

New York State Regents Examination in English Language Arts (Common Core)

2013 Field Test Analysis, Equating Procedure, and Scaling of Operational Test Forms

Technical Report



Prepared for the New York State Education Department
by Pearson

December 2013

Copyright

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

Copyright © 2013 by the New York State Education Department.

Table of Contents

Table of Contents	i
List of Tables	iii
List of Figures.....	iii
Section I: Introduction.....	1
PURPOSE	1
Section II: Field Test Analysis	1
FILE PROCESSING AND DATA CLEANUP	2
CLASSICAL ANALYSIS	3
<i>Item Difficulty</i>	4
<i>Item Discrimination</i>	4
<i>Test Reliability</i>	5
<i>Scoring Reliability</i>	5
<i>Inter-Rater Agreement</i>	6
<i>Constructed-Response Item Means and Standard Deviations</i>	7
<i>Intraclass Correlation</i>	7
<i>Weighted Kappa</i>	8
ITEM RESPONSE THEORY (IRT) AND THE CALIBRATION AND EQUATING OF THE FIELD TEST ITEMS	8
<i>Item Calibration</i>	10
<i>Item Fit Evaluation</i>	11
DIFFERENTIAL ITEM FUNCTIONING.....	12
<i>The Mantel Chi-Square and Standardized Mean Difference</i>	13
<i>Multiple-Choice Items</i>	14
<i>The Odds Ratio</i>	14
<i>The Delta Scale</i>	15
<i>DIF Classification for MC Items</i>	15
<i>DIF Classification for CR Items</i>	15
Section III: Equating Procedure.....	16
COMMON ITEM EQUATING DESIGN.....	16
Section IV: Scaling of Operational Test Forms.....	17
References.....	21
Appendix A: Classical Item Analysis	23
Appendix B: Inter-Rater Consistency – Point Differences Between First and Second Reads.....	31
Appendix C: Additional Measures of Inter-Rater Reliability and Agreement.....	32

Appendix D: Partial-Credit Model Item Analysis	33
Appendix E: DIF Statistics	41
Appendix F: Operational Test Maps	47
Appendix G: Scoring Tables	50

List of Tables

Table 1. Need/Resource Capacity Category Definitions	2
Table 2. Classical Item Analysis Summary	5
Table 3. Test and Scoring Reliability	6
Table 4. Criteria to Evaluate Mean-Square Fit Statistics	11
Table 5. Partial-Credit Model Item Analysis Summary	12
Table 6. DIF Classification for MC Items	15
Table 7. DIF Classification for CR Items	15

List of Figures

Figure 1. 2 × t Contingency Table at the k th of K Levels.....	13
---	----

Section I: Introduction

PURPOSE

The purpose of this report is to document the psychometric properties of the New York State Regents Examination in English Language Arts. In addition, this report documents the procedures used to analyze the results of the field test and to equate and scale the operational test forms.

This test is being transitioned to the Common Core-based English Language Arts curricula. Therefore, the items that were field tested were all Common Core-based items while the operational tests all addressed the old New York Regents Comprehensive English framework. For the sake of simplicity, both the Common Core-based and Non-Common Core-based assessments will be referred to as the “New York State Regents Examination in English Language Arts.”

Section II: Field Test Analysis

In May 2013, prospective items for the New York State Regents Examination in English Language Arts were field tested. The results of this testing were used to evaluate item quality. Only items with acceptable statistical characteristics can be selected for use on operational tests.

Representative student samples for participation in this testing were selected to mirror the demographics of the student population that is expected to take the operational test. The Need/Resource Capacity Categories in Table 1 were used as variables in the sampling plan.

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	All districts at or above the 70 th percentile on the index with at least 100 students per square mile or enrollment greater than 2500
Rural	All districts at or above the 70 th percentile on the index with fewer than 50 students per square mile or enrollment of fewer than 2500
Average N/RC Districts	All districts between the 20 th and 70 th percentiles on the index
Low N/RC Districts	All districts below the 20 th percentile on the index
Charter Schools	Each charter school is a district

FILE PROCESSING AND DATA CLEANUP

The Regents examinations utilize both multiple-choice (MC) and constructed-response (CR) item types in order to more fully assess student ability. Multiple field test (FT) forms were given during this administration to allow for a large number of items to be field tested without placing an undue burden on the students participating in the field test; each student took only a small subset of the items being field tested. The NYSED handled all scanning of the MC responses. Scoring of the CR responses was performed by Measurement Incorporated (MI) under contract with the NYSED. The NYSED and MI produced separate data files, which were both provided to Pearson. A test map file that documented the items on each of the FT forms was also provided to Pearson by the NYSED. Finally, student data file layouts containing the position of every field within the student data files from both the NYSED and MI were also provided to Pearson by the NYSED. Upon receipt of these files, Pearson staff checked the data, test map, and layouts for consistency. Any anomalies were referred back to the NYSED for resolution. After these had been resolved and corrected as necessary, final processing of the data file took place. Merging of the NYSED and MI provided data was accomplished through uniquely assigned booklet numbers. This processing included the identification and deletion of invalid student test records through the application of a set of predefined

exclusion rules¹. The original student data file received from the NYSED contained 14,691 records; the final field test data file contained 10,321 records².

Within the final data file used in the field test analyses, MC responses were scored according to the item keys contained in the test map; correct responses received a score of 1 while incorrect responses received a score of 0. CR item scores were taken directly from the student data file, with the exception that out-of-range scores were assigned scores of 0. For Item Response Theory (IRT) calibrations, blanks (i.e., missing data; not omits) were also scored as 0.

In addition to the scored data, the final data file also contained the unscored student responses and scores. Unscored data was used to calculate the percentage of students who selected the various answer choices for the MC items or the percentage of students who received each achievable score point for the CR items. The frequency of students leaving items blank was also calculated. The scored data were used for all other analyses.

CLASSICAL ANALYSIS

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

$$x = t + e$$

All test scores are composed of both a true and an error component. 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 ability would warrant. This error component is random and uncorrelated with (i.e., unrelated to) the student's true score. Across an infinitely large number of administrations, the mean of the error scores would be zero. Thus, the best estimate of a student's true score for any test administration (or their expected score given their [unobservable] true level of ability or true score) is that student's observed score. This expectation is expressed as follows:

$$E(x) = t$$

Item difficulties, point-biserial correlations, reliability estimates, and various statistics related to rater agreement have been calculated and are summarized in the following section.

¹ These exclusion rules flagged records 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 prior to analysis.

² Forms 613 and 626 each had one field test item (a six-point and a four-point CR item, respectively) that was invalidated (Do Not Score [DNS] status). Since these forms no longer had field test items, they (and their associated records) were dropped from the analyses. This accounts for the larger than usual difference between the number of records in the data file received from the NYSED and the final data file.

Item Difficulty

Item difficulty is typically defined as the average of scores for a given item. For MC items, this value (commonly referred to as a p-value) ranges from 0 to 1. For CR items, this value ranges from 0 to the maximum possible score. In order to place all item means on a common metric (ranging from 0 to 1), CR item means were divided by the maximum points possible for the item.

Item Discrimination

Item discrimination is defined as the correlation between a score on a given test question and the overall raw test score. These correlations are Pearson correlation coefficients. For MC items, it is also known as the point-biserial correlation.

Table 2 presents a summary of the classical item analysis for each of the field test forms. The first three columns from the left 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 discrimination). Recall that for CR items, item means were divided by the maximum number of points possible in order to place them in the same metric as the MC items. There were no items with difficulties that were greater than 0.90; 30 items had correlations that were less than 0.25. In addition to the summary information provided in Table 2, further classical item statistics are provided in Appendix A.

Table 2. Classical Item Analysis Summary

Form	N-Count	No. of Items	Item Difficulty			Item Discrimination		
			<0.50	0.50 to 0.90	>0.90	<0.25	0.25 to 0.50	>0.50
N3	10321	10	3	7	0	0	2	8
601	747	30	17	13	0	2	26	2
602	752	30	16	14	0	5	20	5
603	750	30	13	17	0	3	21	6
604	752	30	14	16	0	8	19	3
605	754	30	16	14	0	7	16	7
606	759	30	21	9	0	5	19	6
611	641	1	1	0	0	0	0	1
612	664	1	1	0	0	0	0	1
614	668	1	0	1	0	0	0	1
615	653	1	0	1	0	0	0	1
621	436	1	1	0	0	0	0	1
622	587	1	0	1	0	0	0	1
623	414	1	1	0	0	0	0	1
624	600	1	1	0	0	0	0	1
625	611	1	1	0	0	0	0	1
627	533	1	0	1	0	0	0	1

For some forms, the item counts in the “Item Difficulty” and “Item Discrimination” columns may not sum to the value in the “No. of Items” column due to DNS (Do Not Score) items.

Form N3 was the common anchor form, and so its statistics were aggregated across all forms.

The items field tested with forms 613 and 626 were classed as DNS (Do Not Score). Since these forms no longer had field test items, they were dropped from the analyses.

Test Reliability

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, the higher the reliability.

