New York State Regents Examination in English

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

Table of Contents

List of Tables

Section I: Introduction

Purpose

The purpose of this report is to document the psychometric work on the New York State Regents Examination in English 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 Regents Examination in English.

Section II: Field Test Analysis

In May 2010, field testing was conducted for the New York State Regents Examination in English 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 two entities. The multiple-choice items were scored by the New York State Education Department and the constructed-response items were scored by Measurement Incorporated. Therefore, it was necessary to combine data files for data analysis. Both classical and item response theory analyses were conducted using the data to evaluate the quality of the test items.

File Merging and Data Clean-up

Field test forms contained multiple-choice and constructed-response item types. Response data were contained in two separate files. The multiple-choice data file contained 12,909 student records and the constructed-response data file contained 5,803 student records. To combine the two files, the multiple-choice file served as the base file and constructed-response records were merged to the multiple-choice records using unique test booklet numbers. For multiple-choice records that did not have corresponding constructed-response records, constructed-response items were treated as non-attempted and scored as 0. After the exclusion rules were applied, the resulting field test data file contained 12,120 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 ten, item difficulties were below 0.90. With respect to the point-biserial correlations, all of these correlations were greater than or equal to 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
601	610	8	0	8	0	0	7	1
602	618	8	0	7	1	0	7	1
603	645	8	0	7	1	0	8	0
604	561	8	0	6	2	0	7	1
605	637	8	0	6	2	0	6	2
606	552	12	0	12	0	0	6	6
607	532	12	0	12	0	0	1	11
608	558	12	0	12	0	0	4	8
609	540	12	0	12	0	0	4	8
610	550	12	0	12	0	0	2	10
612	616	7	2	4	1	0	4	3
613	615	1	1	0	0	0	0	1
614	607	7	0	6	1	0	5	2
615	602	1	0	1	0	0	0	1
616	600	7	0	6	1	0	5	2
617	612	1	0	1	0	0	0	1
618	599	7	0	7	0	0	4	3
619	601	1	0	1	0	0	0	1
620	579	7	1	5	1	0	3	4
621	588	1	0	1	0	0	0	1
N3	11,822	10	3	7	0	0	6	4

In addition to the summary information provided in Table 2, all of the classical item statistics are provided in Appendix A. ‘Max’ is the maximum number of possible points. ‘N-Count’ refers to the number of student records in the analysis. ‘Alpha’ contains the internal consistency statistics discussed below. For multiple-choice items, ‘B’ represents the proportion of students who left the item blank and ‘M1’ through ‘M4’ are the proportions of students who selected each of the four answer choices. For constructed-response items, ‘B’ represents the proportion of students who left the item blank and ‘M0’ through ‘M6’ are the proportions of students who received scores 0 through 6. ‘Mean’ is the average of the scores received by the students. The final column contains the point-biserial correlation for each item. There are some instances of items missing statistics; this occurs when an item was not scored.

Test Reliability

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

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

Table 3 contains the internal consistency statistics for all of the field test forms. These statistics ranged from 0.67 to 0.81 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
601	0.74	n/a
602	0.72	n/a
603	0.70	n/a
604	0.75	n/a
605	0.74	n/a
606	0.80	n/a
607	0.81	n/a
608	0.80	n/a
609	0.80	n/a
610	0.81	n/a
612	0.79	0.73
613	0.71	0.82
614	0.77	0.66
615	0.67	0.78
616	0.78	0.69
617	0.72	0.82
618	0.76	0.72
619	0.71	0.81
620	0.80	0.76
621	0.69	0.76

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 0.66 to 0.82, 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. There were no instances of the first read and second read differing by more than 2.

Table 4. Point Differences Between First and Second Reads

			Difference (First Read minus Second Read)						
Form	**Item**	**Score Points**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**
612	6	2	0.00	0.00	0.10	0.79	0.11	0.00	0.00
612	7	2	0.00	0.00	0.10	0.82	0.08	0.00	0.00
613	Es	6	0.00	0.00	0.18	0.63	0.19	0.00	0.00
614	6	2	0.00	0.00	0.14	0.79	0.07	0.00	0.00
614	7	2	0.00	0.00	0.13	0.75	0.13	0.00	0.00
615	Es	6	0.00	0.00	0.21	0.64	0.14	0.01	0.00
616	6	2	0.00	0.00	0.08	0.84	0.08	0.00	0.00
616	7	2	0.00	0.00	0.15	0.77	0.07	0.01	0.00
617	Es	6	0.00	0.00	0.17	0.66	0.16	0.01	0.00
618	6	2	0.00	0.00	0.05	0.87	0.08	0.00	0.00
618	7	2	0.00	0.00	0.13	0.77	0.10	0.00	0.00
619	Es	6	0.00	0.00	0.22	0.64	0.13	0.01	0.00
620	6	2	0.00	0.00	0.10	0.80	0.10	0.00	0.00
620	7	2	0.00	0.00	0.04	0.90	0.06	0.00	0.00
621	Es	6	0.00	0.00	0.14	0.71	0.15	0.00	0.00

Table 5 contains additional summary information regarding the first and second reads. In the fifth 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 98.6% and 100%. These values indicate a high degree of agreement.

Table 5. First and Second Read Descriptive Statistics and Agreement

				Agreement (%)			Raw Score Mean		Raw Score Standard Deviation			
Form	Item	Score Points	Total N-Count	Exact	Adj.	Total	First Read	Second Read	First Read	Second Read	Intra-Class Correlation	Wt Kappa
612	6	2	82	79.3	20.7	100.0	1.1	1.1	0.65	0.60	0.74	0.67
612	7	2	73	82.2	17.8	100.0	1.2	1.2	0.57	0.55	0.71	0.67
613	Es	6	94	62.8	37.2	100.0	3.2	3.2	1.05	1.01	0.82	0.66
614	6	2	87	79.3	20.7	100.0	1.3	1.4	0.58	0.60	0.71	0.65
614	7	2	80	75.0	25.0	100.0	1.3	1.3	0.55	0.57	0.60	0.55
615	Es	6	92	64.1	34.8	98.9	3.5	3.5	0.95	0.95	0.78	0.63
616	6	2	86	83.7	16.3	100.0	1.4	1.4	0.61	0.60	0.78	0.73
616	7	2	74	77.0	21.6	98.6	1.4	1.4	0.54	0.58	0.57	0.57
617	Es	6	87	65.5	33.3	98.9	3.4	3.4	1.05	1.00	0.82	0.67
618	6	2	86	87.2	12.8	100.0	1.5	1.5	0.57	0.59	0.81	0.78
618	7	2	78	76.9	23.1	100.0	1.3	1.3	0.53	0.54	0.60	0.56
619	Es	6	87	64.4	34.5	98.9	3.4	3.5	0.96	1.03	0.81	0.65
620	6	2	81	80.2	19.8	100.0	1.3	1.3	0.67	0.61	0.76	0.70
620	7	2	77	89.6	10.4	100.0	1.2	1.2	0.45	0.46	0.75	0.73
621	Es	6	86	70.9	29.1	100.0	3.6	3.6	0.79	0.78	0.76	0.62

* Adj. = difference of one

Constructed-Response Item Means and Standard Deviations

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

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

Intra-class Correlation

The intra-class correlation was computed for each item. This correlation is an estimate of the reliability of scoring based on an average of the first and second reads. Correlations greater than 0.60 are considered very strong because they explain more than one-third of the variance in scores. All but three items had intra-class correlations greater than 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 0.55 to 0.78, 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. The majority of items fell within the moderate -2.0 to +2.0 difficulty range and only five items had an INFIT statistic outside the typical range.

