		0.07	0.70	0.21	0.01			0.70	0.37
2012_G4Sc_FT	509	MC	03	1	1,130	0.52	0.00		0.20	0.02	0.05	0.72			0.72	0.51
2012_G4Sc_FT	509	MC	04	1	1,130	0.52	0.00		0.03	0.33	0.54	0.10			0.54	0.34
2012_G4Sc_FT	509	MC	05	1	1,130	0.52	0.00		0.78	0.08	0.05	0.09			0.78	0.40
2012_G4Sc_FT	509	MC	06	1	1,130	0.52	0.01		0.28	0.50	0.12	0.09			0.50	0.48
2012_G4Sc_FT	509	CR	41	1	1,130	0.52	0.00	0.14	0.86						0.86	0.39
2012_G4Sc_FT	509	CR	42	1	1,130	0.52	0.01	0.46	0.53						0.53	0.54
2012_G4Sc_FT	509	CR	43	1	1,130	0.52	0.02	0.37	0.61						0.61	0.42
2012_G4Sc_FT	509	CR	44	1	1,130	0.52	0.02	0.20	0.78						0.78	0.41
2012_G4Sc_FT	510	MC	01	1	1,118	0.63	0.00		0.06	0.57	0.32	0.05			0.57	0.28

Test	Form	Type	Item,	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	MEAN	Point-Biserial
2012_G4Sc_FT	510	MC	02	1	1,118	0.63	0.00		0.76	0.18	0.06	0.00			0.76	0.40
2012_G4Sc_FT	510	MC	03	1	1,118	0.63	0.00		0.10	0.02	0.06	0.82			0.82	0.45
2012_G4Sc_FT	510	MC	04	1	1,118	0.63	0.00		0.07	0.13	0.68	0.11			0.68	0.60
2012_G4Sc_FT	510	MC	05	1	1,118	0.63	0.00		0.86	0.12	0.01	0.02			0.86	0.47
2012_G4Sc_FT	510	MC	06	1	1,118	0.63	0.00		0.05	0.92	0.02	0.01			0.92	0.47
2012_G4Sc_FT	510	MC	07	1	1,118	0.63	0.00		0.08	0.04	0.07	0.80			0.80	0.54
2012_G4Sc_FT	510	MC	08	1	1,118	0.63	0.01		0.90	0.03	0.02	0.05			0.90	0.51
2012_G4Sc_FT	510	CR	41	1	1,118	0.63	0.01	0.30	0.69						0.69	0.51
2012_G4Sc_FT	510	CR	42	1	1,118	0.63	0.03	0.48	0.50						0.50	0.51
2012_G4Sc_FT	510	CR	43	1	1,118	0.63	0.01	0.12	0.87						0.87	0.41
2012_G4Sc_FT	511	MC	01	1	1,120	0.60	0.00		0.03	0.06	0.73	0.18			0.73	0.52
2012_G4Sc_FT	511	MC	02	1	1,120	0.60	0.00		0.62	0.29	0.08	0.00			0.62	0.54
2012_G4Sc_FT	511	MC	03	1	1,120	0.60	0.00		0.02	0.02	0.46	0.50			0.46	0.52
2012_G4Sc_FT	511	MC	04	1	1,120	0.60	0.01		0.33	0.07	0.14	0.45			0.45	0.37
2012_G4Sc_FT	511	MC	05	1	1,120	0.60	0.00		0.04	0.93	0.01	0.01			0.93	0.37
2012_G4Sc_FT	511	MC	06	1	1,120	0.60	0.01		0.04	0.06	0.37	0.52			0.52	0.36
2012_G4Sc_FT	511	MC	07	1	1,120	0.60	0.01		0.07	0.63	0.25	0.04			0.63	0.34
2012_G4Sc_FT	511	MC	08	1	1,120	0.60	0.03		0.18	0.75	0.02	0.02			0.75	0.51
2012_G4Sc_FT	511	CR	41	1	1,120	0.60	0.00	0.18	0.81						0.81	0.46
2012_G4Sc_FT	511	CR	42	1	1,120	0.60	0.01	0.12	0.87						0.87	0.35
2012_G4Sc_FT	511	CR	43	1	1,120	0.60	0.01	0.35	0.64						0.64	0.44
2012_G4Sc_FT	511	CR	44	1	1,120	0.60	0.02	0.11	0.87						0.87	0.46
2012_G4Sc_FT	511	CR	45													
2012_G4Sc_FT	512	MC	01	1	1,115	0.60	0.00		0.08	0.15	0.68	0.09			0.68	0.56