Table 3 contains the internal consistency statistics for each of the field test forms under the heading “Test Reliability.” These statistics ranged from 0.643 to 0.891. It should be noted that operational tests generally are composed of more items and would be expected to have higher reliabilities than do these field test forms.

Scoring Reliability

One concern with CR items is the reliability of the scoring process (i.e., consistency of the score assignment). CR items must be read by scorers who assign scores based

on a comparison between the rubric and student responses. Consistency between scorers is a critical part of the reliability of the assessment. To track scorer consistency, approximately 10% of the test booklets are scored a second time (these are termed “second read scores”) and compared to the original set of scores (also known as “first read scores”).

As an overall measure of scoring reliability, the Pearson correlation coefficient between the first and second scores for all CR items with second read scores was computed for each form. This statistic is often used as an overall indicator of scoring reliability, and it generally ranges from 0 to 1. Table 3 contains these values in the column headed “Scoring Reliability.” They ranged from 0.657 to 0.891, indicating a fair to high degree of reliability, especially considering that the CR items were all four-point or six-point items. Note that forms 601–606 did not have any CR items, and so this statistic does not apply to those forms.

Table 3. Test and Scoring Reliability

Form Number	Test Reliability	Scoring Reliability
601	0.884	N/A
602	0.873	N/A
603	0.891	N/A
604	0.842	N/A
605	0.877	N/A
606	0.873	N/A
611	0.681	0.884
612	0.656	0.816
614	0.680	0.853
615	0.682	0.891
621	0.684	0.732
622	0.657	0.801
623	0.669	0.660
624	0.669	0.828
625	0.643	0.657
627	0.643	0.737

Inter-Rater Agreement

For each CR item, the difference between the first and second reads was tracked, and the number of times each possible difference between the scores occurred was tabulated. These values were then used to calculate the percentage of times each possible difference occurred. 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. Blank cells in the table indicate out-of-range differences (e.g., it is impossible for two raters to differ by more than one point in their scores on an item

with a maximum possible score of one; cells in the table other than -1, 0, and 1 would therefore be blanked out).

Appendix B contains the proportion of occurrence of these differences for each CR item. All items had a maximum point value of either four or six. Only one each of the four-point and six-point items had ratings that differed by more than one point, and this only occurred for between 1–2% of the sample that received dual reads. Appendix C contains additional summary information regarding the first and second reads, including the percentage of first and second scores that were exact or adjacent matches. The percentage of exact matches ranged from 61.2 to 75.2%, and the percentage of exact plus adjacent matches ranged from 98.1 to 100%. In general, the four-point items had higher rates of exact matches compared to the rates for the six-point items.

Constructed-Response Item Means and Standard Deviations

Appendix C also contains the mean and standard deviation of the first and second scores for each CR item. While there were minimal differences between the standard deviation statistics, the largest difference between the item means for the first and second read scores was 0.1.

Intraclass Correlation

In addition, Appendix C contains the intraclass correlations for the items. These correlations are calculated using a formulation given by Shrout and Fleiss (1979). Specifically, they described six different models based on various configurations of judges and targets (in this case, papers that are being scored). For this assessment, the purpose of the statistic is to describe the reliability of single ratings, and each paper is scored by two judges, who are randomly assigned from the larger pool of judges, and who score multiple papers. This description fits their “Case 1.” Further, they distinguish between situations where the score assigned to the paper is that of a single rater versus that when the score is the mean of k raters. Since the students’ operational scores are those from single (i.e., the first) raters, the proper intraclass correlation in this instance is termed by Shrout and Fleiss as “ICC(1,1).” It will be referred to herein simply as the “intraclass correlation” (ICC).

While the ICC is a bona fide correlation coefficient, it differs from a regular correlation coefficient in that its value remains the same regardless of how the raters are ordered. A regular Pearson correlation coefficient would change values if, for example, half of the second raters were switched to the first position, while the ICC would maintain a consistent value. Because the papers were randomly assigned to the judges, ordering was arbitrary, and thus the ICC is a more appropriate measure of reliability than the Pearson correlation coefficient in this situation. The ICC ranges from zero (the scores given by the two judges are unrelated) to one (the scores from the two judges match perfectly); negative values are possible, but rare, and have essentially the same meaning as values of zero. It should also be noted that the ICC can be affected by low degrees of variance in the scores being related, which is similar to the way that regular Pearson correlation coefficients are affected. ICCs for items where almost every examinee achieved the same score point (e.g., an extremely easy dichotomous item

where almost every examinee was able to answer it correctly) may have a low or negative ICC even though almost all ratings by the judges matched exactly.

McGraw and Wong (1996, Table 4, p. 35) state that the ICC can be interpreted as “the degree of absolute agreement among measurements made on randomly selected objects. It estimates the correlation of any two measurements.” Since it is a correlation coefficient, its square indicates the percent of variance in the scores that is accounted for by the relationship between the two sets of scores (i.e., the two measurements). In this case, these scores are those of the pair of judges. ICC values greater than 0.60 indicate that at least 36% (0.60^2) of the variation in the scores given by the raters is accounted for by variations in the responses to the items that are being scored (e.g., variations in the ability being measured) rather than by variations caused by a combination of differences in the severity of the judges, interactions between judge severity and the items, and random error (e.g., variations exterior to the ability being measured). It is generally preferred that items have ICCs at this level or higher. There were no items with ICCs below 0.60. Consistent with other information provided in the table, these values indicate a high to very high level of scoring reliability for almost all of the items in the field test.

Weighted Kappa

Weighted Kappa (Cohen, 1968) was also calculated for each item, based on the first and second reads and is included in Appendix C as well. This statistic is an estimate of the agreement of the score classifications over and above that which would be expected to occur by chance. Similar to the ICC, its value can range between zero (the scores given by the judges agree as often as would be expected by chance) and one (scores given by the judges agree perfectly). In addition, negative values are possible, but rare, and have the same interpretation as zero values. One set of guidelines for the evaluation of this statistic is (Fleiss, 1981):

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

The results show excellent reproducibility between the first and second reads for one of the 10 items field tested, and good reproducibility for the remaining nine. With the lowest kappa being equal to 0.55, there were no items displaying marginal reproducibility. The scoring reliability analyses offer strong evidence that the scoring of the CR items was performed in a highly reliable manner.

ITEM RESPONSE THEORY (IRT) AND THE CALIBRATION AND EQUATING OF THE FIELD TEST ITEMS

While classical test theory-based statistical measures are useful for assessing the suitability of items for operational use (i.e., use as part of an assessment used to measure student ability and thus having real-world consequences for students, teachers, schools, and administrators), their values are dependent on both the

psychometric properties of the items and the ability distributions of the samples upon which they are based. In other words, classical test theory-based statistics are *sample-dependent statistics*.

In contrast, Item Response Theory (IRT) based statistics are not dependent on the sample over which they are estimated—they are invariant across different samples (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). This invariance allows student ability to be estimated on a common metric even if different sets of items are used (as with different test forms over different test administrations).

The process of estimating IRT-based item parameters is referred to as “item calibration,” and the placing of these parameters on a common metric or scale is termed “equating.” While one reason for the field testing of items is to allow their suitability for use in the operational measurement of student ability to be assessed, the data resulting from field testing is also used to place items on the scale of the operational test (i.e., they are equated to the operational metric). Once items are on this common metric, any form composed of items from this pool can be scaled (the process through which scale score equivalents for each achievable raw score are derived) and the resulting scale scores will be directly comparable to those from other administrations, even though the underlying test forms are composed of different sets of items.

There are several variations of IRT that differ mainly in the way item behavior is modeled. The New York State Regents Examinations use the Rasch family of IRT statistics to calibrate, scale, and equate all subjects (Rasch, 1980; Masters, 1982).

The most basic expression of the Rasch model is in the item characteristic curve. It conceptualizes the probability of a correct response to an item as a function of the ability level and the item’s difficulty. The probability of a correct response is bounded by “1” (certainty of a correct response) and “0” (certainty of an incorrect response). The ability scale is theoretically unbounded. In practice, the ability scale ranges from approximately -4 to $+4$ logits. The relationship between examinee ability θ , item difficulty D_i , and probability of answering the item correctly P_i is shown in the equation below:

$$P_i(\theta) = \frac{\exp(\theta - D_i)}{1 + \exp(\theta - D_i)}$$

Examinee ability (θ) and item difficulty (D_i) are on the same scale. This is useful for certain purposes. An examinee with an ability level equal to the item difficulty will have a 50% chance of answering the item correctly; if his or her ability level is higher than the item difficulty, then the probability of answering the item correctly is commensurately higher, and the converse is also true.

The Rasch Partial Credit Model (PCM) (Masters, 1982) is a direct extension of the dichotomous one-parameter IRT model above. For an item involving m score

categories, the general expression for the probability of achieving a score of x on the item is given by

$$P_x(\theta) = \frac{\exp[\sum_{k=0}^x (\theta - D_k)]}{\sum_{h=0}^m \exp[\sum_{k=0}^h (\theta - D_k)]}$$

where

$$D_0 \equiv 0.0$$

In the above equation, P_x is the probability of achieving a score of x given an ability of θ ; m is the number of achievable score points minus one (note that the subscript k runs from 0 to m); and D_k is the step parameter for step k . The steps are numbered from 0 to the number of achievable score points minus one, and step 0 (D_0) is defined as being equal to zero. Note that a four-point item, for example, usually has five achievable score points (0, 1, 2, 3, and 4), thus the step numbers usually mirror the achievable point values.