Table 6. Partial Credit Model Item Analysis

			Rasch			INFIT		
Form	N-Count	No of Items	<-2.0	-2.0 to 2.0	>2.0	<-0.70	-0.70 to 1.30	>1.30
601	610	8	0	8	0	0	8	0
602	618	8	2	6	0	0	8	0
603	645	8	2	6	0	0	8	0
604	561	8	2	6	0	0	8	0
605	637	8	2	6	0	0	8	0
606	552	12	0	12	0	0	12	0
607	532	12	0	12	0	0	12	0
608	558	12	0	12	0	0	12	0
609	540	12	0	12	0	0	12	0
610	550	12	0	12	0	0	12	0
612	616	7	2	5	0	0	7	0
613	615	1	0	1	0	0	0	1
614	607	7	1	6	0	0	7	0
615	602	1	0	1	0	0	0	1
616	600	7	3	4	0	0	7	0
617	612	1	0	1	0	0	0	1
618	599	7	1	6	0	0	7	0
619	601	1	0	1	0	0	0	1
620	579	7	3	4	0	0	7	0
621	588	1	0	1	0	0	0	1
N3	11,822	10	0	10	0	0	10	0

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 Regents Examination in English consisted of 21 field test forms numbered 601–621 and an anchor form labeled N3. All students participating in the field test were administered the anchor form and one of the 21 field test forms. The field 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.

Each field test form was administered with the anchor form. The field test data were arranged in an incomplete data matrix so that the anchor items were in each data line along with the unique items for each field test form. Items not appearing on the field test form are left blank and treated as not administered when item parameters are calibrated. The entire data set was then calibrated using WINSTEPS and applying the Partial Credit Model. In this calibration, the anchor items were fixed to their 2009 bank values. This places all of the item parameters on the bank scale.

Table 7 is a sample matrix equating design for three of the forms where 'X' represents the presence of data and '—' represents the absence of data.

Table 7. Incomplete Data Matrix Structure

Anchor	Form 601	Form 602	Form 603
X	X	—	—
X	—	X	—
X	—	—	X

An item-stability check is performed on the anchor items by examining displacement values. The displacement values indicate the difference between the bank values for the anchor items and the difficulty values for those items as if they were not fixed to the bank values. 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 resulted in no items having a displacement value with a magnitude greater than 0.30. This indicates a strong level of stability in the items used on the anchor form.

The equated 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 2010 operational test maps with content information regarding each item included on the January 2010, June 2010, and August 2010 operational test forms.

In order to limit wide fluctuations of raw scores that correspond to scale scores of 65 and 85 across administrations, the average Rasch item difficulty for the operational test is considered. For this examination, an average Rasch difficulty of approximately 0.451 is used as a target for each administration. In most cases, meeting this target will provide raw scores of similar magnitude to other forms. However, differences with these scores also occur due to the distribution of the Rasch item difficulty parameters.

Scoring tables display the relationship between raw scores on the operational test and assigned scale scores. Appendix E contains the scoring tables used for January, June, and August 2010 operational test forms. Four steps are taken in order to produce these tables and resulting conversion charts.

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

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

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

The second step is to develop a raw score to theta relationship for the new operational test form using the field test equated PCM item parameters. This is accomplished by doing a calibration where all items are anchored to their field test parameters. One modification that is made is that for 6-point items, a constant based on existing bank values is used in place of the field test parameters. The number of points on the test form (i.e., raw score) expected across student ability levels is based on the difficulty of the items on the form. Thus, given a particular student ability level (i.e., theta), if the points are more difficult to earn on the new test than the points on the June 2004 test, the number of points expected of this student on the new test will be less than the number of points expected of this student on the baseline form.

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

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

References

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

Cohen, J. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–20.

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

Dorans, N.J. and Holland, P.W. 1992. DIF Detection and Description: Mantel–Haenszel and Standardization. In *Differential Item Functioning: Theory and Practice,* edited by P.W. Holland and H. Wainer, 35–66. Hillsdale, NJ: Erlbaum.

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

Appendix A: Classical Item Analysis

Table 8. Classical Item Analysis

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Biserial
2010_Engl_FT	601	MC	1	1	610	0.74	0.01	0.00	0.04	0.09	0.86	0.01			0.86	0.36
2010_Engl_FT	601	MC	2	1	610	0.74	0.01	0.00	0.82	0.13	0.02	0.02			0.82	0.45
2010_Engl_FT	601	MC	3	1	610	0.74	0.01	0.00	0.04	0.05	0.03	0.85			0.85	0.45
2010_Engl_FT	601	MC	4	1	610	0.74	0.01	0.00	0.07	0.73	0.03	0.15			0.73	0.56
2010_Engl_FT	601	MC	5	1	610	0.74	0.01	0.00	0.82	0.05	0.05	0.07			0.82	0.35
2010_Engl_FT	601	MC	6	1	610	0.74	0.02	0.00	0.04	0.31	0.61	0.02			0.61	0.47
2010_Engl_FT	601	MC	7	1	610	0.74	0.01	0.00	0.06	0.04	0.03	0.86			0.86	0.48
2010_Engl_FT	601	MC	8	1	610	0.74	0.01	0.00	0.02	0.79	0.04	0.14			0.79	0.50
2010_Engl_FT	602	MC	1	1	618	0.72	0.02	0.00	0.00	0.10	0.81	0.07			0.81	0.50
2010_Engl_FT	602	MC	2	1	618	0.72	0.01	0.00	0.78	0.07	0.03	0.11			0.78	0.45
2010_Engl_FT	602	MC	3	1	618	0.72	0.02	0.00	0.16	0.62	0.07	0.13			0.62	0.45
2010_Engl_FT	602	MC	4	1	618	0.72	0.02	0.00	0.03	0.07	0.04	0.85			0.85	0.53
2010_Engl_FT	602	MC	5	1	618	0.72	0.01	0.00	0.06	0.03	0.86	0.04			0.86	0.45
2010_Engl_FT	602	MC	6	1	618	0.72	0.02	0.00	0.20	0.02	0.10	0.66			0.66	0.47
2010_Engl_FT	602	MC	7	1	618	0.72	0.02	0.00	0.89	0.06	0.02	0.01			0.89	0.43
2010_Engl_FT	602	MC	8	1	618	0.72	0.02	0.00	0.02	0.93	0.02	0.02			0.93	0.50
2010_Engl_FT	603	MC	1	1	645	0.70	0.01	0.00	0.02	0.03	0.00	0.93			0.93	0.28
2010_Engl_FT	603	MC	2	1	645	0.70	0.02	0.00	0.04	0.85	0.05	0.03			0.85	0.41
2010_Engl_FT	603	MC	3	1	645	0.70	0.02	0.00	0.68	0.07	0.15	0.08			0.68	0.38
2010_Engl_FT	603	MC	4	1	645	0.70	0.01	0.00	0.02	0.09	0.85	0.02			0.85	0.34
2010_Engl_FT	603	MC	5	1	645	0.70	0.02	0.00	0.11	0.06	0.04	0.77			0.77	0.30

Table 8. Classical Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Biserial
2010_Engl_FT	603	MC	6	1	645	0.70	0.01	0.00	0.09	0.68	0.11	0.11			0.68	0.47
2010_Engl_FT	603	MC	7	1	645	0.70	0.01	0.00	0.86	0.02	0.09	0.02			0.86	0.39
2010_Engl_FT	603	MC	8	1	645	0.70	0.01	0.00	0.03	0.05	0.04	0.87			0.87	0.40
2010_Engl_FT	604	MC	1	1	561	0.75	0.01	0.00	0.68	0.01	0.04	0.26			0.68	0.34
2010_Engl_FT	604	MC	2	1	561	0.75	0.01	0.00	0.02	0.04	0.93	0.01			0.93	0.35
2010_Engl_FT	604	MC	3	1	561	0.75	0.01	0.00	0.05	0.02	0.14	0.78			0.78	0.33
2010_Engl_FT	604	MC	4	1	561	0.75	0.01	0.00	0.03	0.91	0.03	0.02			0.91	0.45
2010_Engl_FT	604	MC	5	1	561	0.75	0.01	0.00	0.85	0.09	0.03	0.02			0.85	0.51
2010_Engl_FT	604	MC	6	1	561	0.75	0.01	0.00	0.06	0.76	0.08	0.09			0.76	0.42
2010_Engl_FT	604	MC	7	1	561	0.75	0.01	0.00	0.17	0.15	0.04	0.62			0.62	0.46
2010_Engl_FT	604	MC	8	1	561	0.75	0.01	0.00	0.20	0.14	0.52	0.13			0.52	0.47
2010_Engl_FT	605	MC	1	1	637	0.74	0.01	0.00	0.06	0.82	0.05	0.05			0.82	0.49
2010_Engl_FT	605	MC	2	1	637	0.74	0.01	0.00	0.08	0.06	0.03	0.82			0.82	0.43
2010_Engl_FT	605	MC	3	1	637	0.74	0.00	0.00	0.06	0.03	0.87	0.03			0.87	0.51
2010_Engl_FT	605	MC	4	1	637	0.74	0.01	0.00	0.69	0.07	0.04	0.19			0.69	0.46
2010_Engl_FT	605	MC	5	1	637	0.74	0.01	0.00	0.03	0.03	0.92	0.02			0.92	0.49
2010_Engl_FT	605	MC	6	1	637	0.74	0.01	0.00	0.04	0.11	0.07	0.78			0.78	0.47
2010_Engl_FT	605	MC	7	1	637	0.74	0.01	0.00	0.05	0.05	0.83	0.07			0.83	0.51
2010_Engl_FT	605	MC	8	1	637	0.74	0.01	0.00	0.02	0.93	0.02	0.03			0.93	0.46
2010_Engl_FT	606	MC	11	1	552	0.80	0.00	0.00	0.06	0.75	0.15	0.03			0.75	0.48
2010_Engl_FT	606	MC	12	1	552	0.80	0.00	0.00	0.15	0.11	0.13	0.60			0.60	0.50
2010_Engl_FT	606	MC	13	1	552	0.80	0.01	0.00	0.71	0.10	0.11	0.08			0.71	0.47
2010_Engl_FT	606	MC	14	1	552	0.80	0.01	0.00	0.13	0.13	0.69	0.05			0.69	0.56