Test	Form	Type	Item,	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	MEAN	Point-Biserial
2012_G4Sc_FT	512	MC	02	1	1,115	0.60	0.00		0.77	0.10	0.03	0.10			0.77	0.45
2012_G4Sc_FT	512	MC	03	1	1,115	0.60	0.00		0.82	0.02	0.04	0.12			0.82	0.45
2012_G4Sc_FT	512	MC	04	1	1,115	0.60	0.00		0.05	0.91	0.01	0.02			0.91	0.29
2012_G4Sc_FT	512	MC	05	1	1,115	0.60	0.00		0.03	0.05	0.90	0.02			0.90	0.39
2012_G4Sc_FT	512	MC	06	1	1,115	0.60	0.00		0.04	0.31	0.13	0.52			0.31	0.42
2012_G4Sc_FT	512	MC	07	1	1,115	0.60	0.00		0.52	0.21	0.15	0.12			0.52	0.36
2012_G4Sc_FT	512	MC	08	1	1,115	0.60	0.02		0.93	0.02	0.02	0.03			0.93	0.26
2012_G4Sc_FT	512	CR	41	1	1,115	0.60	0.01	0.39	0.60						0.60	0.51
2012_G4Sc_FT	512	CR	42	1	1,115	0.60	0.01	0.37	0.62						0.62	0.49
2012_G4Sc_FT	512	CR	43	1	1,115	0.60	0.03	0.23	0.74						0.74	0.54
2012_G4Sc_FT	512	CR	44	1	1,115	0.60	0.03	0.33	0.64						0.64	0.41
2012_G4Sc_FT	513	MC	01	1	1,089	0.66	0.00		0.70	0.10	0.03	0.16			0.70	0.50
2012_G4Sc_FT	513	MC	02	1	1,089	0.66	0.00		0.03	0.06	0.01	0.88			0.88	0.36
2012_G4Sc_FT	513	MC	03	1	1,089	0.66	0.00		0.02	0.68	0.11	0.18			0.68	0.47
2012_G4Sc_FT	513	MC	04	1	1,089	0.66	0.00		0.05	0.02	0.89	0.02			0.89	0.30
2012_G4Sc_FT	513	MC	05	1	1,089	0.66	0.00		0.72	0.16	0.11	0.00			0.72	0.48
2012_G4Sc_FT	513	MC	06	1	1,089	0.66	0.00		0.03	0.90	0.05	0.03			0.90	0.37
2012_G4Sc_FT	513	MC	07	1	1,089	0.66	0.00		0.22	0.04	0.08	0.66			0.66	0.59
2012_G4Sc_FT	513	MC	08	1	1,089	0.66	0.01		0.03	0.02	0.87	0.07			0.87	0.38
2012_G4Sc_FT	513	MC	09	1	1,089	0.66	0.02		0.60	0.10	0.10	0.18			0.60	0.51
2012_G4Sc_FT	513	CR	41	1	1,089	0.66	0.05	0.72	0.23						0.23	0.42
2012_G4Sc_FT	513	CR	42	1	1,089	0.66	0.01	0.34	0.65						0.65	0.54
2012_G4Sc_FT	513	CR	43	1	1,089	0.66	0.05	0.33	0.62						0.62	0.54
2012_G4Sc_FT	514	MC	01	1	1,096	0.58	0.00		0.20	0.49	0.13	0.18			0.49	0.44

Test	Form	Type	Item,	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	MEAN	Point-Biserial
2012_G4Sc_FT	514	MC	02	1	1,096	0.58	0.00		0.10	0.03	0.74	0.13			0.74	0.46
2012_G4Sc_FT	514	MC	03	1	1,096	0.58	0.00		0.91	0.06	0.01	0.02			0.91	0.41
2012_G4Sc_FT	514	MC	04	1	1,096	0.58	0.00		0.07	0.03	0.88	0.02			0.88	0.30
2012_G4Sc_FT	514	MC	05	1	1,096	0.58	0.00		0.05	0.04	0.06	0.84			0.84	0.40
2012_G4Sc_FT	514	MC	06	1	1,096	0.58	0.00		0.10	0.03	0.82	0.04			0.82	0.45
2012_G4Sc_FT	514	MC	07	1	1,096	0.58	0.00		0.17	0.09	0.54	0.19			0.54	0.52
2012_G4Sc_FT	514	MC	08	1	1,096	0.58	0.02		0.01	0.02	0.04	0.91			0.91	0.27
2012_G4Sc_FT	514	CR	41	1	1,096	0.58	0.00	0.22	0.77						0.77	0.39
2012_G4Sc_FT	514	CR	42	1	1,096	0.58	0.02	0.13	0.85						0.85	0.29
2012_G4Sc_FT	514	CR	43	1	1,096	0.58	0.03	0.19	0.78						0.78	0.47
2012_G4Sc_FT	514	CR	44	1	1,096	0.58	0.03	0.47	0.50						0.50	0.40
2012_G4Sc_FT	514	CR	45	1	1,096	0.58	0.04	0.56	0.40						0.40	0.45