According to this model, the probability of an examinee scoring in a particular category (step) is the sum of the logit (log-odds) differences between θ and D_k of all the completed steps, divided by the sum of the differences of all the steps of an item. Thissen and Steinberg (1986) refer to this model as a divide-by-total model. The parameters estimated by this model are $m_i - 1$ threshold (difficulty) estimates and represent the points on the ability continuum where the probability of the examinee achieving score m_i exceeds that of m_{i-1} . The mean of these threshold estimates is used as an overall summary of the polytomous item's difficulty.

If the number of achievable score points is one (i.e., the item is dichotomous), then the PCM reduces to the basic Rasch IRT model for dichotomous items. This means that dichotomous and polytomous items are being scaled using a common model and therefore can be calibrated, equated, and scaled together. It should be noted that the Rasch model assumes that all items have equal levels of discrimination and that there is no guessing on MC items. However, it is robust to violations of these assumptions, and items that violate these assumptions to a large degree are usually flagged for item-model misfit.

Item Calibration

When interpreting IRT item parameters, it is important to remember that they do not have an absolute scale—rather, their scale (in terms of mean and standard deviation) is purely arbitrary. It is conventional to set the mean of the item difficulties to zero when an assessment is scaled for the first time. Rasch IRT scales the theta measures in terms of *logits*, or “log-odds units.” The length of a logit varies from test to test, but generally the standard deviation of the item difficulties of a test scaled for the first time will be somewhere in the area of 0.6–0.8. While the item difficulties are invariant with respect to one another, the absolute level of difficulty represented by their mean is dependent on the overall difficulty of the group of items with which it was tested. In addition, there is

no basis for assuming that the difficulty values are normally distributed around their mean—their distribution again depends solely upon the intrinsic difficulties of the items themselves. Thus, if a particularly difficult set of items (relative to the set of items originally calibrated) was field tested, their overall mean would most probably be greater than zero, and their standard deviation would be considerably less than one. In addition, they would most probably not be normally distributed.

Rasch item difficulties generally range from -3.0 to 3.0 , although very easy or difficult items can fall outside of this range. Items should not be discounted solely on the basis of their difficulty. A particular topic may require either a difficult or an easy item. Items are usually most useful if their difficulty is close to a cut score, as items provide the highest level of information at the ability level equal to their difficulty. Items with difficulties farther away from the cuts provide less information about students with abilities close to the cut scores (and hence more susceptible to misclassification), but they are still useful. In general, items should be selected for use based on their content, with their Rasch difficulty being only a secondary consideration.

Item Fit Evaluation

The INFIT statistic is used to assess how well items fit the Rasch model. Rasch theory models the probability of a student being able to answer an item correctly as a function of the student's level of ability and the item's difficulty, as stated previously. The Rasch model also assumes that items' discriminations do not differ, and that the items are not susceptible to guessing. If these assumptions do not hold (if, for example, an item has an extremely high or low level of discrimination), then the item's behavior will not be well modeled by Rasch IRT. Guidelines for interpretation of the INFIT statistic are taken from Linacre (2005) and can be found in Table 4 below.

Table 4. Criteria to Evaluate Mean-Square Fit Statistics

INFIT	Interpretation
>2.0	Distorts or degrades the measurement system
1.5–2.0	Unproductive for construction of measurement, but not degrading
0.5–1.5	Productive for measurement
<0.5	Unproductive for measurement, but not degrading. May produce misleadingly good reliabilities and separations

INFIT is an information-weighted fit statistic, which is more sensitive to unexpected behavior affecting responses to items near the person's measure (or ability) level. In general, values near 1.0 indicate little distortion of the measurement system, while values less than 1.0 indicate observations are too predictable (redundancy, model overfit). Values greater than 1.0 indicate unpredictability (unmodeled noise, model underfit).

Table 5 contains a summary of the analysis for each of the field test forms. The first column from the left 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 following columns show the frequency of items at three levels of difficulty (easier

items with a Rasch difficulty <-2.0 , moderate items with a Rasch difficulty between -2.0 and 2.0 , and more difficult items with a Rasch difficulty >2.0), and frequencies of item misfits as classified in the preceding table. Nearly all of the items fell within the moderate -2.0 to $+2.0$ difficulty range, and there were no items with an INFIT statistic outside the range most productive for measurement. Item level results of the analysis can be found in Appendix D.

Table 5. Partial-Credit Model Item Analysis Summary

Form	N-Count	No. of Items	Rasch			INFIT			
			<-2.0	-2.0 to 2.0	>2.0	<0.5	0.5 to 1.5	1.5 to 2.0	>2.0
N3	746	10	0	10	0	0	10	0	0
601	752	30	0	30	0	0	30	0	0
602	750	30	0	29	1	0	30	0	0
603	750	30	0	30	0	0	30	0	0
604	753	30	0	30	0	0	30	0	0
605	759	30	0	29	1	0	30	0	0
606	636	30	0	30	0	0	30	0	0
611	664	1	0	1	0	0	1	0	0
612	665	1	0	1	0	0	1	0	0
614	648	1	0	1	0	0	1	0	0
615	432	1	0	1	0	0	1	0	0
621	578	1	0	1	0	0	1	0	0
622	410	1	0	1	0	0	1	0	0
623	596	1	0	1	0	0	1	0	0
624	608	1	0	1	0	0	1	0	0
625	522	1	0	1	0	0	1	0	0
627	522	1	0	1	0	0	1	0	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. Also, “N-Count” does not include students with zero or perfect scores.

DIFFERENTIAL ITEM FUNCTIONING

Differential Item Functioning (DIF) occurs when members of a particular group have a different probability of success than members of another group who have the same level of ability for reasons unrelated to the academic skill or construct being measured. For example, items testing English Language Arts grammar skills may be more difficult for LEP students as opposed to non-LEP students, but such differences are likely due to the fact that the item measures an academic skill related to English Language Arts language proficiency. Such items would not be considered to be functioning differentially.

The Mantel Chi-Square and Standardized Mean Difference

The Mantel χ^2 is a conditional mean comparison of the ordered-response categories for reference and focal groups combined over values of the matching variable score. “Ordered” means that a response earning a score of “1” on an item is better than a response earning a score of “0” or “2” is better than “1,” and so on. “Conditional,” on the other hand, refers to the comparison of members from the two groups who received the same score on the matching variable, that is, the total test score in our analysis.

Group	Item Score				Total
	y_1	y_2	...	y_T	
Reference	n_{R1k}	n_{R2k}	...	n_{Rtk}	n_{R+k}
Focal	n_{F1k}	n_{F2k}	...	n_{Ftk}	n_{F+k}
Total	n_{+1k}	n_{+2k}	...	n_{+tk}	n_{++k}

Figure 1. $2 \times t$ Contingency Table at the k^{th} of K Levels.

Figure 1 (from Zwick, Donoghue, & Grima, 1993) shows a $2 \times t$ contingency table at the k^{th} of K levels, where t represents the number of response categories and k represents the number of levels of the matching variable. The values y_1, y_2, \dots, y_T represent the t scores that can be gained on the item. The values n_{Ftk} and n_{Rtk} represent the numbers of focal and reference groups who are at the k^{th} level of the matching variable and gain an item score of y_t . The “+” indicates the total number over a particular index (Zwick et al., 1993). The Mantel statistic is defined as the following formula:

$$\text{Mantel}\chi^2 = \frac{\left(\sum_k F_k - \sum_k E(F_k) \right)^2}{\sum_k \text{Var}(F_k)}$$

in which F_k represents the sum of scores for the focal group at the k^{th} level of the matching variable and is defined as follows:

$$F_k = \sum_t y_t n_{Ftk}$$

The expectation of F_k under the null hypothesis is

$$E(F_k) = \frac{n_{F+k}}{n_{++k}} \sum_t y_t n_{Ftk}$$

The variance of F_k under the null hypothesis is as follows:

$$\text{Var}(F_k) = \frac{n_{R+k}n_{F+k}}{n_{++k}^2(n_{++k}-1)} \left[(n_{++k} \sum_t y_t^2 n_{+tk}) - \left(\sum_t y_t n_{+tk} \right)^2 \right]$$

Under H_0 , the Mantel statistic has a chi-square distribution with one degree of freedom. In DIF applications, rejecting H_0 suggests that the students of the reference and focal groups who are similar in overall test performance tend to differ in their mean performance on the item. For dichotomous items, the statistic is identical to the Mantel-Haenszel (MH) (1959) statistic without the continuity correction (Zwick et al., 1993).