Table 8. Classical Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Biserial
2010_Engl_FT	606	MC	15	1	552	0.80	0.01	0.00	0.70	0.09	0.04	0.16			0.70	0.49
2010_Engl_FT	606	MC	16	1	552	0.80	0.01	0.00	0.08	0.51	0.13	0.28			0.51	0.44
2010_Engl_FT	606	MC	17	1	552	0.80	0.02	0.00	0.11	0.15	0.05	0.67			0.67	0.56
2010_Engl_FT	606	MC	18	1	552	0.80	0.03	0.00	0.15	0.63	0.15	0.05			0.63	0.52
2010_Engl_FT	606	MC	19	1	552	0.80	0.03	0.00	0.54	0.08	0.24	0.11			0.54	0.57
2010_Engl_FT	606	MC	20	1	552	0.80	0.04	0.00	0.16	0.08	0.09	0.63			0.63	0.46
2010_Engl_FT	606	MC	21	1	552	0.80	0.04	0.00	0.17	0.12	0.62	0.06			0.62	0.60
2010_Engl_FT	606	MC	22	1	552	0.80	0.04	0.00	0.61	0.17	0.12	0.05			0.61	0.62
2010_Engl_FT	607	MC	11	1	532	0.81	0.00	0.00	0.07	0.09	0.10	0.73			0.73	0.62
2010_Engl_FT	607	MC	12	1	532	0.81	0.01	0.00	0.09	0.73	0.12	0.06			0.73	0.52
2010_Engl_FT	607	MC	13	1	532	0.81	0.01	0.00	0.72	0.13	0.06	0.08			0.72	0.51
2010_Engl_FT	607	MC	14	1	532	0.81	0.01	0.00	0.07	0.09	0.78	0.05			0.78	0.57
2010_Engl_FT	607	MC	15	1	532	0.81	0.01	0.00	0.07	0.04	0.05	0.83			0.83	0.61
2010_Engl_FT	607	MC	16	1	532	0.81	0.01	0.00	0.63	0.09	0.07	0.20			0.63	0.53
2010_Engl_FT	607	MC	17	1	532	0.81	0.02	0.00	0.11	0.09	0.64	0.14			0.64	0.48
2010_Engl_FT	607	MC	18	1	532	0.81	0.02	0.00	0.71	0.08	0.08	0.10			0.71	0.56
2010_Engl_FT	607	MC	19	1	532	0.81	0.03	0.00	0.24	0.08	0.12	0.53			0.53	0.58
2010_Engl_FT	607	MC	20	1	532	0.81	0.04	0.00	0.12	0.61	0.08	0.15			0.61	0.51
2010_Engl_FT	607	MC	21	1	532	0.81	0.04	0.00	0.14	0.12	0.64	0.06			0.64	0.54
2010_Engl_FT	607	MC	22	1	532	0.81	0.04	0.00	0.06	0.68	0.12	0.10			0.68	0.55
2010_Engl_FT	608	MC	11	1	558	0.80	0.02	0.00	0.74	0.09	0.10	0.07			0.74	0.56
2010_Engl_FT	608	MC	12	1	558	0.80	0.00	0.00	0.05	0.08	0.76	0.11			0.76	0.59
2010_Engl_FT	608	MC	13	1	558	0.80	0.01	0.00	0.16	0.57	0.11	0.16			0.57	0.46

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_Engl_FT	608	MC	14	1	558	0.80	0.01	0.00	0.65	0.09	0.21	0.04			0.65	0.50
2010_Engl_FT	608	MC	15	1	558	0.80	0.01	0.00	0.10	0.06	0.17	0.67			0.67	0.50
2010_Engl_FT	608	MC	16	1	558	0.80	0.01	0.00	0.68	0.13	0.06	0.13			0.68	0.49
2010_Engl_FT	608	MC	17	1	558	0.80	0.02	0.00	0.09	0.05	0.09	0.75			0.75	0.58
2010_Engl_FT	608	MC	18	1	558	0.80	0.03	0.00	0.07	0.77	0.09	0.05			0.77	0.56
2010_Engl_FT	608	MC	19	1	558	0.80	0.04	0.00	0.08	0.11	0.06	0.71			0.71	0.54
2010_Engl_FT	608	MC	20	1	558	0.80	0.04	0.00	0.08	0.08	0.66	0.14			0.66	0.53
2010_Engl_FT	608	MC	21	1	558	0.80	0.04	0.00	0.79	0.05	0.06	0.06			0.79	0.60
2010_Engl_FT	608	MC	22	1	558	0.80	0.04	0.00	0.05	0.08	0.76	0.06			0.76	0.57
2010_Engl_FT	609	MC	11	1	540	0.80	0.00	0.00	0.71	0.17	0.06	0.05			0.71	0.55
2010_Engl_FT	609	MC	12	1	540	0.80	0.03	0.00	0.10	0.64	0.06	0.20			0.64	0.49
2010_Engl_FT	609	MC	13	1	540	0.80	0.03	0.00	0.10	0.11	0.08	0.71			0.71	0.59
2010_Engl_FT	609	MC	14	1	540	0.80	0.01	0.00	0.08	0.22	0.62	0.07			0.62	0.47
2010_Engl_FT	609	MC	15	1	540	0.80	0.00	0.00	0.10	0.08	0.12	0.70			0.70	0.53
2010_Engl_FT	609	MC	16	1	540	0.80	0.01	0.00	0.09	0.66	0.09	0.15			0.66	0.49
2010_Engl_FT	609	MC	17	1	540	0.80	0.03	0.00	0.10	0.09	0.11	0.67			0.67	0.60
2010_Engl_FT	609	MC	18	1	540	0.80	0.03	0.00	0.08	0.65	0.15	0.09			0.65	0.58
2010_Engl_FT	609	MC	19	1	540	0.80	0.04	0.00	0.75	0.07	0.07	0.07			0.75	0.59
2010_Engl_FT	609	MC	20	1	540	0.80	0.04	0.00	0.67	0.13	0.07	0.10			0.67	0.57
2010_Engl_FT	609	MC	21	1	540	0.80	0.05	0.00	0.09	0.12	0.11	0.63			0.63	0.58
2010_Engl_FT	609	MC	22	1	540	0.80	0.06	0.00	0.19	0.17	0.50	0.08			0.50	0.45
2010_Engl_FT	610	MC	11	1	550	0.81	0.00	0.00	0.74	0.11	0.09	0.05			0.74	0.53
2010_Engl_FT	610	MC	12	1	550	0.81	0.00	0.00	0.10	0.24	0.15	0.52			0.52	0.46