Appendix B: Partial-Credit Model Item Analysis

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2012_G4Sc_FT	501	MC	01	1	1,116	−0.3239							1.08
2012_G4Sc_FT	501	MC	02	1	1,116	−0.7400							1.09
2012_G4Sc_FT	501	MC	03	1	1,116	0.9800							1.06
2012_G4Sc_FT	501	MC	04	1	1,116	0.9632							1.04
2012_G4Sc_FT	501	MC	05	1	1,116	1.0300							1.18
2012_G4Sc_FT	501	MC	06	1	1,116	−0.7700							1.16
2012_G4Sc_FT	501	MC	07	1	1,116	2.0900							1.12
2012_G4Sc_FT	501	MC	08	1	1,116	0.5600							1.15
2012_G4Sc_FT	501	MC	09	1	1,116	−0.7500							1.07
2012_G4Sc_FT	501	MC	10	1	1,116	−0.9100							0.94
2012_G4Sc_FT	501	MC	11	1	1,116	0.4100							0.90
2012_G4Sc_FT	501	MC	12	1	1,116	−0.4506							1.00
2012_G4Sc_FT	501	MC	13	1	1,116	0.8600							1.03
2012_G4Sc_FT	501	MC	14	1	1,116	1.1400							0.91
2012_G4Sc_FT	501	MC	15	1	1,116	−0.9000							0.94
2012_G4Sc_FT	501	MC	16	1	1,116	−0.0600							0.80
2012_G4Sc_FT	501	MC	17	1	1,116	−1.0600							0.94
2012_G4Sc_FT	501	MC	18	1	1,116	−0.2000							0.98
2012_G4Sc_FT	501	MC	19	1	1,116	0.0900							1.02
2012_G4Sc_FT	501	MC	20	1	1,116	0.1310							1.04
2012_G4Sc_FT	501	MC	21	1	1,116	0.2700							0.96
2012_G4Sc_FT	501	MC	22	1	1,116	−1.3091							0.91
2012_G4Sc_FT	501	MC	23	1	1,116	−0.5600							0.97
2012_G4Sc_FT	501	MC	24	1	1,116	0.9300							1.04

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2012_G4Sc_FT	501	MC	25	1	1,116	−0.3300							0.98
2012_G4Sc_FT	502	MC	01	1	1,138	0.1226							0.99
2012_G4Sc_FT	502	MC	02	1	1,138	0.4523							0.92
2012_G4Sc_FT	502	MC	03	1	1,138	0.1455							1.07
2012_G4Sc_FT	502	MC	04	1	1,138	1.1073							1.03
2012_G4Sc_FT	502	MC	05	1	1,138	−0.5652							1.01
2012_G4Sc_FT	502	MC	06	1	1,138	−0.2975							0.94
2012_G4Sc_FT	502	MC	07	1	1,138	0.4829							1.12
2012_G4Sc_FT	502	MC	08	1	1,138	1.1253							0.91
2012_G4Sc_FT	502	MC	09	1	1,138	0.4317							1.03
2012_G4Sc_FT	502	CR	41	1	1,138	1.6270							1.08
2012_G4Sc_FT	502	CR	42	1	1,138	−0.4488							0.93
2012_G4Sc_FT	502	CR	43	1	1,138	0.9896							0.94
2012_G4Sc_FT	502	CR	44	1	1,138	−0.6056							0.95
2012_G4Sc_FT	503	MC	01	1	1,118	0.9531							1.00
2012_G4Sc_FT	503	MC	02	1	1,118	−0.0118							1.02
2012_G4Sc_FT	503	MC	03	1	1,118	1.1330							1.16
2012_G4Sc_FT	503	MC	04	1	1,118	0.7200							0.98
2012_G4Sc_FT	503	MC	05	1	1,118	0.2723							0.99
2012_G4Sc_FT	503	MC	06	1	1,118	−0.6179							0.97
2012_G4Sc_FT	503	MC	07	1	1,118	0.8068							1.02
2012_G4Sc_FT	503	MC	08	1	1,118	−0.0552							0.95
2012_G4Sc_FT	503	CR	41										
2012_G4Sc_FT	503	CR	42	1	1,118	−0.2646							0.94