A summary statistic to accompany the Mantel approach is the standardized mean difference (SMD) between the reference and focal groups proposed by Dorans and Schmitt (1991). This statistic compares the means of the reference and focal groups, adjusting for differences in the distribution of the reference and focal group members across the values of the matching variable. The *SMD* has the following form:

$$SMD = \sum_k p_{Fk} m_{Fk} - \sum_k p_{Rk} m_{Rk}$$

in which

$$p_{Fk} = \frac{n_{F+k}}{n_{F++}}$$

is the proportion of the focal group members who are at the k^{th} level of the matching variable and

$$m_{Fk} = \frac{1}{n_{F+k} \sum_t y_t n_{Ftk}}$$

is the mean item score of the focal group members at the k^{th} level; and m_{Rk} is the analogous value for the reference group. As can be seen from the equation above, the SMD is the difference between the unweighted item mean of the focal group and the weighted item mean of the reference group. The weights for the reference group are applied to make the weighted number of the reference-group students the same as in the focal group within the same level of ability. A negative SMD value implies that the focal group has a lower mean item score than the reference group, conditional on the matching variable.

Multiple-Choice Items

For the MC items, the MH odds ratio (converted to the ETS delta scale [D]) is used to classify items into one of three categories of DIF.

The Odds Ratio

The odds of a correct response (proportion passing divided by proportion failing) are P/Q or $P/(1-P)$. The *odds ratio* is the odds of a correct response of the reference group divided by the odds of a correct response of the focal group. For a given item, the odds ratio is defined as follows:

$$\alpha_{MH} = \frac{P_r/Q_r}{P_f/Q_f}$$

and the corresponding null hypothesis is that the odds of getting the item correct are equal for the two groups. Thus, the odds ratio is equal to 1:

$$\alpha_{MH} = \frac{P_r/Q_r}{P_f/Q_f} = 1$$

The Delta Scale

To make the odds ratio symmetrical around zero with its range being in the interval $-\infty$ to $+\infty$, the odds ratio is transformed into a log odds ratio according to this equation:

$$\beta_{MH} = \ln(\alpha_{MH})$$

This simple natural logarithm transformation of the odds ratio is symmetrical around zero. This DIF measure is a signed index. A positive value signifies DIF in favor of the reference group, a negative value indicates DIF in favor of the focal group, and zero has the interpretation of equal odds of success on the item. β_{MH} also has the advantage of a linear relationship to other interval scale metrics (Camilli & Shepard, 1994). β_{MH} is placed on the ETS delta scale (D) using the following equation:

$$D = -2.35\beta_{MH}.$$

DIF Classification for MC Items

Table 6 depicts DIF classifications for MC items. Classification depends on the delta (D) value and the significance of its difference from zero ($p < 0.05$). The criteria are derived from those used by the National Assessment of Educational Progress (Allen, Carlson, & Zalanak, 1999) in the development of their assessments.

Table 6. DIF Classification for MC Items

Category	Description	Criterion
A	No DIF	D not significantly different from zero or $ D < 1.0$
B	Moderate DIF	$1.0 \leq D < 1.5$ or not otherwise A or C
C	High DIF	D is significantly different from zero and $ D \geq 1.5$

DIF Classification for CR Items

The SMD is divided by the total group item standard deviation to obtain an effect-size value for the SMD (ES_{SMD}). The value of ES_{SMD} and the significance of the Mantel χ^2 statistic ($p < 0.05$) are then used to determine the DIF category of the item as depicted in Table 7 below.

Table 7. DIF Classification for CR Items

Category	Description	Criterion
AA	No DIF	Nonsignificant Mantel χ^2 or $ ES_{SMD} \leq 0.17$
BB	Moderate DIF	Significant Mantel χ^2 and $0.17 < ES_{SMD} \leq 0.25$
CC	High DIF	Significant Mantel χ^2 and $0.25 < ES_{SMD} $

Reliable DIF results are dependent on the number of examinees in both the focal and reference groups. Clauser and Mazor (1998) state that a minimum of 200 to 250 examinees per group are sufficient to provide reliable results. Some testing organizations require as many as 300 to 400 examinees per group (Zwick, 2012) in some applications. For the field testing of the Regents examinations, the sample sizes were such that only comparisons based on gender (e.g., males vs. females) were possible. Even for gender, sample sizes were only moderately large, and so the results should be interpreted with caution.

The DIF statistics for gender are shown in Appendix E. MC items in DIF categories “B” and “C” and CR items in categories “BB” and “CC” were flagged. These flags are shown in the “DIF Category” column (“A” and “AA” category items will have blank cells here). The “Favored Group” column indicates which gender is favored for items that are flagged.

Section III: Equating Procedure

Students participating in the 2013 field test administration for the New York State Regents Examination in English Language Arts received one of 18 test forms (numbered 601–606, 611–615, and 621–627)³. Each form included an embedded anchor form composed of 10 items that had been administered in previous administrations⁴. Because the items had been previously administered they had known parameters on the operational scale. The remaining items on the forms were items that had not been administered to New York State students. Test forms were spiraled within classrooms, so that students had an equal chance of receiving any of the 18 forms⁵, depending solely on their ordinal position within the classroom. In essence, students were randomly assigned to test forms, forming randomly equivalent groups taking each of the forms. Appendices A and D (with the classical and Rasch IRT item level statistics) may be consulted to determine the characteristics of the items (e.g., item type and maximum number of points possible) that made up each form.

COMMON ITEM EQUATING DESIGN

An equating design utilizing common items (as was done with this assessment) does not require the assumption of randomly equivalent groups, but that assumption is tenable and so should be noted. All that is required is a set of representative common items. In this case, 10 items (Form “N3”) formed this anchor set and were administered to all examinees along with their assigned set of field test items.

Using this design, the field test forms can either be calibrated individually or together. For the English Language Arts assessment, all forms were calibrated together. The data file for such an approach contains one record per examinee. Within the data records, each distinct column corresponds to one item that was administered in one or

³ Forms 613 and 626 each had one field test item (a six-point and a four-point CR item, respectively) that was invalidated (DNS status). Since these forms no longer had field test items, they (and their associated records) were dropped from the analyses.

⁴ Even though the field test items were based on the Common Core curriculum, the anchor test items were based on the previous Comprehensive English framework. In this case, the function of the anchor test was solely to fix the scale to a concrete metric. The Common Core-based assessments will be scaled after a standard setting meeting in June 2014, and even though the underlying theta distributions will be on a “common” scale, the scaled scores of the Common Core-based and non-Common Core-based assessments will exist in two entirely different metrics and will in no sense be comparable.

⁵ The field test sample was actually composed of three samples. Each sample received a different type of item(s) (either 30 MC items [Forms 601–606], one four-point item [Forms 611–615], or one six-point item [Forms 621–627]). Each set of forms was spiraled only across its own sample.

more of the forms. If the examinee was presented with the item, then his or her item score appears in that column. If the examinee did not encounter the item, then the score is replaced with a “not-presented” code (most commonly a blank space). Thus, all examinees had scores for the columns corresponding to the common item set, scores for the items that they were presented with, and blanks for items that were not on their individual test forms. The calibration only used the examinees that were presented with an item when estimating the items’ Rasch difficulties.

In either case, the difficulty parameters for the anchor items were fixed at their “known” parameters from their most recent use (in this case, the 2012 English Language Arts field test administration), and the parameters of the field test items were estimated relative to those of the anchor items. All forms were calibrated using Winsteps, version 3.60 (Linacre, 2005). Because it is possible for item parameters to “drift” (shift their difficulties relative to one another over time), a stability check was integrated into the analysis.

Winsteps provides an item level statistic, termed “displacement.” Linacre (2011, p. 545) describes this statistic as:

...the size of the change in the parameter estimate that would be observed in the next estimation iteration if this parameter was free (unanchored) and all other parameter estimates were anchored at their current values. For a parameter (item or person) that is anchored in the main estimation, (the displacement value) indicates the size of disagreement between an estimate based on the current data and the anchor value.

This statistic was used to identify items with difficulties that had shifted, relative to the difficulties of the other items on the form. After the initial calibration run, the Winsteps displacement values for all anchor form items were examined for absolute values greater than 0.30. If present, the item with the largest absolute displacement value was removed from anchored status but remained on the test form. Its difficulty value was subsequently reestimated relative to the difficulties of the remaining anchored items. The Winsteps calibration was then rerun with the reduced anchor set, after which the displacement values were again checked for absolute values in excess of 0.30. If another was found, it was also removed from anchored status and the calibration rerun. This iterative procedure continued until all anchored items had displacements of 0.30 or less. When the iterative procedure finishes, the parameters resulting from the final run are then in the operational metric, and the calibrations analyses are complete. No items were identified as having drifted for the 2013 analyses.

Section IV: Scaling of Operational Test Forms

Operational test items were selected based on content coverage, content accuracy, and statistical quality. The sets of items on each operational test conformed to the

coverage determined by content experts working from the learning standards established by the New York State Education Department and explicated in the test blueprint. Each item's classical and Rasch statistics were used to assess item quality. Items were selected to vary in difficulty to accurately measure students' abilities across the ability continuum. Appendix F contains the 2013 operational test maps for the January, June, and August administrations.