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_Engl_FT	610	MC	13	1	550	0.81	0.01	0.00	0.09	0.68	0.13	0.08			0.68	0.51
2010_Engl_FT	610	MC	14	1	550	0.81	0.01	0.00	0.10	0.14	0.66	0.09			0.66	0.57
2010_Engl_FT	610	MC	15	1	550	0.81	0.00	0.00	0.14	0.10	0.13	0.63			0.63	0.48
2010_Engl_FT	610	MC	16	1	550	0.81	0.01	0.00	0.80	0.07	0.05	0.06			0.80	0.57
2010_Engl_FT	610	MC	17	1	550	0.81	0.03	0.00	0.07	0.13	0.70	0.07			0.70	0.56
2010_Engl_FT	610	MC	18	1	550	0.81	0.03	0.00	0.56	0.17	0.18	0.06			0.56	0.53
2010_Engl_FT	610	MC	19	1	550	0.81	0.03	0.00	0.58	0.22	0.06	0.11			0.58	0.56
2010_Engl_FT	610	MC	20	1	550	0.81	0.03	0.00	0.09	0.66	0.08	0.14			0.66	0.62
2010_Engl_FT	610	MC	21	1	550	0.81	0.06	0.00	0.15	0.06	0.11	0.61			0.61	0.54
2010_Engl_FT	610	MC	22	1	550	0.81	0.07	0.00	0.04	0.07	0.77	0.05			0.77	0.56
2010_Engl_FT	612	MC	1	1	616	0.79	0.01	0.00	0.10	0.74	0.05	0.10			0.74	0.47
2010_Engl_FT	612	MC	2	1	616	0.79	0.01	0.00	0.16	0.05	0.77	0.01			0.77	0.55
2010_Engl_FT	612	MC	3	1	616	0.79	0.01	0.00	0.76	0.08	0.13	0.03			0.76	0.48
2010_Engl_FT	612	MC	4	1	616	0.79	0.01	0.00	0.02	0.93	0.02	0.02			0.93	0.45
2010_Engl_FT	612	MC	5	1	616	0.79	0.01	0.00	0.02	0.04	0.04	0.89			0.89	0.47
2010_Engl_FT	612	CR	6	2	616	0.79	0.11	0.13	0.50	0.23					0.97	0.61
2010_Engl_FT	612	CR	7	2	616	0.79	0.17	0.05	0.48	0.23					0.95	0.64
2010_Engl_FT	613	Essay	Essay	6	615	0.71	0.09	0.00	0.07	0.11	0.32	0.34	0.07	0.00	2.93	0.68
2010_Engl_FT	614	MC	1	1	607	0.77	0.01	0.00	0.03	0.02	0.01	0.94			0.94	0.41
2010_Engl_FT	614	MC	2	1	607	0.77	0.01	0.00	0.78	0.06	0.13	0.01			0.78	0.38
2010_Engl_FT	614	MC	3	1	607	0.77	0.01	0.00	0.03	0.07	0.05	0.84			0.84	0.50
2010_Engl_FT	614	MC	4	1	607	0.77	0.01	0.00	0.17	0.18	0.50	0.14			0.50	0.45
2010_Engl_FT	614	MC	5	1	607	0.77	0.01	0.00	0.05	0.80	0.11	0.04			0.80	0.36

Table 8. Classical Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Biserial
2010_Engl_FT	614	CR	6	2	607	0.77	0.13	0.07	0.41	0.39					1.19	0.58
2010_Engl_FT	614	CR	7	2	607	0.77	0.22	0.04	0.39	0.35					1.09	0.64
2010_Engl_FT	615	Essay	Essay	6	602	0.67	0.10	0.00	0.07	0.11	0.29	0.38	0.07	0.00	3.00	0.66
2010_Engl_FT	616	MC	1	1	600	0.78	0.01	0.00	0.01	0.03	0.05	0.90			0.90	0.44
2010_Engl_FT	616	MC	2	1	600	0.78	0.01	0.00	0.92	0.04	0.02	0.01			0.92	0.40
2010_Engl_FT	616	MC	3	1	600	0.78	0.01	0.00	0.03	0.02	0.90	0.04			0.90	0.45
2010_Engl_FT	616	MC	4	1	600	0.78	0.01	0.00	0.08	0.02	0.05	0.83			0.83	0.44
2010_Engl_FT	616	MC	5	1	600	0.78	0.01	0.00	0.02	0.59	0.30	0.06			0.59	0.39
2010_Engl_FT	616	CR	6	2	600	0.78	0.11	0.07	0.37	0.45					1.26	0.59
2010_Engl_FT	616	CR	7	2	600	0.78	0.23	0.03	0.44	0.30					1.04	0.62
2010_Engl_FT	617	Essay	Essay	6	612	0.72	0.11	0.00	0.06	0.09	0.26	0.37	0.10	0.01	3.06	0.72
2010_Engl_FT	618	MC	1	1	599	0.76	0.01	0.00	0.02	0.14	0.78	0.05			0.78	0.29
2010_Engl_FT	618	MC	2	1	599	0.76	0.01	0.00	0.06	0.89	0.01	0.03			0.89	0.32
2010_Engl_FT	618	MC	3	1	599	0.76	0.01	0.00	0.69	0.10	0.03	0.17			0.69	0.34
2010_Engl_FT	618	MC	4	1	599	0.76	0.01	0.00	0.19	0.07	0.10	0.64			0.64	0.51
2010_Engl_FT	618	MC	5	1	599	0.76	0.01	0.00	0.75	0.15	0.08	0.02			0.75	0.37
2010_Engl_FT	618	CR	6	2	599	0.76	0.10	0.05	0.42	0.43					1.27	0.60
2010_Engl_FT	618	CR	7	2	599	0.76	0.18	0.05	0.45	0.33					1.10	0.62
2010_Engl_FT	619	Essay	Essay	6	601	0.71	0.10	0.00	0.04	0.11	0.29	0.37	0.09	0.01	3.07	0.66
2010_Engl_FT	620	MC	1	1	579	0.80	0.01	0.00	0.08	0.79	0.06	0.06			0.79	0.45
2010_Engl_FT	620	MC	2	1	579	0.80	0.01	0.00	0.12	0.02	0.05	0.80			0.80	0.55
2010_Engl_FT	620	MC	3	1	579	0.80	0.01	0.00	0.91	0.03	0.02	0.03			0.91	0.53

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_Engl_FT	620	MC	4	1	579	0.80	0.01	0.00	0.04	0.90	0.02	0.02			0.90	0.48
2010_Engl_FT	620	MC	5	1	579	0.80	0.01	0.00	0.03	0.02	0.89	0.04			0.89	0.50
2010_Engl_FT	620	CR	6	2	579	0.80	0.18	0.06	0.46	0.31					1.07	0.64
2010_Engl_FT	620	CR	7	2	579	0.80	0.24	0.03	0.51	0.22					0.95	0.65
2010_Engl_FT	621	Essay	Essay	6	588	0.69	0.12	0.00	0.04	0.08	0.28	0.43	0.07	0.01	3.11	0.69
2010_Engl_FT	N3	MC	1	1	11,822		0.03	0.00	0.16	0.04	0.05	0.72				0.45
2010_Engl_FT	N3	MC	2	1	11,822		0.03	0.00	0.24	0.22	0.44	0.07				0.39
2010_Engl_FT	N3	MC	3	1	11,822		0.03	0.00	0.05	0.81	0.03	0.07				0.51
2010_Engl_FT	N3	MC	4	1	11,822		0.03	0.00	0.69	0.14	0.06	0.08				0.45
2010_Engl_FT	N3	MC	5	1	11,822		0.04	0.00	0.07	0.65	0.19	0.06				0.45
2010_Engl_FT	N3	MC	6	1	11,822		0.05	0.00	0.09	0.17	0.23	0.46				0.50
2010_Engl_FT	N3	MC	7	1	11,822		0.05	0.00	0.59	0.13	0.13	0.10				0.54
2010_Engl_FT	N3	MC	8	1	11,822		0.05	0.00	0.24	0.10	0.13	0.48				0.44
2010_Engl_FT	N3	MC	9	1	11,822		0.05	0.00	0.20	0.60	0.05	0.11				0.53
2010_Engl_FT	N3	MC	10	1	11,822		0.06	0.00	0.17	0.09	0.54	0.16				0.56