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2012_G4Sc_FT	503	CR	43	1	1,118	2.2622							0.93
2012_G4Sc_FT	503	CR	44	1	1,118	1.5129							1.00
2012_G4Sc_FT	504	MC	01	1	1,127	−0.3395							0.98
2012_G4Sc_FT	504	MC	02	1	1,127	0.4598							1.12
2012_G4Sc_FT	504	MC	03	1	1,127	−0.0697							1.01
2012_G4Sc_FT	504	MC	04	1	1,127	1.0645							0.96
2012_G4Sc_FT	504	MC	05	1	1,127	−0.2311							0.88
2012_G4Sc_FT	504	MC	06	1	1,127	0.8126							0.99
2012_G4Sc_FT	504	MC	07	1	1,127	0.6327							0.85
2012_G4Sc_FT	504	MC	08	1	1,127	0.5013							0.96
2012_G4Sc_FT	504	MC	09	1	1,127	−0.4470							1.00
2012_G4Sc_FT	504	MC	10	1	1,127	1.2146							1.10
2012_G4Sc_FT	504	CR	41	1	1,127	0.0980							1.07
2012_G4Sc_FT	504	CR	42	1	1,127	0.6526							1.05
2012_G4Sc_FT	504	CR	43										
2012_G4Sc_FT	505	MC	01	1	1,147	0.9109							0.99
2012_G4Sc_FT	505	MC	02	1	1,147	1.8256							1.07
2012_G4Sc_FT	505	MC	03	1	1,147	−0.7399							0.92
2012_G4Sc_FT	505	MC	04	1	1,147	0.4266							1.07
2012_G4Sc_FT	505	MC	05										
2012_G4Sc_FT	505	MC	06	1	1,147	−0.8007							0.98
2012_G4Sc_FT	505	MC	07	1	1,147	−1.1049							0.86
2012_G4Sc_FT	505	MC	08	1	1,147	1.3875							1.06
2012_G4Sc_FT	505	CR	41	1	1,147	1.9192							1.00

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2012_G4Sc_FT	505	CR	42	1	1,147	0.8767							0.96
2012_G4Sc_FT	505	CR	43	1	1,147	0.2385							0.99
2012_G4Sc_FT	505	CR	44	1	1,147	−0.2497							0.94
2012_G4Sc_FT	506	MC	01	1	1,113	−0.3179							0.86
2012_G4Sc_FT	506	MC	02	1	1,113	−0.8117							0.83
2012_G4Sc_FT	506	MC	03	1	1,113	0.3653							0.97
2012_G4Sc_FT	506	MC	04	1	1,113	1.1619							1.18
2012_G4Sc_FT	506	MC	05	1	1,113	−0.7790							1.11
2012_G4Sc_FT	506	MC	06	1	1,113	0.1479							0.96
2012_G4Sc_FT	506	CR	41	1	1,113	−1.4325							0.99
2012_G4Sc_FT	506	CR	42	1	1,113	0.6748							0.95
2012_G4Sc_FT	506	CR	43	1	1,113	1.2378							0.95
2012_G4Sc_FT	506	CR	44	1	1,113	0.9713							1.03
2012_G4Sc_FT	506	CR	45	1	1,113	0.2137							1.13
2012_G4Sc_FT	507	MC	01	1	1,135	0.8787							1.08
2012_G4Sc_FT	507	MC	02	1	1,135	−0.4452							0.88
2012_G4Sc_FT	507	MC	03	1	1,135	−0.3555							0.95
2012_G4Sc_FT	507	MC	04	1	1,135	1.2513							0.98
2012_G4Sc_FT	507	MC	05	1	1,135	−0.0339							1.11
2012_G4Sc_FT	507	MC	06	1	1,135	−1.0390							0.89
2012_G4Sc_FT	507	MC	07	1	1,135	1.1895							0.98
2012_G4Sc_FT	507	CR	41	1	1,135	1.9248							1.11
2012_G4Sc_FT	507	CR	42	1	1,135	0.4326							0.90
2012_G4Sc_FT	507	CR	43	1	1,135	0.0273							1.05

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2012_G4Sc_FT	507	CR	44										
2012_G4Sc_FT	508	MC	01	1	1,128	−0.1612							1.04
2012_G4Sc_FT	508	MC	02	1	1,128	1.4351							1.07
2012_G4Sc_FT	508	MC	03	1	1,128	1.5905							1.04
2012_G4Sc_FT	508	MC	04	1	1,128	−0.5770							1.15
2012_G4Sc_FT	508	MC	05	1	1,128	−1.5772							0.94
2012_G4Sc_FT	508	MC	06	1	1,128	0.2401							1.02
2012_G4Sc_FT	508	MC	07	1	1,128	0.6049							0.91
2012_G4Sc_FT	508	CR	41	1	1,128	0.0302							0.94
2012_G4Sc_FT	508	CR	42	1	1,128	−0.0727							0.91
2012_G4Sc_FT	508	CR	43	1	1,128	2.6446							0.93
2012_G4Sc_FT	508	CR	44	1	1,128	0.4506							1.00
2012_G4Sc_FT	509	MC	01	1	1,130	0.7678							0.97
2012_G4Sc_FT	509	MC	02	1	1,130	0.4968							1.07
2012_G4Sc_FT	509	MC	03	1	1,130	0.4063							0.90
2012_G4Sc_FT	509	MC	04	1	1,130	1.2976							1.14
2012_G4Sc_FT	509	MC	05	1	1,130	0.0772							1.00
2012_G4Sc_FT	509	MC	06	1	1,130	1.4871							0.99
2012_G4Sc_FT	509	CR	41	1	1,130	−0.4889							0.96
2012_G4Sc_FT	509	CR	42	1	1,130	1.3387							0.92
2012_G4Sc_FT	509	CR	43	1	1,130	0.9483							1.04
2012_G4Sc_FT	509	CR	44	1	1,130	0.0383							0.99
2012_G4Sc_FT	510	MC	01	1	1,118	1.3194							1.35
2012_G4Sc_FT	510	MC	02	1	1,118	0.2784							1.11