All Regents examinations have two cut scores, which are set at the scale scores of 65 and 85. One of the primary considerations during test construction was to select items so as to minimize changes in the raw scores corresponding to these two scale scores. Maintaining a consistent mean Rasch difficulty level from administration to administration facilitates this. For this assessment, the target value for the mean Rasch difficulty was set at 0.451. It should be noted that the raw scores corresponding to the scale score cut scores may still fluctuate even if the mean Rasch difficulty level is maintained at the target value due to differences in the distributions of the Rasch difficulty values amongst the items from administration to administration.

The relationship between raw and scale scores is explicated in the scoring tables for each administration. These tables can be found in Appendix G and cover the January, June, and August administrations. These tables are the end product of the following scaling procedure.

All Regents examinations are equated back to a base scale that is held constant from year to year. Specifically, they are equated to the base scale through the use of a calibrated item pool. The Rasch difficulties from the items' initial administration in a previous year's field test are used to equate the scale for the current administration to the base administration. For this examination, the base administration was the January 2011 administration. Scale scores from the 2013 administrations are on the same scale and can be directly compared to scale scores on all previous administrations back to and including the January 2011 administration.

When the base administration was concluded, the initial raw score-to-scale score relationship was established. Four raw scores were fixed at specific scale scores. Scale scores of 0 and 100 were fixed to correspond to the minimum and maximum possible raw scores. In addition, a standard setting had been held to determine the passing and passing with distinction cut scores in the raw score metric. The scale score points of 65 and 85 were set to correspond to those raw score cuts. A third degree polynomial is required in order to fit a line exactly to four arbitrary points (e.g., the raw scores corresponding to the four critical scale scores of 0, 65, 85, and 100). The general form of this best-fitting line is:

$$SS = m_3 * RS^3 + m_2 * RS^2 + m_1 * RS + m_0$$

where SS is the scaled score, RS is the raw score, and m_0 through m_3 are the transformation constants that convert the raw score into the scale score (please note that m_0 will always be equal to zero in this application since a raw score of zero

corresponds to a scale score of zero). The above relationship and the values of m_1 to m_3 specific to this subject were then used to determine the scale scores corresponding to the remainder of the raw scores on the examination. This initial relationship between the raw and scale scores became the base scale.

The Rasch difficulty parameters for the items on the base form were used to derive a raw score to Rasch student ability (theta score) relationship. This allowed the relationship between the Rasch theta score and the scale score to be known, mediated through their common relationship with the raw scores.

In succeeding years, each test form was selected from the pool of items that had been tested in previous years' field tests, each of which had known Rasch item difficulty parameter(s). These known parameters were used to construct the relationship between the raw and Rasch theta scores for that particular form⁶. Because the Rasch difficulty parameters are all on a common scale, the Rasch theta scores were also on a common scale with previously administered forms. The remaining step in the scaling process was to find the scale score equivalent for the Rasch theta score corresponding to each raw score point on the new form using the theta-to-scale score relationship established in the base year. This was done via linear interpolation.

This process results in a relationship between the raw scores on the form and the overall scale scores. The scale scores corresponding to each raw score are then rounded to the nearest integer for reporting on the conversion chart (posted at the close of each administration). The only exceptions are for the minimum and maximum raw scores and the raw scores that correspond to the scaled cut scores of 65 and 85.

The minimum (zero) and maximum possible raw scores are assigned scale scores of 0 and 100, respectively. In the event that there are raw scores less than the maximum with scale scores that round to 100, their scale scores are set equal to 99. A similar process is followed with the minimum score; if any raw scores other than zero have scale scores that round to zero, their scale scores are instead set equal to one.

With regard to the cuts, if two or more scale scores round to either 65 or 85, the lowest raw score's scale score is set equal to a 65 or 85 and the scale scores corresponding to the higher raw scores are set to 66 or 86 as appropriate. If no scale score rounds to either of these two critical cuts, then the raw score with the largest scale score that is less than the cut is set equal to the cut. The overarching principle when two raw scores both round to either scale score cut is that the lower of the raw scores is

⁶ All Regents examinations are pre-equated, meaning that the parameters used to derive the relationship between the raw and scale scores are estimated prior to the construction and administration of the operational form. These field tests are administered to as small a sample of students as possible in order to minimize the impact on student instructional time across the state. The small N-Counts associated with such administrations are sufficient for reasonably accurate estimation of most items' parameters; however, for the six-point essay item, its parameters can be unstable when estimated across as small a sample as is typically used. Therefore, a set of constants is used for these items' parameters on operational examinations. These constants were set by the NYSED and are based on the values in the bank for all essay items.

always assigned to be equal to the cut so that students are never penalized for this ambiguity.

References

- Allen, N. L., Carlson, J. E., & Zalanak, C. A. (1999). *The NAEP 1996 technical report*. Washington, DC: National Center for Education Statistics.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, 17(1), 31–44.
- 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., & Schmitt, A. P. (1991). *Constructed-response and differential item functioning: A pragmatic approach* (ETS Research Report No. 91-49). Princeton, NJ: Educational Testing Service.
- Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. 2nd ed. New York: John Wiley.
- Hambleton, R. K, Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.
- Linacre, J. M. (2005). WINSTEPS Rasch measurement computer program and manual (PDF file) v 3.60. Chicago: Winsteps.com
- Linacre, J. M. (2011). *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs: Program manual 3.73.0* (PDF file). Chicago: Winsteps.com
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30–46.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.

- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567–577.
- Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (ETS Research Report No. 12-08). Princeton, NJ: Educational Testing Service.
- Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30, 233–251.

Appendix A: Classical Item Analysis

In the following table, “Max” is the maximum number of possible points. “N-Count” refers to the number of student records in the analysis. “Alpha” contains Cronbach’s Coefficient α (since this is a test [form] level statistic, it has the same value for all items within each form). For MC items, “B” represents the proportion of students who left the item blank, and “M1” through “M6” are the proportions of students who selected each of the four answer choices. For CR items, “B” represents the proportion of students who left the item blank, and “M0” through “M6” are the proportions of students who received scores of 0 through 6. “Mean” is the average of the scores received by the students. The final (right) column contains the Point-Biserial correlation for each item. There may be some instances of items with missing statistics; this occurs when an item was not scored.

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Bis
2013_Engl	601	MC	01	1	747	0.88	0.00		0.07	0.78	0.04	0.10			0.78	0.38
2013_Engl	601	MC	02	1	747	0.88	0.00		0.16	0.06	0.11	0.67			0.67	0.44
2013_Engl	601	MC	03	1	747	0.88	0.00		0.22	0.18	0.50	0.10			0.50	0.41
2013_Engl	601	MC	04	1	747	0.88	0.00		0.59	0.06	0.21	0.14			0.59	0.41
2013_Engl	601	MC	05	1	747	0.88	0.00		0.16	0.38	0.37	0.09			0.37	0.30
2013_Engl	601	MC	06	1	747	0.88	0.01		0.23	0.52	0.15	0.09			0.52	0.32
2013_Engl	601	MC	07	1	747	0.88	0.01		0.19	0.15	0.13	0.52			0.52	0.42
2013_Engl	601	MC	08	1	747	0.88	0.01		0.34	0.09	0.23	0.34			0.34	0.13
2013_Engl	601	MC	09	1	747	0.88	0.01		0.56	0.10	0.18	0.14			0.56	0.50
2013_Engl	601	MC	10	1	747	0.88	0.01		0.23	0.13	0.18	0.45			0.45	0.42
2013_Engl	601	MC	11	1	747	0.88	0.01		0.54	0.22	0.13	0.09			0.54	0.47
2013_Engl	601	MC	12	1	747	0.88	0.01		0.13	0.08	0.71	0.07			0.71	0.44
2013_Engl	601	MC	13	1	747	0.88	0.01		0.17	0.41	0.09	0.32			0.41	0.44
2013_Engl	601	MC	14	1	747	0.88	0.01		0.14	0.52	0.16	0.17			0.52	0.44
2013_Engl	601	MC	15	1	747	0.88	0.01		0.29	0.14	0.37	0.18			0.29	0.38
2013_Engl	601	MC	16	1	747	0.88	0.01		0.14	0.43	0.20	0.21			0.43	0.19
2013_Engl	601	MC	17	1	747	0.88	0.02		0.28	0.12	0.13	0.45			0.45	0.36