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_Engl_FT	601	MC	1	1	610	-1.93							0.86
2010_Engl_FT	601	MC	2	1	610	-1.57							0.82
2010_Engl_FT	601	MC	3	1	610	-1.91							0.85
2010_Engl_FT	601	MC	4	1	610	-0.92							0.73
2010_Engl_FT	601	MC	5	1	610	-1.57							0.82
2010_Engl_FT	601	MC	6	1	610	-0.17							0.61
2010_Engl_FT	601	MC	7	1	610	-1.96							0.86
2010_Engl_FT	601	MC	8	1	610	-1.32							0.79
2010_Engl_FT	602	MC	1	1	618	-1.52							0.81
2010_Engl_FT	602	MC	2	1	618	-1.25							0.78
2010_Engl_FT	602	MC	3	1	618	-0.20							0.62
2010_Engl_FT	602	MC	4	1	618	-1.83							0.85
2010_Engl_FT	602	MC	5	1	618	-1.98							0.86
2010_Engl_FT	602	MC	6	1	618	-0.40							0.66
2010_Engl_FT	602	MC	7	1	618	-2.36							0.89
2010_Engl_FT	602	MC	8	1	618	-3.08							0.93
2010_Engl_FT	603	MC	1	1	645	-2.88							0.93
2010_Engl_FT	603	MC	2	1	645	-1.82							0.85
2010_Engl_FT	603	MC	3	1	645	-0.59							0.68
2010_Engl_FT	603	MC	4	1	645	-1.87							0.85
2010_Engl_FT	603	MC	5	1	645	-1.15							0.77
2010_Engl_FT	603	MC	6	1	645	-0.60							0.68

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

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2010_Engl_FT	603	MC	7	1	645	-1.93							0.86
2010_Engl_FT	603	MC	8	1	645	-2.07							0.87
2010_Engl_FT	604	MC	1	1	561	-0.56							0.68
2010_Engl_FT	604	MC	2	1	561	-3.03							0.93
2010_Engl_FT	604	MC	3	1	561	-1.20							0.78
2010_Engl_FT	604	MC	4	1	561	-2.68							0.91
2010_Engl_FT	604	MC	5	1	561	-1.82							0.85
2010_Engl_FT	604	MC	6	1	561	-1.09							0.76
2010_Engl_FT	604	MC	7	1	561	-0.21							0.62
2010_Engl_FT	604	MC	8	1	561	0.37							0.52
2010_Engl_FT	605	MC	1	1	637	-1.48							0.82
2010_Engl_FT	605	MC	2	1	637	-1.45							0.82
2010_Engl_FT	605	MC	3	1	637	-1.95							0.87
2010_Engl_FT	605	MC	4	1	637	-0.42							0.69
2010_Engl_FT	605	MC	5	1	637	-2.65							0.92
2010_Engl_FT	605	MC	6	1	637	-1.04							0.78
2010_Engl_FT	605	MC	7	1	637	-1.57							0.83
2010_Engl_FT	605	MC	8	1	637	-2.84							0.93
2010_Engl_FT	606	MC	11	1	552	-0.68							0.73
2010_Engl_FT	606	MC	12	1	552	0.22							0.58
2010_Engl_FT	606	MC	13	1	552	-0.41							0.69
2010_Engl_FT	606	MC	14	1	552	-0.28							0.67
2010_Engl_FT	606	MC	15	1	552	-0.38							0.69

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

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2010_Engl_FT	606	MC	16	1	552	0.70							0.49
2010_Engl_FT	606	MC	17	1	552	-0.17							0.65
2010_Engl_FT	606	MC	18	1	552	0.02							0.62
2010_Engl_FT	606	MC	19	1	552	0.53							0.53
2010_Engl_FT	606	MC	20	1	552	0.02							0.62
2010_Engl_FT	606	MC	21	1	552	0.12							0.60
2010_Engl_FT	606	MC	22	1	552	0.13							0.60
2010_Engl_FT	607	MC	11	1	532	-0.61							0.72
2010_Engl_FT	607	MC	12	1	532	-0.58							0.71
2010_Engl_FT	607	MC	13	1	532	-0.55							0.71
2010_Engl_FT	607	MC	14	1	532	-0.96							0.76
2010_Engl_FT	607	MC	15	1	532	-1.33							0.81
2010_Engl_FT	607	MC	16	1	532	-0.01							0.62
2010_Engl_FT	607	MC	17	1	532	-0.04							0.62
2010_Engl_FT	607	MC	18	1	532	-0.51							0.70
2010_Engl_FT	607	MC	19	1	532	0.55							0.52
2010_Engl_FT	607	MC	20	1	532	0.12							0.60
2010_Engl_FT	607	MC	21	1	532	-0.08							0.63
2010_Engl_FT	607	MC	22	1	532	-0.32							0.67
2010_Engl_FT	608	MC	11	1	558	-0.75							0.73
2010_Engl_FT	608	MC	12	1	558	-0.88							0.74
2010_Engl_FT	608	MC	13	1	558	0.28							0.56
2010_Engl_FT	608	MC	14	1	558	-0.19							0.64
2010_Engl_FT	608	MC	15	1	558	-0.32							0.66

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

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2010_Engl_FT	608	MC	16	1	558	-0.37							0.67
2010_Engl_FT	608	MC	17	1	558	-0.80							0.73
2010_Engl_FT	608	MC	18	1	558	-0.95							0.75
2010_Engl_FT	608	MC	19	1	558	-0.58							0.70
2010_Engl_FT	608	MC	20	1	558	-0.25							0.65
2010_Engl_FT	608	MC	21	1	558	-1.10							0.78
2010_Engl_FT	608	MC	22	1	558	-0.90							0.75
2010_Engl_FT	609	MC	11	1	540	-0.54							0.69
2010_Engl_FT	609	MC	12	1	540	-0.12							0.62
2010_Engl_FT	609	MC	13	1	540	-0.55							0.70
2010_Engl_FT	609	MC	14	1	540	-0.01							0.61
2010_Engl_FT	609	MC	15	1	540	-0.44							0.68
2010_Engl_FT	609	MC	16	1	540	-0.23							0.64
2010_Engl_FT	609	MC	17	1	540	-0.27							0.65
2010_Engl_FT	609	MC	18	1	540	-0.15							0.63
2010_Engl_FT	609	MC	19	1	540	-0.79							0.73
2010_Engl_FT	609	MC	20	1	540	-0.27							0.65
2010_Engl_FT	609	MC	21	1	540	-0.05							0.61
2010_Engl_FT	609	MC	22	1	540	0.64							0.49
2010_Engl_FT	610	MC	11	1	550	-0.72							0.73
2010_Engl_FT	610	MC	12	1	550	0.58							0.51
2010_Engl_FT	610	MC	13	1	550	-0.34							0.67
2010_Engl_FT	610	MC	14	1	550	-0.21							0.65
2010_Engl_FT	610	MC	15	1	550	-0.02							0.62

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

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2010_Engl_FT	610	MC	16	1	550	-1.13							0.79
2010_Engl_FT	610	MC	17	1	550	-0.44							0.69
2010_Engl_FT	610	MC	18	1	550	0.34							0.55
2010_Engl_FT	610	MC	19	1	550	0.27							0.56
2010_Engl_FT	610	MC	20	1	550	-0.19							0.64
2010_Engl_FT	610	MC	21	1	550	0.07							0.60
2010_Engl_FT	610	MC	22	1	550	-0.93							0.76
2010_Engl_FT	612	MC	1	1	616	-0.92							0.74
2010_Engl_FT	612	MC	2	1	616	-1.10							0.77
2010_Engl_FT	612	MC	3	1	616	-1.05							0.76
2010_Engl_FT	612	MC	4	1	616	-2.99							0.93
2010_Engl_FT	612	MC	5	1	616	-2.27							0.89
2010_Engl_FT	612	CR	6	2	616	0.66	-1.34	1.34					0.97
2010_Engl_FT	612	CR	7	2	616	0.70	-1.26	1.26					0.95
2010_Engl_FT	613	Essay	Essay	6	615	0.61	-1.21	-1.32	-1.38	0.45	3.46		2.93
2010_Engl_FT	614	MC	1	1	607	-2.96							0.94
2010_Engl_FT	614	MC	2	1	607	-1.20							0.78
2010_Engl_FT	614	MC	3	1	607	-1.67							0.84
2010_Engl_FT	614	MC	4	1	607	0.52							0.50
2010_Engl_FT	614	MC	5	1	607	-1.33							0.80
2010_Engl_FT	614	CR	6	2	607	-0.02	-0.94	0.94					1.19
2010_Engl_FT	614	CR	7	2	607	0.27	-0.80	0.80					1.09
2010_Engl_FT	615	Essay	Essay	6	602	0.58	-0.89	-1.24	-1.29	0.04	3.38		3.00