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2012_G4Sc_FT	510	MC	03	1	1,118	−0.1793							1.00
2012_G4Sc_FT	510	MC	04	1	1,118	0.7353							0.85
2012_G4Sc_FT	510	MC	05	1	1,118	−0.5011							0.94
2012_G4Sc_FT	510	MC	06	1	1,118	−1.2559							0.86
2012_G4Sc_FT	510	MC	07	1	1,118	−0.0427							0.91
2012_G4Sc_FT	510	MC	08	1	1,118	−0.9279							0.85
2012_G4Sc_FT	510	CR	41	1	1,118	0.6641							0.97
2012_G4Sc_FT	510	CR	42	1	1,118	1.7166							0.98
2012_G4Sc_FT	510	CR	43	1	1,118	−0.5957							1.01
2012_G4Sc_FT	511	MC	01	1	1,120	0.2940							0.91
2012_G4Sc_FT	511	MC	02	1	1,120	0.9054							0.91
2012_G4Sc_FT	511	MC	03	1	1,120	1.6954							0.93
2012_G4Sc_FT	511	MC	04	1	1,120	1.7301							1.14
2012_G4Sc_FT	511	MC	05	1	1,120	−1.4493							0.94
2012_G4Sc_FT	511	MC	06	1	1,120	1.3780							1.13
2012_G4Sc_FT	511	MC	07	1	1,120	0.8652							1.15
2012_G4Sc_FT	511	MC	08	1	1,120	0.1818							0.92
2012_G4Sc_FT	511	CR	41	1	1,120	−0.2102							0.94
2012_G4Sc_FT	511	CR	42	1	1,120	−0.6582							1.02
2012_G4Sc_FT	511	CR	43	1	1,120	0.8111							1.03
2012_G4Sc_FT	511	CR	44	1	1,120	−0.7093							0.89
2012_G4Sc_FT	511	CR	45										
2012_G4Sc_FT	512	MC	01	1	1,115	0.6027							0.88
2012_G4Sc_FT	512	MC	02	1	1,115	0.0793							0.97

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2012_G4Sc_FT	512	MC	03	1	1,115	−0.3021							0.96
2012_G4Sc_FT	512	MC	04	1	1,115	−1.1915							1.02
2012_G4Sc_FT	512	MC	05	1	1,115	−1.0225							0.93
2012_G4Sc_FT	512	MC	06	1	1,115	2.5155							1.03
2012_G4Sc_FT	512	MC	07	1	1,115	1.4043							1.16
2012_G4Sc_FT	512	MC	08	1	1,115	−1.3568							1.05
2012_G4Sc_FT	512	CR	41	1	1,115	1.0065							0.97
2012_G4Sc_FT	512	CR	42	1	1,115	0.9116							0.98
2012_G4Sc_FT	512	CR	43	1	1,115	0.2463							0.89
2012_G4Sc_FT	512	CR	44	1	1,115	0.7921							1.08
2012_G4Sc_FT	513	MC	01	1	1,089	0.4671							0.99
2012_G4Sc_FT	513	MC	02	1	1,089	−0.8855							1.00
2012_G4Sc_FT	513	MC	03	1	1,089	0.5720							1.03
2012_G4Sc_FT	513	MC	04	1	1,089	−0.9967							1.04
2012_G4Sc_FT	513	MC	05	1	1,089	0.3374							1.00
2012_G4Sc_FT	513	MC	06	1	1,089	−1.0072							0.98
2012_G4Sc_FT	513	MC	07	1	1,089	0.7097							0.88
2012_G4Sc_FT	513	MC	08	1	1,089	−0.7633							1.01
2012_G4Sc_FT	513	MC	09	1	1,089	1.0433							1.00
2012_G4Sc_FT	513	CR	41	1	1,089	3.2260							1.06
2012_G4Sc_FT	513	CR	42	1	1,089	0.7648							0.95
2012_G4Sc_FT	513	CR	43	1	1,089	0.9175							0.96
2012_G4Sc_FT	514	MC	01	1	1,096	1.5604							1.02
2012_G4Sc_FT	514	MC	02	1	1,096	0.2570							0.96