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Bis
2013_Engl	601	MC	18	1	747	0.88	0.03		0.22	0.18	0.34	0.23			0.23	0.41
2013_Engl	601	MC	19	1	747	0.88	0.05		0.37	0.27	0.15	0.16			0.37	0.28
2013_Engl	601	MC	20	1	747	0.88	0.05		0.14	0.11	0.32	0.37			0.37	0.33
2013_Engl	601	MC	21	1	747	0.88	0.05		0.14	0.55	0.18	0.07			0.55	0.48
2013_Engl	601	MC	22	1	747	0.88	0.05		0.28	0.32	0.09	0.26			0.32	0.36
2013_Engl	601	MC	23	1	747	0.88	0.05		0.13	0.12	0.22	0.48			0.48	0.39
2013_Engl	601	MC	24	1	747	0.88	0.05		0.29	0.14	0.39	0.12			0.39	0.43
2013_Engl	601	MC	25	1	747	0.88	0.06		0.59	0.16	0.11	0.08			0.59	0.47
2013_Engl	601	MC	26	1	747	0.88	0.05		0.20	0.49	0.10	0.16			0.49	0.46
2013_Engl	601	MC	27	1	747	0.88	0.06		0.46	0.18	0.23	0.07			0.46	0.46
2013_Engl	601	MC	28	1	747	0.88	0.06		0.10	0.14	0.59	0.11			0.59	0.52
2013_Engl	601	MC	29	1	747	0.88	0.06		0.21	0.30	0.36	0.06			0.36	0.33
2013_Engl	601	MC	30	1	747	0.88	0.06		0.16	0.15	0.11	0.52			0.52	0.51
2013_Engl	602	MC	01	1	752	0.87	0.01		0.07	0.45	0.05	0.43			0.45	0.33
2013_Engl	602	MC	02	1	752	0.87	0.01		0.54	0.25	0.12	0.08			0.54	0.31
2013_Engl	602	MC	03	1	752	0.87	0.01		0.17	0.13	0.55	0.13			0.55	0.44
2013_Engl	602	MC	04	1	752	0.87	0.01		0.74	0.10	0.11	0.05			0.74	0.49
2013_Engl	602	MC	05	1	752	0.87	0.01		0.15	0.34	0.18	0.32			0.32	0.18
2013_Engl	602	MC	06	1	752	0.87	0.00		0.19	0.46	0.09	0.26			0.46	0.17
2013_Engl	602	MC	07	1	752	0.87	0.01		0.36	0.43	0.12	0.08			0.43	0.23
2013_Engl	602	MC	08	1	752	0.87	0.01		0.14	0.17	0.59	0.09			0.59	0.28
2013_Engl	602	MC	09	1	752	0.87	0.01		0.16	0.43	0.16	0.25			0.43	0.34
2013_Engl	602	MC	10	1	752	0.87	0.01		0.12	0.07	0.13	0.68			0.68	0.53
2013_Engl	602	MC	11	1	752	0.87	0.02		0.62	0.16	0.11	0.10			0.62	0.50
2013_Engl	602	MC	12	1	752	0.87	0.01		0.51	0.23	0.14	0.10			0.23	0.18
2013_Engl	602	MC	13	1	752	0.87	0.02		0.12	0.69	0.09	0.09			0.69	0.43
2013_Engl	602	MC	14	1	752	0.87	0.02		0.07	0.04	0.75	0.13			0.75	0.38

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Bis
2013_Engl	602	MC	15	1	752	0.87	0.02		0.64	0.12	0.10	0.12			0.64	0.52
2013_Engl	602	MC	16	1	752	0.87	0.02		0.08	0.20	0.10	0.60			0.60	0.55
2013_Engl	602	MC	17	1	752	0.87	0.02		0.15	0.06	0.10	0.67			0.67	0.51
2013_Engl	602	MC	18	1	752	0.87	0.02		0.50	0.21	0.09	0.18			0.50	0.37
2013_Engl	602	MC	19	1	752	0.87	0.03		0.16	0.19	0.54	0.09			0.54	0.35
2013_Engl	602	MC	20	1	752	0.87	0.05		0.07	0.10	0.71	0.07			0.71	0.49
2013_Engl	602	MC	21	1	752	0.87	0.05		0.35	0.19	0.25	0.15			0.35	0.28
2013_Engl	602	MC	22	1	752	0.87	0.06		0.39	0.15	0.25	0.15			0.15	0.16
2013_Engl	602	MC	23	1	752	0.87	0.06		0.11	0.11	0.13	0.58			0.58	0.50
2013_Engl	602	MC	24	1	752	0.87	0.07		0.25	0.23	0.34	0.12			0.34	0.29
2013_Engl	602	MC	25	1	752	0.87	0.07		0.18	0.17	0.32	0.26			0.32	0.33
2013_Engl	602	MC	26	1	752	0.87	0.07		0.40	0.13	0.30	0.10			0.40	0.41
2013_Engl	602	MC	27	1	752	0.87	0.07		0.12	0.28	0.09	0.45			0.45	0.42
2013_Engl	602	MC	28	1	752	0.87	0.07		0.15	0.33	0.20	0.24			0.33	0.38
2013_Engl	602	MC	29	1	752	0.87	0.07		0.32	0.24	0.12	0.25			0.32	0.35
2013_Engl	602	MC	30	1	752	0.87	0.07		0.26	0.12	0.40	0.15			0.40	0.47
2013_Engl	603	MC	01	1	750	0.89	0.00		0.07	0.68	0.10	0.14			0.68	0.36
2013_Engl	603	MC	02	1	750	0.89	0.00		0.30	0.23	0.38	0.09			0.30	0.22
2013_Engl	603	MC	03	1	750	0.89	0.00		0.16	0.10	0.09	0.65			0.65	0.30
2013_Engl	603	MC	04	1	750	0.89	0.01		0.19	0.59	0.14	0.07			0.59	0.35
2013_Engl	603	MC	05	1	750	0.89	0.01		0.13	0.11	0.67	0.09			0.67	0.43
2013_Engl	603	MC	06	1	750	0.89	0.01		0.76	0.08	0.06	0.09			0.76	0.48
2013_Engl	603	MC	07	1	750	0.89	0.01		0.11	0.38	0.16	0.34			0.34	0.32
2013_Engl	603	MC	08	1	750	0.89	0.01		0.64	0.15	0.06	0.14			0.64	0.47
2013_Engl	603	MC	09	1	750	0.89	0.01		0.14	0.51	0.17	0.17			0.51	0.41
2013_Engl	603	MC	10	1	750	0.89	0.02		0.12	0.12	0.17	0.57			0.57	0.47
2013_Engl	603	MC	11	1	750	0.89	0.02		0.42	0.16	0.16	0.23			0.42	0.26

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Bis
2013_Engl	603	MC	12	1	750	0.89	0.02		0.45	0.19	0.21	0.13			0.21	0.15
2013_Engl	603	MC	13	1	750	0.89	0.02		0.23	0.21	0.12	0.44			0.44	0.36
2013_Engl	603	MC	14	1	750	0.89	0.02		0.12	0.65	0.07	0.14			0.65	0.36
2013_Engl	603	MC	15	1	750	0.89	0.02		0.12	0.07	0.13	0.66			0.66	0.53
2013_Engl	603	MC	16	1	750	0.89	0.02		0.32	0.29	0.32	0.05			0.29	0.25
2013_Engl	603	MC	17	1	750	0.89	0.03		0.47	0.06	0.36	0.08			0.47	0.30
2013_Engl	603	MC	18	1	750	0.89	0.04		0.11	0.42	0.21	0.22			0.42	0.46
2013_Engl	603	MC	19	1	750	0.89	0.04		0.08	0.20	0.56	0.11			0.56	0.35
2013_Engl	603	MC	20	1	750	0.89	0.04		0.36	0.10	0.13	0.37			0.37	0.45
2013_Engl	603	MC	21	1	750	0.89	0.06		0.10	0.13	0.56	0.15			0.56	0.59
2013_Engl	603	MC	22	1	750	0.89	0.06		0.59	0.08	0.12	0.16			0.59	0.55
2013_Engl	603	MC	23	1	750	0.89	0.06		0.47	0.17	0.11	0.18			0.47	0.58
2013_Engl	603	MC	24	1	750	0.89	0.06		0.14	0.52	0.15	0.13			0.52	0.42
2013_Engl	603	MC	25	1	750	0.89	0.06		0.07	0.12	0.69	0.06			0.69	0.62
2013_Engl	603	MC	26	1	750	0.89	0.06		0.18	0.26	0.29	0.21			0.26	0.21
2013_Engl	603	MC	27	1	750	0.89	0.06		0.17	0.10	0.61	0.05			0.61	0.49
2013_Engl	603	MC	28	1	750	0.89	0.07		0.21	0.10	0.13	0.49			0.49	0.42
2013_Engl	603	MC	29	1	750	0.89	0.07		0.64	0.18	0.06	0.05			0.64	0.63
2013_Engl	603	MC	30	1	750	0.89	0.07		0.37	0.35	0.12	0.10			0.37	0.30
2013_Engl	604	MC	01	1	752	0.84	0.01		0.13	0.51	0.21	0.13			0.51	0.26
2013_Engl	604	MC	02	1	752	0.84	0.00		0.13	0.22	0.07	0.57			0.57	0.33
2013_Engl	604	MC	03	1	752	0.84	0.01		0.59	0.14	0.17	0.10			0.59	0.29
2013_Engl	604	MC	04	1	752	0.84	0.01		0.21	0.12	0.54	0.13			0.54	0.22
2013_Engl	604	MC	05	1	752	0.84	0.01		0.59	0.32	0.04	0.05			0.59	0.30
2013_Engl	604	MC	06	1	752	0.84	0.01		0.27	0.50	0.13	0.10			0.50	0.36
2013_Engl	604	MC	07	1	752	0.84	0.01		0.14	0.19	0.10	0.56			0.56	0.49
2013_Engl	604	MC	08	1	752	0.84	0.01		0.11	0.24	0.44	0.21			0.44	0.11