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

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2010_Engl_FT	616	MC	1	1	600	-2.48							0.90
2010_Engl_FT	616	MC	2	1	600	-2.73							0.92
2010_Engl_FT	616	MC	3	1	600	-2.39							0.90
2010_Engl_FT	616	MC	4	1	600	-1.66							0.83
2010_Engl_FT	616	MC	5	1	600	-0.03							0.59
2010_Engl_FT	616	CR	6	2	600	-0.27	-0.86	0.86					1.26
2010_Engl_FT	616	CR	7	2	600	0.38	-1.05	1.05					1.04
2010_Engl_FT	617	Essay	Essay	6	612	1.40	-1.37	-2.05	-2.23	-0.91	1.72	4.84	3.06
2010_Engl_FT	618	MC	1	1	599	-1.07							0.78
2010_Engl_FT	618	MC	2	1	599	-2.06							0.89
2010_Engl_FT	618	MC	3	1	599	-0.42							0.69
2010_Engl_FT	618	MC	4	1	599	-0.12							0.64
2010_Engl_FT	618	MC	5	1	599	-0.79							0.75
2010_Engl_FT	618	CR	6	2	599	-0.18	-1.07	1.07					1.27
2010_Engl_FT	618	CR	7	2	599	0.36	-1.09	1.09					1.10
2010_Engl_FT	619	Essay	Essay	6	601	1.67	-1.73	-3.00	-2.44	-1.03	1.68	6.51	3.07
2010_Engl_FT	620	MC	1	1	579	-1.47							0.79
2010_Engl_FT	620	MC	2	1	579	-1.55							0.80
2010_Engl_FT	620	MC	3	1	579	-2.79							0.90
2010_Engl_FT	620	MC	4	1	579	-2.66							0.90
2010_Engl_FT	620	MC	5	1	579	-2.56							0.89
2010_Engl_FT	620	CR	6	2	579	0.20	-1.22	1.22					1.07
2010_Engl_FT	620	CR	7	2	579	0.62	-1.48	1.48					0.95

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

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2010_Engl_FT	621	Essay	Essay	6	588	0.43	-0.38	-1.60	-1.47	-0.02	3.47		3.11
2010_Engl_FT	N3	MC	1	1	1,182	-0.71							1.01
2010_Engl_FT	N3	MC	2	1	1,182	1.04							1.20
2010_Engl_FT	N3	MC	3	1	1,182	-1.33							0.85
2010_Engl_FT	N3	MC	4	1	1,182	-0.25							0.98
2010_Engl_FT	N3	MC	5	1	1,182	-0.19							1.04
2010_Engl_FT	N3	MC	6	1	1,182	0.79							1.02
2010_Engl_FT	N3	MC	7	1	1,182	-0.21							0.98
2010_Engl_FT	N3	MC	8	1	1,182	0.69							1.12
2010_Engl_FT	N3	MC	9	1	1,182	-0.12							0.97
2010_Engl_FT	N3	MC	10	1	1,182	0.29							0.91

Appendix C: DIF Statistics

Table 10. DIF Statistics

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
601	1	MC	-0.51	0.80	-0.08		
601	2	MC	0.56	1.17	0.08		
601	3	MC	-0.47	0.71	-0.05		
601	4	MC	0.26	0.31	0.04		
601	5	MC	0.36	0.52	0.07		
601	6	MC	0.92	5.28	0.18		
601	7	MC	0.26	0.21	0.04		
601	8	MC	0.92	3.49	0.14		
602	1	MC	-0.30	0.33	-0.05		
602	2	MC	0.25	0.28	0.02		
602	3	MC	-0.10	0.06	-0.02		
602	4	MC	0.55	0.90	0.05		
602	5	MC	2.13	13.37	0.28	C	F
602	6	MC	0.98	5.51	0.17		
602	7	MC	0.33	0.25	0.03		
602	8	MC	0.67	0.64	0.06		
603	1	MC	1.05	1.77	0.09		
603	2	MC	0.75	1.86	0.09		
603	3	MC	-1.39	11.32	-0.27	B	M
603	4	MC	-0.48	0.77	-0.08		

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
603	5	MC	0.07	0.03	0.03		
603	6	MC	-0.68	2.60	-0.12		
603	7	MC	-1.49	6.80	-0.19	B	M
603	8	MC	0.61	1.09	0.07		
604	1	MC	-1.43	10.22	-0.26	B	M
604	2	MC	-2.45	7.34	-0.22		
604	3	MC	0.16	0.10	0.01		
604	4	MC	-1.78	4.84	-0.17		
604	5	MC	0.05	0.01	0.00		
604	6	MC	-0.28	0.30	-0.07		
604	7	MC	-0.05	0.02	-0.02		
604	8	MC	0.44	1.10	0.08		
605	1	MC	0.11	0.05	0.02		
605	2	MC	-0.08	0.03	-0.02		
605	3	MC	-0.74	1.51	-0.07		
605	4	MC	0.18	0.19	0.04		
605	5	MC	0.46	0.37	0.04		
605	6	MC	-0.51	1.17	-0.09		
605	7	MC	0.30	0.28	0.03		
605	8	MC	0.35	0.20	0.03		

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
606	11	MC	-0.02	0.00	0.01		
606	12	MC	1.02	5.35	0.18	B	F
606	13	MC	0.46	1.02	0.09		
606	14	MC	0.72	2.26	0.11		
606	15	MC	0.59	1.65	0.11		
606	16	MC	0.19	0.20	0.05		
606	17	MC	-0.26	0.31	-0.05		
606	18	MC	-0.16	0.14	-0.05		
606	19	MC	0.08	0.04	0.03		
606	20	MC	0.13	0.08	0.00		
606	21	MC	0.33	0.50	0.06		
606	22	MC	-0.60	1.75	-0.10		
607	11	MC	-0.17	0.10	-0.03		
607	12	MC	0.84	2.92	0.14		
607	13	MC	-0.15	0.10	-0.03		
607	14	MC	0.08	0.02	0.01		
607	15	MC	0.74	1.63	0.10		
607	16	MC	1.44	10.23	0.26	B	F
607	17	MC	-0.11	0.07	-0.03		
607	18	MC	1.19	5.90	0.18	B	F
607	19	MC	0.07	0.02	0.03		
607	20	MC	-0.42	0.87	-0.07		

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
607	21	MC	-0.16	0.12	-0.04		
607	22	MC	0.31	0.43	0.05		
608	11	MC	0.30	0.39	0.01		
608	12	MC	0.28	0.29	0.01		
608	13	MC	-0.90	4.58	-0.20		
608	14	MC	1.48	10.84	0.26	B	F
608	15	MC	0.16	0.13	0.02		
608	16	MC	0.00	0.00	-0.01		
608	17	MC	-0.14	0.08	-0.04		
608	18	MC	-0.45	0.81	-0.07		
608	19	MC	-0.51	1.23	-0.10		
608	20	MC	0.52	1.37	0.10		
608	21	MC	0.65	1.54	0.09		
608	22	MC	1.16	5.27	0.14	B	F
609	11	MC	-0.22	0.19	-0.03		
609	12	MC	0.48	1.15	0.10		
609	13	MC	0.54	1.11	0.07		
609	14	MC	-0.80	3.38	-0.17		
609	15	MC	0.35	0.56	0.06		
609	16	MC	0.28	0.39	0.05		
609	17	MC	-0.83	2.82	-0.12		
609	18	MC	-1.07	4.93	-0.19	B	M