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2012_G4Sc_FT	514	MC	03	1	1,096	−1.0821							0.90
2012_G4Sc_FT	514	MC	04	1	1,096	−0.7285							1.04
2012_G4Sc_FT	514	MC	05	1	1,096	−0.3744							0.98
2012_G4Sc_FT	514	MC	06	1	1,096	−0.2617							0.94
2012_G4Sc_FT	514	MC	07	1	1,096	1.3099							0.93
2012_G4Sc_FT	514	MC	08	1	1,096	−1.1052							1.04
2012_G4Sc_FT	514	CR	41	1	1,096	0.0776							1.02
2012_G4Sc_FT	514	CR	42	1	1,096	−0.5103							1.07
2012_G4Sc_FT	514	CR	43	1	1,096	0.0007							0.93
2012_G4Sc_FT	514	CR	44	1	1,096	1.4811							1.08
2012_G4Sc_FT	514	CR	45	1	1,096	2.0081							1.02

Appendix C: DIF Statistics

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
501	01	MC	0.15	0.14	0.03		
501	02	MC	−0.38	0.66	−0.05		
501	03	MC	−0.55	2.61	−0.08		
501	04	MC	−0.22	0.42	−0.03		
501	05	MC	0.37	1.36	0.06		
501	06	MC	−0.64	2.00	−0.07		
501	07	MC	−0.82	6.27	−0.14		
501	08	MC	−0.13	0.15	−0.01		
501	09	MC	−0.12	0.07	−0.02		
501	10	MC	−0.09	0.03	0.00		
501	11	MC	−0.72	3.39	−0.11		
501	12	MC	0.61	1.99	0.09		
501	13	MC	−0.06	0.03	0.00		
501	14	MC	0.19	0.27	0.03		
501	15	MC	0.35	0.45	0.04		
501	16	MC	0.14	0.10	0.03		
501	17	MC	1.03	3.08	0.09		
501	18	MC	1.00	5.75	0.11		
501	19	MC	0.86	5.00	0.13		
501	20	MC	0.18	0.23	0.04		
501	21	MC	−0.59	2.34	−0.10		
501	22	MC	−0.60	0.90	−0.05		
501	23	MC	0.37	0.62	0.03		
501	24	MC	0.38	1.28	0.05		
501	25	MC	−0.05	0.01	−0.03		
502	01	MC	0.91	5.61	0.13		
502	02	MC	−0.36	0.93	−0.05		
502	03	MC	0.04	0.01	0.01		
502	04	MC	−0.78	5.53	−0.13		
502	05	MC	0.13	0.08	0.02		
502	06	MC	−0.15	0.13	−0.02		
502	07	MC	0.42	1.56	0.06		
502	08	MC	−0.30	0.70	−0.04		
502	09	MC	−0.36	1.02	−0.06		
502	41	OE		0.11	0.02		
502	42	OE		0.58	0.04		
502	43	OE		1.82	0.07		
502	44	OE		0.57	−0.05		
503	01	MC	1.15	11.25	0.17	B	Female
503	02	MC	−0.56	2.08	−0.09		
503	03	MC	−0.65	4.21	−0.12		
503	04	MC	0.24	0.46	0.04		
503	05	MC	−1.53	16.16	−0.22	C	Male
503	06	MC	0.51	1.18	0.06		
503	07	MC	−1.14	11.17	−0.18	B	Male
503	08	MC	0.94	5.32	0.12		
503	41	OE					
503	42	OE		11.33	0.19	BB	Female
503	43	OE		0.03	0.01		
503	44	OE		0.92	0.06		
504	01	MC	−0.13	0.09	−0.02		
504	02	MC	−0.97	7.89	−0.16		
504	03	MC	0.38	0.90	0.05		
504	04	MC	−0.35	1.06	−0.06		
504	05	MC	0.69	2.36	0.08		
504	06	MC	0.38	1.23	0.06		