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Bis
2013_Engl	604	MC	09	1	752	0.84	0.01		0.37	0.19	0.13	0.30			0.37	0.39
2013_Engl	604	MC	10	1	752	0.84	0.01		0.10	0.40	0.32	0.17			0.40	0.29
2013_Engl	604	MC	11	1	752	0.84	0.01		0.44	0.19	0.19	0.17			0.44	0.27
2013_Engl	604	MC	12	1	752	0.84	0.01		0.14	0.14	0.48	0.23			0.48	0.17
2013_Engl	604	MC	13	1	752	0.84	0.01		0.14	0.48	0.34	0.03			0.48	0.04
2013_Engl	604	MC	14	1	752	0.84	0.02		0.23	0.31	0.15	0.29			0.29	0.19
2013_Engl	604	MC	15	1	752	0.84	0.01		0.79	0.03	0.03	0.14			0.79	0.35
2013_Engl	604	MC	16	1	752	0.84	0.02		0.30	0.31	0.16	0.21			0.31	0.20
2013_Engl	604	MC	17	1	752	0.84	0.02		0.11	0.11	0.62	0.14			0.62	0.45
2013_Engl	604	MC	18	1	752	0.84	0.02		0.24	0.40	0.30	0.03			0.30	0.22
2013_Engl	604	MC	19	1	752	0.84	0.02		0.28	0.59	0.09	0.02			0.59	0.26
2013_Engl	604	MC	20	1	752	0.84	0.04		0.57	0.13	0.21	0.05			0.57	0.35
2013_Engl	604	MC	21	1	752	0.84	0.04		0.10	0.10	0.23	0.53			0.53	0.49
2013_Engl	604	MC	22	1	752	0.84	0.05		0.55	0.23	0.09	0.08			0.55	0.47
2013_Engl	604	MC	23	1	752	0.84	0.05		0.08	0.67	0.14	0.06			0.67	0.50
2013_Engl	604	MC	24	1	752	0.84	0.06		0.49	0.20	0.14	0.11			0.49	0.39
2013_Engl	604	MC	25	1	752	0.84	0.06		0.18	0.28	0.40	0.08			0.40	0.36
2013_Engl	604	MC	26	1	752	0.84	0.06		0.14	0.11	0.09	0.60			0.60	0.62
2013_Engl	604	MC	27	1	752	0.84	0.06		0.10	0.63	0.10	0.11			0.63	0.61
2013_Engl	604	MC	28	1	752	0.84	0.06		0.20	0.20	0.11	0.43			0.43	0.47
2013_Engl	604	MC	29	1	752	0.84	0.07		0.24	0.27	0.28	0.14			0.27	0.20
2013_Engl	604	MC	30	1	752	0.84	0.06		0.51	0.25	0.11	0.07			0.51	0.39
2013_Engl	605	MC	01	1	754	0.88	0.01		0.28	0.28	0.21	0.22			0.21	0.28
2013_Engl	605	MC	02	1	754	0.88	0.00		0.62	0.09	0.05	0.23			0.62	0.51
2013_Engl	605	MC	03	1	754	0.88	0.01		0.27	0.32	0.34	0.05			0.34	0.31
2013_Engl	605	MC	04	1	754	0.88	0.01		0.03	0.11	0.03	0.82			0.82	0.45
2013_Engl	605	MC	05	1	754	0.88	0.01		0.04	0.61	0.20	0.14			0.61	0.37

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Bis
2013_Engl	605	MC	06	1	754	0.88	0.01		0.72	0.17	0.03	0.06			0.72	0.43
2013_Engl	605	MC	07	1	754	0.88	0.01		0.35	0.18	0.07	0.38			0.38	0.38
2013_Engl	605	MC	08	1	754	0.88	0.01		0.09	0.08	0.79	0.03			0.79	0.44
2013_Engl	605	MC	09	1	754	0.88	0.02		0.45	0.17	0.21	0.15			0.45	0.47
2013_Engl	605	MC	10	1	754	0.88	0.02		0.09	0.34	0.21	0.34			0.34	0.31
2013_Engl	605	MC	11	1	754	0.88	0.02		0.19	0.47	0.04	0.28			0.47	0.44
2013_Engl	605	MC	12	1	754	0.88	0.02		0.15	0.20	0.24	0.40			0.40	0.23
2013_Engl	605	MC	13	1	754	0.88	0.02		0.68	0.07	0.10	0.12			0.68	0.51
2013_Engl	605	MC	14	1	754	0.88	0.03		0.06	0.06	0.36	0.50			0.50	0.17
2013_Engl	605	MC	15	1	754	0.88	0.02		0.80	0.09	0.07	0.02			0.80	0.42
2013_Engl	605	MC	16	1	754	0.88	0.02		0.28	0.33	0.12	0.25			0.33	0.21
2013_Engl	605	MC	17	1	754	0.88	0.02		0.17	0.47	0.10	0.23			0.47	0.28
2013_Engl	605	MC	18	1	754	0.88	0.03		0.14	0.53	0.17	0.13			0.17	0.08
2013_Engl	605	MC	19	1	754	0.88	0.03		0.08	0.62	0.20	0.07			0.62	0.24
2013_Engl	605	MC	20	1	754	0.88	0.06		0.22	0.37	0.20	0.15			0.22	0.09
2013_Engl	605	MC	21	1	754	0.88	0.05		0.52	0.14	0.15	0.14			0.52	0.41
2013_Engl	605	MC	22	1	754	0.88	0.05		0.13	0.53	0.22	0.07			0.53	0.47
2013_Engl	605	MC	23	1	754	0.88	0.05		0.06	0.19	0.17	0.53			0.53	0.54
2013_Engl	605	MC	24	1	754	0.88	0.05		0.10	0.29	0.43	0.14			0.43	0.45
2013_Engl	605	MC	25	1	754	0.88	0.05		0.12	0.13	0.07	0.63			0.63	0.59
2013_Engl	605	MC	26	1	754	0.88	0.06		0.11	0.08	0.08	0.68			0.68	0.62
2013_Engl	605	MC	27	1	754	0.88	0.06		0.21	0.20	0.42	0.11			0.42	0.34
2013_Engl	605	MC	28	1	754	0.88	0.06		0.11	0.18	0.16	0.48			0.48	0.54
2013_Engl	605	MC	29	1	754	0.88	0.06		0.08	0.11	0.66	0.10			0.66	0.56
2013_Engl	605	MC	30	1	754	0.88	0.06		0.27	0.12	0.05	0.50			0.27	0.20
2013_Engl	606	MC	01	1	759	0.87	0.01		0.12	0.12	0.15	0.61			0.61	0.46
2013_Engl	606	MC	02	1	759	0.87	0.01		0.68	0.12	0.11	0.08			0.68	0.39

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Bis
2013_Engl	606	MC	03	1	759	0.87	0.01		0.16	0.57	0.14	0.12			0.57	0.32
2013_Engl	606	MC	04	1	759	0.87	0.01		0.19	0.13	0.20	0.47			0.47	0.27
2013_Engl	606	MC	05	1	759	0.87	0.01		0.28	0.22	0.33	0.16			0.28	0.21
2013_Engl	606	MC	06	1	759	0.87	0.01		0.13	0.10	0.70	0.07			0.70	0.40
2013_Engl	606	MC	07	1	759	0.87	0.01		0.20	0.15	0.57	0.07			0.57	0.50
2013_Engl	606	MC	08	1	759	0.87	0.01		0.15	0.09	0.33	0.43			0.43	0.37
2013_Engl	606	MC	09	1	759	0.87	0.01		0.47	0.23	0.11	0.19			0.47	0.40
2013_Engl	606	MC	10	1	759	0.87	0.01		0.10	0.69	0.08	0.11			0.69	0.50
2013_Engl	606	MC	11	1	759	0.87	0.02		0.13	0.13	0.19	0.53			0.53	0.53
2013_Engl	606	MC	12	1	759	0.87	0.02		0.23	0.33	0.33	0.08			0.33	0.12
2013_Engl	606	MC	13	1	759	0.87	0.02		0.16	0.15	0.46	0.20			0.46	0.18
2013_Engl	606	MC	14	1	759	0.87	0.01		0.17	0.40	0.10	0.31			0.40	0.41
2013_Engl	606	MC	15	1	759	0.87	0.02		0.15	0.45	0.19	0.19			0.45	0.49
2013_Engl	606	MC	16	1	759	0.87	0.02		0.29	0.13	0.38	0.18			0.29	0.30
2013_Engl	606	MC	17	1	759	0.87	0.02		0.14	0.39	0.25	0.20			0.39	0.20
2013_Engl	606	MC	18	1	759	0.87	0.02		0.32	0.12	0.12	0.41			0.41	0.37
2013_Engl	606	MC	19	1	759	0.87	0.02		0.23	0.19	0.33	0.23			0.23	0.37
2013_Engl	606	MC	20	1	759	0.87	0.05		0.25	0.14	0.49	0.06			0.49	0.48
2013_Engl	606	MC	21	1	759	0.87	0.05		0.14	0.17	0.11	0.52			0.52	0.55
2013_Engl	606	MC	22	1	759	0.87	0.05		0.16	0.22	0.19	0.37			0.37	0.42
2013_Engl	606	MC	23	1	759	0.87	0.06		0.37	0.29	0.17	0.11			0.37	0.34
2013_Engl	606	MC	24	1	759	0.87	0.06		0.15	0.19	0.45	0.14			0.45	0.39
2013_Engl	606	MC	25	1	759	0.87	0.06		0.54	0.19	0.09	0.11			0.54	0.53
2013_Engl	606	MC	26	1	759	0.87	0.06		0.16	0.40	0.18	0.20			0.40	0.43
2013_Engl	606	MC	27	1	759	0.87	0.07		0.29	0.14	0.19	0.31			0.31	0.30
2013_Engl	606	MC	28	1	759	0.87	0.07		0.47	0.19	0.15	0.12			0.47	0.57
2013_Engl	606	MC	29	1	759	0.87	0.07		0.20	0.24	0.19	0.31			0.24	0.09
2013_Engl	606	MC	30	1	759	0.87	0.07		0.13	0.31	0.36	0.13			0.31	0.41