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
609	19	MC	0.39	0.60	0.05		
609	20	MC	-0.63	1.86	-0.13		
609	21	MC	-0.06	0.02	0.02		
609	22	MC	0.11	0.07	0.04		
610	11	MC	0.90	3.33	0.14		
610	12	MC	0.04	0.01	0.02		
610	13	MC	1.15	6.27	0.18	B	F
610	14	MC	-0.01	0.00	0.00		
610	15	MC	1.09	6.17	0.21	B	F
610	16	MC	1.12	4.19	0.16	B	F
610	17	MC	-0.02	0.00	-0.01		
610	18	MC	-0.19	0.19	-0.04		
610	19	MC	0.67	2.28	0.11		
610	20	MC	0.39	0.67	0.06		
610	21	MC	0.16	0.14	0.02		
610	22	MC	0.15	0.08	0.01		
612	1	MC	0.22	0.22	0.04		
612	2	MC	-1.00	4.04	-0.13	B	M
612	3	MC	0.83	3.05	0.14		
612	4	MC	-1.18	2.13	-0.09		
612	5	MC	0.12	0.04	0.02		
612	6	OE		5.99	0.19		

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
612	7	OE		13.42	0.28	CC	F
613	Es	OE		7.30	0.20		
614	1	MC	0.82	0.87	0.05		
614	2	MC	0.01	0.00	0.00		
614	3	MC	0.40	0.52	0.05		
614	4	MC	-0.31	0.60	-0.07		
614	5	MC	1.79	13.16	0.30	C	F
614	6	OE		13.73	0.29	CC	F
614	7	OE		11.29	0.25	CC	F
615	Es	OE		24.14	0.39	CC	F
616	1	MC	1.31	3.59	0.14		
616	2	MC	0.68	0.82	0.07		
616	3	MC	0.87	1.67	0.10		
616	4	MC	0.82	2.33	0.12		
616	5	MC	0.36	0.78	0.07		
616	6	OE		24.39	0.39	CC	F
616	7	OE		3.89	0.16		
617	Es	OE		19.42	0.34	CC	F
618	1	MC	0.41	0.76	0.08		
618	2	MC	0.13	0.04	0.03		
618	3	MC	-1.06	5.92	-0.20	B	M
618	4	MC	0.31	0.49	0.06		

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
618	5	MC	0.25	0.29	0.03		
618	6	OE		7.65	0.22	BB	F
618	7	OE		3.61	0.14		
619	Es	OE		7.83	0.22		
620	1	MC	0.60	1.35	0.10		
620	2	MC	0.89	2.44	0.10		

*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
620	3	MC	0.98	1.66	0.09		
620	4	MC	0.09	0.02	0.02		
620	5	MC	1.48	4.54	0.17	B	F
620	6	OE		15.31	0.28	CC	F
620	7	OE		13.86	0.27	CC	F
621	Es	OE		10.82	0.24		

Appendix D: Operational Test Maps

Table 11. Operational Test Map for January 2010

Position	Item Type	Max Points	Weight	Strand	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
1	MC	1	1	1	0.88	0.39	-2.14						
2	MC	1	1	1	0.88	0.43	-2.16						
3	MC	1	1	1	0.84	0.50	-1.75						
4	MC	1	1	1	0.68	0.49	-0.57						
5	MC	1	1	1	0.91	0.44	-2.51						
6	MC	1	1	1	0.63	0.38	-0.27						
Es	Es	6	2	1	3.28	0.72	0.47	-2.33	-2.07	-1.74	-0.40	2.93	3.61
7	MC	1	1	1	0.88	0.45	-1.57						
8	MC	1	1	1	0.93	0.43	-2.32						
9	MC	1	1	1	0.86	0.42	-1.32						
10	MC	1	1	1	0.78	0.45	-0.66						
11	MC	1	1	1	0.69	0.51	-0.03						
12	MC	1	1	1	0.84	0.35	-1.15						
13	MC	1	1	1	0.90	0.44	-1.80						
14	MC	1	1	1	0.90	0.41	-1.80						
15	MC	1	1	1	0.82	0.41	-0.92						
16	MC	1	1	1	0.89	0.46	-1.69						
Es	Es	6	2	1	3.22	0.75	0.90	-2.29	-1.50	-1.62	0.58	1.59	3.24
1	MC	1	1	2	0.85	0.42	-1.41						
2	MC	1	1	2	0.93	0.42	-2.43						

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

Position	Item Type	Max Points	Weight	Strand	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
3	MC	1	1	2	0.80	0.47	-0.94						
4	MC	1	1	2	0.77	0.35	-0.71						
5	MC	1	1	2	0.86	0.34	-1.50						
6	MC	1	1	2	0.88	0.39	-1.73						
7	MC	1	1	2	0.93	0.44	-2.34						
8	MC	1	1	2	0.66	0.49	-0.05						
9	MC	1	1	2	0.77	0.25	-0.76						
10	MC	1	1	2	0.92	0.46	-2.17						
Es	Es	6	2	2	3.14	0.70	1.04	-2.89	-2.54	-2.15	0.56	2.58	4.45
Es	Es	6	2	3	2.99	0.67	2.01	-4.10	-2.94	-2.72	-0.10	3.53	6.33

Table 12. Operational Test Map for June 2010

Position	Item Type	Max Points	Weight	Strand	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
1	MC	1	1	1	0.90	0.31	-2.03						
2	MC	1	1	1	0.84	0.42	-1.36						
3	MC	1	1	1	0.84	0.36	-1.43						
4	MC	1	1	1	0.81	0.32	-1.14						
5	MC	1	1	1	0.90	0.34	-2.06						
6	MC	1	1	1	0.64	0.48	0.00						
Es	Es	6	2	1	3.21	0.70	1.05	-2.67	-2.38	-2.59	0.46	2.22	4.96
7	MC	1	1	1	0.92	0.41	-2.04						
8	MC	1	1	1	0.82	0.45	-0.97						
9	MC	1	1	1	0.70	0.53	-0.07						
10	MC	1	1	1	0.86	0.50	-1.32						
11	MC	1	1	1	0.68	0.47	0.04						
12	MC	1	1	1	0.80	0.53	-0.77						
13	MC	1	1	1	0.62	0.46	0.37						
14	MC	1	1	1	0.72	0.45	-0.20						
15	MC	1	1	1	0.69	0.50	-0.04						
16	MC	1	1	1	0.69	0.49	-0.03						
Es	Es	6	2	1	3.15	0.73	1.08	-2.36	-1.54	-1.44	0.44	1.43	3.49

Table 12. Operational Test Map for June 2010 *(continued)*

Position	Item Type	Max Points	Weight	Strand	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
1	MC	1	1	2	0.82	0.43	-1.30						
2	MC	1	1	2	0.80	0.40	-1.10						
3	MC	1	1	2	0.93	0.29	-2.57						
4	MC	1	1	2	0.74	0.31	-0.70						
5	MC	1	1	2	0.84	0.44	-1.49						
6	MC	1	1	2	0.84	0.46	-1.43						
7	MC	1	1	2	0.89	0.40	-2.05						
8	MC	1	1	2	0.90	0.41	-2.07						
9	MC	1	1	2	0.70	0.50	-0.44						
10	MC	1	1	2	0.86	0.46	-1.62						
Es	Es	6	2	2	3.09	0.59	0.21	-1.60	-2.41	-1.93	1.60	4.35	
Es	Es	6	2	3	2.97	0.64	0.87	-3.21	-1.19	-1.72	1.10	4.92	