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
504	07	MC	−0.39	1.06	−0.05		
504	08	MC	−0.47	1.61	−0.07		
504	09	MC	0.24	0.30	0.03		
504	10	MC	−0.03	0.01	0.00		
504	41	OE		1.61	0.07		
504	42	OE		0.09	0.02		
504	43	OE					
505	01	MC	−0.35	1.17	−0.06		
505	02	MC	0.74	5.47	0.13		
505	03	MC	1.11	4.94	0.12	B	Female
505	04	MC	−0.19	0.31	−0.03		
505	05	MC					
505	06	MC	−0.69	2.09	−0.08		
505	07	MC	−0.47	0.62	−0.04		
505	08	MC	−0.57	3.40	−0.09		
505	41	OE		5.39	−0.12		
505	42	OE		4.45	0.11		
505	43	OE		1.77	0.07		
505	44	OE		0.01	0.00		
506	01	MC	1.41	7.52	0.15	B	Female
506	02	MC	1.43	5.29	0.12	B	Female
506	03	MC	−0.26	0.43	−0.04		
506	04	MC	−0.22	0.42	−0.03		
506	05	MC	−0.59	1.35	−0.08		
506	06	MC	0.52	1.53	0.08		
506	41	OE		0.14	0.03		
506	42	OE		0.07	−0.02		
506	43	OE		0.69	−0.04		
506	44	OE		1.59	−0.08		
506	45	OE		0.02	−0.02		
507	01	MC	−1.01	9.23	−0.17	B	Male
507	02	MC	0.21	0.20	0.02		
507	03	MC	0.19	0.19	0.02		
507	04	MC	−0.29	0.72	−0.05		
507	05	MC	0.76	3.88	0.12		
507	06	MC	−0.80	1.93	−0.08		
507	07	MC	0.07	0.04	0.02		
507	41	OE		1.12	0.05		
507	42	OE		0.18	0.02		
507	43	OE		1.29	0.07		
507	44	OE					
508	01	MC	0.29	0.53	0.05		
508	02	MC	−0.71	4.80	−0.11		
508	03	MC	−0.68	4.28	−0.11		
508	04	MC	0.10	0.05	0.00		
508	05	MC	1.43	4.42	0.13	B	Female
508	06	MC	−0.11	0.09	−0.02		
508	07	MC	0.00	0.00	0.01		
508	41	OE		0.03	0.00		
508	42	OE		0.03	0.01		
508	43	OE		0.20	0.02		
508	44	OE		5.14	0.12		
509	01	MC	−0.16	0.23	−0.03		
509	02	MC	−0.05	0.03	−0.02		
509	03	MC	0.52	1.95	0.08		
509	04	MC	0.25	0.68	0.04		

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
509	05	MC	−0.51	1.88	−0.08		
509	06	MC	0.09	0.08	0.02		
509	41	OE		0.83	0.06		
509	42	OE		9.60	0.16		
509	43	OE		0.03	0.01		
509	44	OE		26.56	−0.29	CC	Male
510	01	MC	0.03	0.01	0.01		
510	02	MC	−0.50	1.81	−0.07		
510	03	MC	0.21	0.24	0.02		
510	04	MC	0.55	2.04	0.07		
510	05	MC	−0.09	0.04	−0.02		
510	06	MC	−0.25	0.15	−0.02		
510	07	MC	−0.18	0.17	−0.02		
510	08	MC	−1.01	3.01	−0.08		
510	41	OE		0.44	−0.04		
510	42	OE		1.57	0.06		
510	43	OE		0.58	0.04		
511	01	MC	0.12	0.10	0.02		
511	02	MC	−0.54	2.35	−0.08		
511	03	MC	−0.14	0.17	−0.02		
511	04	MC	−0.34	1.24	−0.06		
511	05	MC	−0.19	0.09	−0.02		
511	06	MC	0.09	0.09	0.02		
511	07	MC	−0.09	0.08	−0.01		
511	08	MC	−0.16	0.17	−0.03		
511	41	OE		2.09	0.08		
511	42	OE		1.51	0.06		
511	43	OE		0.00	0.00		

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
511	44	OE		3.85	0.11		
511	45	OE					
512	01	MC	−0.45	1.44	−0.06		
512	02	MC	0.17	0.20	0.02		
512	03	MC	0.46	1.23	0.06		
512	04	MC	−0.33	0.38	−0.03		
512	05	MC	−0.84	2.48	−0.09		
512	06	MC	−0.60	2.84	−0.09		
512	07	MC	0.99	10.27	0.19		
512	08	MC	0.49	0.79	0.05		
512	41	OE		1.15	0.06		
512	42	OE		0.08	0.02		
512	43	OE		6.71	−0.13		
512	44	OE		0.12	−0.03		
513	01	MC	0.16	0.19	0.04		
513	02	MC	1.43	7.83	0.17	B	Female
513	03	MC	−0.36	1.00	−0.06		
513	04	MC	1.00	3.95	0.12		
513	05	MC	−0.42	1.26	−0.06		
513	06	MC	0.05	0.01	0.02		
513	07	MC	−0.44	1.35	−0.07		
513	08	MC	−0.82	2.89	−0.09		
513	09	MC	0.21	0.35	0.03		
513	41	OE		1.65	−0.07		
513	42	OE		0.20	0.02		
513	43	OE		0.35	0.02		
514	01	MC	−0.42	1.67	−0.07		
514	02	MC	0.30	0.64	0.05		

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
514	03	MC	−1.23	4.58	−0.12	B	Male
514	04	MC	1.22	6.66	0.14	B	Female
514	05	MC	−0.08	0.03	−0.02		
514	06	MC	−1.54	13.24	−0.18	C	Male
514	07	MC	−0.57	2.71	−0.08		
514	08	MC	0.96	3.23	0.12		

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

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
514	41	OE		7.68	0.14		
514	42	OE		0.77	0.05		
514	43	OE		0.42	−0.02		
514	44	OE		2.57	0.09		
514	45	OE		0.73	−0.05		

Appendix D: Operational Test Map

June 2012

Position	Item Type	Max Points	Weight	Strand	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
1	MC	1	1	4	0.85	0.42	−0.51						
2	MC	1	1	4	0.41	0.53	2.07						
3	MC	1	1	4	0.81	0.41	−0.15						
4	MC	1	1	4	0.80	0.53	−0.07						
5	MC	1	1	4	0.79	0.32	−0.09						
6	MC	1	1	4	0.77	0.47	−0.09						
7	MC	1	1	4	0.89	0.42	−1.23						
8	MC	1	1	4	0.84	0.27	−0.50						
9	MC	1	1	4	0.54	0.45	1.15						
10	MC	1	1	4	0.61	0.53	1.02						
11	MC	1	1	4	0.66	0.50	0.61						
12	MC	1	1	4	0.49	0.47	1.62						
13	MC	1	1	4	0.68	0.43	0.70						
14	MC	1	1	4	0.93	0.42	−1.40						
15	MC	1	1	4	0.89	0.36	−0.94						
16	MC	1	1	4	0.59	0.48	0.95						
17	MC	1	1	4	0.90	0.46	−1.25						
18	MC	1	1	4	0.76	0.31	0.13						
19	MC	1	1	4	0.79	0.51	−0.27						
20	MC	1	1	4	0.78	0.44	0.04						
21	MC	1	1	4	0.89	0.40	−0.93						
22	MC	1	1	4	0.76	0.49	0.08						
23	MC	1	1	4	0.78	0.23	−0.28						

Position	Item Type	Max Points	Weight	Strand	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
24	MC	1	1	4	0.74	0.47	0.34						
25	MC	1	1	4	0.65	0.62	0.80						
26	MC	1	1	4	0.94	0.27	−2.17						
27	MC	1	1	6	0.53	0.31	1.09						
28	MC	1	1	6	0.86	0.42	−0.55						
29	MC	1	1	1	0.72	0.55	0.32						
30	MC	1	1	1	0.54	0.50	1.05						
31	CR	1	1	1	0.79	0.24	−0.28						
32	CR	1	1	1	0.90	0.24	−1.24						
33	CR	1	1	4	0.62	0.44	0.52						
34	CR	1	1	4	0.59	0.45	1.16						
35	CR	1	1	4	0.54	0.49	1.39						
36	CR	2	1	4	1.72	0.67	−0.44	0.49	−0.49				
37	CR	1	1	4	0.85	0.42	−0.52						
38	CR	2	1	4	1.12	0.61	1.21	−0.71	0.71				
39	CR	2	1	4	1.61	0.60	−0.65	−1.39	1.39				
40	CR	1	1	4	0.81	0.44	−0.49						
41	CR	1	1	4	0.81	0.45	−0.28						
42	CR	1	1	4	0.50	0.50	1.41						

Appendix E: Scoring Table

June 2012

Raw Score	Ability	Scale Score
0	−5.359	0.000
1	−4.118	1.683
2	−3.374	3.787
3	−2.919	5.757
4	−2.582	7.820
5	−2.310	9.970
6	−2.080	12.185
7	−1.879	14.466
8	−1.698	16.797
9	−1.534	19.175
10	−1.382	21.598
11	−1.240	24.067
12	−1.106	26.551
13	−0.979	29.061
14	−0.857	31.611
15	−0.739	34.171

Raw Score	Ability	Scale Score
16	−0.625	36.748
17	−0.514	39.327
18	−0.405	41.927
19	−0.297	44.526
20	−0.191	47.137
21	−0.086	49.735
22	0.019	52.313
23	0.124	54.900
24	0.230	57.462
25	0.337	59.996
26	0.444	62.485
27	0.553	64.958
28	0.664	67.374
29	0.778	69.761
30	0.894	72.089
31	1.014	74.364

Raw Score	Ability	Scale Score
32	1.138	76.584
33	1.268	78.742
34	1.403	80.840
35	1.546	82.870
36	1.699	84.842
37	1.863	86.741
38	2.042	88.584
39	2.242	90.356
40	2.469	92.073
41	2.736	93.727
42	3.068	95.327
43	3.517	96.877
44	4.253	98.397
45	5.487	100.000