Table 13. Operational Test Map for August 2010

Position	Item Type	Max Points	Weight	Strand	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
1	MC	1	1	1	0.81	0.42	-1.29						
2	MC	1	1	1	0.83	0.33	-1.46						
3	MC	1	1	1	0.81	0.39	-1.23						
4	MC	1	1	1	0.88	0.40	-1.87						
5	MC	1	1	1	0.89	0.39	-2.03						
6	MC	1	1	1	0.84	0.44	-1.52						
Es	Es	6	2	1	3.32	0.69	0.64	-3.18	-2.26	-1.68	-0.38	3.27	4.22
7	MC	1	1	1	0.82	0.49	-0.93						
8	MC	1	1	1	0.87	0.42	-1.42						
9	MC	1	1	1	0.64	0.43	0.26						
10	MC	1	1	1	0.81	0.38	-0.89						
11	MC	1	1	1	0.86	0.48	-1.30						
12	MC	1	1	1	0.87	0.42	-1.40						
13	MC	1	1	1	0.85	0.41	-1.22						
14	MC	1	1	1	0.83	0.46	-1.03						
15	MC	1	1	1	0.85	0.46	-1.22						
16	MC	1	1	1	0.81	0.49	-0.85						
Es	Es	6	2	1	3.33	0.74	0.72	-2.90	-1.84	-1.18	0.64	1.68	3.60
1	MC	1	1	2	0.70	0.35	-0.38						
2	MC	1	1	2	0.82	0.43	-1.20						

Table 13. Operational Test Map for August 2010 *(continued)*

Position	Item Type	Max Points	Weight	Strand	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
3	MC	1	1	2	0.73	0.49	-0.58						
4	MC	1	1	2	0.61	0.37	0.12						
5	MC	1	1	2	0.84	0.49	-1.45						
6	MC	1	1	2	0.81	0.53	-1.16						
7	MC	1	1	2	0.84	0.36	-1.39						
8	MC	1	1	2	0.83	0.39	-1.37						
9	MC	1	1	2	0.93	0.45	-2.48						
10	MC	1	1	2	0.63	0.30	0.03						
Es	Es	6	2	2	3.13	0.67	0.93	-2.68	-2.52	-2.20	0.70	2.13	4.58
Es	Es	6	2	3	3.04	0.71	1.63	-3.49	-2.40	-2.12	0.24	2.37	5.40

Appendix E: Scoring Tables

Table 14. Scoring Table for January 2010

Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score
0	-5.624	0.000		19	-0.649	10.056		38	0.783	43.471		57	1.377	79.486
1	-4.877	0.123		20	-0.514	11.303		39	0.817	45.563		58	1.414	81.027
2	-4.130	1.000		21	-0.384	13.259		40	0.850	47.625		59	1.453	83.105
3	-3.668	1.000		22	-0.260	14.387		41	0.882	49.733		60	1.495	84.610
4	-3.323	1.000		23	-0.144	16.030		42	0.914	51.867		61	1.539	85.636
5	-3.040	1.000		24	-0.037	17.641		43	0.944	53.867		62	1.587	87.362
6	-2.796	1.607		25	0.061	19.529		44	0.975	56.000		63	1.640	88.745
7	-2.578	2.000		26	0.150	20.873		45	1.004	58.000		64	1.698	89.804
8	-2.379	2.000		27	0.230	22.944		46	1.034	59.143		65	1.764	91.746
9	-2.193	2.000		28	0.304	24.164		47	1.063	61.207		66	1.838	92.930
10	-2.018	2.544		29	0.370	26.351		48	1.093	63.286		67	1.926	94.011
11	-1.850	3.000		30	0.430	28.491		49	1.122	65.357		68	2.030	95.087
12	-1.688	3.200		31	0.486	30.333		50	1.152	67.483		69	2.159	96.170
13	-1.531	4.077		32	0.537	31.826		51	1.182	69.276		70	2.326	97.000
14	-1.377	4.994		33	0.584	33.930		52	1.212	70.600		71	2.554	97.389
15	-1.226	5.949		34	0.629	36.073		53	1.243	72.645		72	2.897	98.526
16	-1.078	6.927		35	0.670	38.158		54	1.275	74.710		73	3.527	99.000
17	-0.932	7.944		36	0.710	39.649		55	1.308	76.406		74	4.157	99.657
18	-0.789	8.985		37	0.747	41.371		56	1.342	77.909				

Table 15. Scoring Table for June 2010

Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score
0	-5.3520	0.000		19	-0.4060	13.063		38	0.8200	45.750		57	1.4040	80.486
1	-4.5990	0.493		20	-0.2880	14.123		39	0.8520	47.750		58	1.4410	82.474
2	-3.8460	1.000		21	-0.1760	15.384		40	0.8830	49.800		59	1.4810	84.268
3	-3.3800	1.000		22	-0.0730	17.250		41	0.9140	51.867		60	1.5240	85.295
4	-3.0310	1.000		23	0.0230	18.635		42	0.9440	53.867		61	1.5690	86.596
5	-2.7450	1.754		24	0.1110	20.380		43	0.9740	55.931		62	1.6190	88.333
6	-2.4980	2.000		25	0.1910	21.861		44	1.0030	57.931		63	1.6730	89.357
7	-2.2780	2.000		26	0.2640	23.478		45	1.0320	59.000		64	1.7340	90.794
8	-2.0770	2.270		27	0.3310	25.049		46	1.0610	61.069		65	1.8020	92.423
9	-1.8900	3.000		28	0.3930	27.158		47	1.0900	63.071		66	1.8800	93.451
10	-1.7140	3.056		29	0.4490	29.208		48	1.1190	65.143		67	1.9710	94.484
11	-1.5460	3.989		30	0.5010	30.646		49	1.1480	67.207		68	2.0800	95.522
12	-1.3860	4.940		31	0.5490	32.348		50	1.1770	69.103		69	2.2160	96.574
13	-1.2310	5.918		32	0.5950	34.442		51	1.2070	70.267		70	2.3900	97.000
14	-1.0820	6.901		33	0.6370	36.463		52	1.2370	72.258		71	2.6270	97.684
15	-0.9370	7.909		34	0.6770	38.526		53	1.2680	74.258		72	2.9820	98.759
16	-0.7970	8.927		35	0.7150	39.784		54	1.3000	76.156		73	3.6260	99.000
17	-0.6610	9.962		36	0.7510	41.600		55	1.3330	77.364		74	4.2700	99.829
18	-0.5310	11.017		37	0.7860	43.647		56	1.3680	79.229				

Table 16. Scoring Table for August 2010

Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score
0	-5.2790	0.000		19	-0.4710	12.025		38	0.8050	44.813		57	1.3920	79.914
1	-4.5350	0.579		20	-0.3510	13.554		39	0.8380	46.875		58	1.4290	81.838
2	-3.7910	1.000		21	-0.2380	14.594		40	0.8700	48.933		59	1.4690	83.947
3	-3.3340	1.000		22	-0.1300	16.313		41	0.9010	51.000		60	1.5110	85.000
4	-2.9930	1.038		23	-0.0300	17.717		42	0.9320	53.067		61	1.5560	86.043
5	-2.7160	1.838		24	0.0620	19.553		43	0.9610	55.034		62	1.6050	88.059
6	-2.4790	2.000		25	0.1470	20.835		44	0.9910	57.103		63	1.6580	89.089
7	-2.2680	2.000		26	0.2250	22.806		45	1.0200	58.571		64	1.7180	90.286
8	-2.0760	2.274		27	0.2960	23.955		46	1.0490	60.214		65	1.7840	92.169
9	-1.8990	3.000		28	0.3600	26.000		47	1.0780	62.241		66	1.8610	93.220
10	-1.7320	3.000		29	0.4200	28.113		48	1.1070	64.286		67	1.9500	94.263
11	-1.5730	3.839		30	0.4750	30.104		49	1.1370	66.429		68	2.0570	95.322
12	-1.4210	4.732		31	0.5250	31.304		50	1.1660	68.448		69	2.1890	96.383
13	-1.2740	5.646		32	0.5720	33.372		51	1.1960	69.759		70	2.3600	97.000
14	-1.1310	6.576		33	0.6170	35.488		52	1.2260	71.533		71	2.5930	97.547
15	-0.9920	7.524		34	0.6580	37.526		53	1.2570	73.548		72	2.9430	98.652
16	-0.8570	8.489		35	0.6970	39.297		54	1.2890	75.613		73	3.5800	99.000
17	-0.7240	9.481		36	0.7350	40.686		55	1.3220	76.844		74	4.2170	99.748
18	-0.5960	10.484		37	0.7700	42.706		56	1.3560	78.758				