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Bis
2013_Engl	611	CR	Ar	6	641	0.68	0.00	0.02	0.18	0.20	0.33	0.20	0.06	0.01	2.73	0.77
2013_Engl	612	CR	Ar	6	664	0.66	0.00	0.02	0.08	0.22	0.41	0.20	0.07	0.00	2.93	0.72
2013_Engl	613	CR	Ar													
2013_Engl	614	CR	Ar	6	668	0.68	0.00	0.02	0.08	0.17	0.45	0.22	0.06	0.01	3.00	0.72
2013_Engl	615	CR	Ar	6	653	0.68	0.00	0.02	0.10	0.18	0.34	0.27	0.08	0.01	3.02	0.72
2013_Engl	621	CR	Re	4	436	0.68	0.00	0.03	0.21	0.58	0.16	0.03			1.95	0.59
2013_Engl	622	CR	Re	4	587	0.66	0.00	0.04	0.20	0.46	0.24	0.05			2.05	0.66
2013_Engl	623	CR	Re	4	414	0.67	0.00	0.03	0.22	0.60	0.13	0.02			1.88	0.62
2013_Engl	624	CR	Re	4	600	0.67	0.00	0.04	0.38	0.45	0.12	0.02			1.69	0.65
2013_Engl	625	CR	Re	4	611	0.64	0.00	0.04	0.35	0.48	0.11	0.02			1.73	0.62
2013_Engl	626	CR	Re													
2013_Engl	627	CR	Re	4	533	0.64	0.00	0.02	0.13	0.59	0.21	0.05			2.14	0.64
2013_Engl	N3	MC	01	1	10321	0.75	0.03		0.18	0.04	0.05	0.70			0.70	0.55
2013_Engl	N3	MC	02	1	10321	0.75	0.04		0.26	0.25	0.37	0.08			0.37	0.41
2013_Engl	N3	MC	03	1	10321	0.75	0.03		0.06	0.79	0.04	0.07			0.79	0.62
2013_Engl	N3	MC	04	1	10321	0.75	0.04		0.71	0.11	0.07	0.07			0.71	0.55
2013_Engl	N3	MC	05	1	10321	0.75	0.04		0.08	0.63	0.19	0.06			0.63	0.54
2013_Engl	N3	MC	06	1	10321	0.75	0.04		0.09	0.16	0.24	0.48			0.48	0.55
2013_Engl	N3	MC	07	1	10321	0.75	0.04		0.59	0.14	0.14	0.09			0.59	0.61
2013_Engl	N3	MC	08	1	10321	0.75	0.04		0.23	0.11	0.14	0.48			0.48	0.48
2013_Engl	N3	MC	09	1	10321	0.75	0.04		0.21	0.60	0.05	0.10			0.60	0.60
2013_Engl	N3	MC	10	1	10321	0.75	0.04		0.16	0.09	0.54	0.17			0.54	0.62

Appendix B: Inter-Rater Consistency – Point Differences Between First and Second Reads

The first three columns from the left contain the form ID, item sequence number, and number of score points for each item. The remaining columns contain the percentage of times each possible difference between the first and second raters' scores occurred. Blank cells indicate out-of-range differences (e.g., differences greater than the maximum possible given the point value of that particular item).

Form	Item	Score Pts	Difference (First Read Minus Second Read)												
			-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
611	11	6	0%	0%	0%	0%	0%	18%	66%	16%	0%	0%	0%	0%	0%
612	11	6	0%	0%	0%	0%	0%	18%	61%	21%	0%	0%	0%	0%	0%
614	11	6	0%	0%	0%	0%	0%	14%	69%	16%	1%	0%	0%	0%	0%
615	11	6	0%	0%	0%	0%	0%	16%	73%	11%	0%	0%	0%	0%	0%
621	11	4			0%	0%	0%	11%	75%	15%	0%	0%	0%		
622	11	4			0%	0%	0%	12%	68%	20%	0%	0%	0%		
623	11	4			0%	0%	0%	14%	74%	13%	0%	0%	0%		
624	11	4			0%	0%	0%	14%	75%	11%	0%	0%	0%		
625	11	4			0%	0%	1%	15%	68%	15%	1%	0%	0%		
627	11	4			0%	0%	0%	8%	74%	19%	0%	0%	0%		

Appendix C: Additional Measures of Inter-Rater Reliability and Agreement

The first four columns from the left contain the form ID, item sequence number, number of score points, and the total count of items receiving a first and second read. In the fifth column, the percent of exact matches between the first and second scores is provided. The following column (“Adj.”) is the percentage of the first and second scores with a difference of –1 or 1. “Total” is the sum of the Exact and Adjacent matches (e.g., the two prior columns).

Form	Item	Score Points	Total N-Count	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intraclass Corr	Wt Kappa
				Exact	Adj	Total	First Read	Second Read	First Read	Second Read		
611	11	6	112	66.1%	33.9%	100.0%	2.7	2.7	1.20	1.23	0.88	0.74
612	11	6	116	61.2%	38.8%	100.0%	3.0	2.9	1.06	0.98	0.81	0.65
614	11	6	118	69.5%	29.7%	99.2%	3.1	3.0	1.05	1.07	0.85	0.72
615	11	6	116	73.3%	26.7%	100.0%	2.9	2.9	1.11	1.11	0.89	0.78
621	11	4	75	74.7%	25.3%	100.0%	1.9	1.9	0.69	0.69	0.73	0.63
622	11	4	105	67.6%	32.4%	100.0%	2.1	2.0	0.91	0.89	0.80	0.66
623	11	4	72	73.6%	26.4%	100.0%	2.0	2.0	0.58	0.66	0.66	0.57
624	11	4	109	75.2%	24.8%	100.0%	1.8	1.8	0.85	0.86	0.83	0.72
625	11	4	106	67.9%	30.2%	98.1%	1.8	1.8	0.76	0.73	0.66	0.55
627	11	4	91	73.6%	26.4%	100.0%	2.2	2.1	0.69	0.70	0.73	0.62

Appendix D: Partial-Credit Model Item Analysis

The first five columns from the left contain the test name, form name, item type, item number on the form, and maximum points possible for the item. The sixth column contains the number of students that the item was administered to. The remaining eight columns contain the Rasch Item Difficulty, step difficulties (for multi-point items only), and the INFIT Rasch model fit statistic. Items without statistics are DNS (Do Not Score) status items.

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2013_Engl	601	MC	01	1	747	-1.3602							1.01
2013_Engl	601	MC	02	1	747	-0.6787							0.98
2013_Engl	601	MC	03	1	747	0.1831							1.03
2013_Engl	601	MC	04	1	747	-0.2779							1.03
2013_Engl	601	MC	05	1	747	0.8485							1.13
2013_Engl	601	MC	06	1	747	0.0770							1.13
2013_Engl	601	MC	07	1	747	0.0770							1.03
2013_Engl	601	MC	08	1	747	1.0149							1.34
2013_Engl	601	MC	09	1	747	-0.1495							0.93
2013_Engl	601	MC	10	1	747	0.4360							1.02
2013_Engl	601	MC	11	1	747	-0.0426							0.97
2013_Engl	601	MC	12	1	747	-0.9364							0.97
2013_Engl	601	MC	13	1	747	0.6461							0.99
2013_Engl	601	MC	14	1	747	0.0638							0.99
2013_Engl	601	MC	15	1	747	1.2823							1.01
2013_Engl	601	MC	16	1	747	0.5235							1.27
2013_Engl	601	MC	17	1	747	0.4427							1.08
2013_Engl	601	MC	18	1	747	1.6278							0.96
2013_Engl	601	MC	19	1	747	0.8485							1.17
2013_Engl	601	MC	20	1	747	0.8556							1.11
2013_Engl	601	MC	21	1	747	-0.0759							0.95

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2013_Engl	601	MC	22	1	747	1.1271							1.05
2013_Engl	601	MC	23	1	747	0.2626							1.05
2013_Engl	601	MC	24	1	747	0.7221							1.00
2013_Engl	601	MC	25	1	747	-0.2643							0.95
2013_Engl	601	MC	26	1	747	0.2228							0.98
2013_Engl	601	MC	27	1	747	0.3825							0.98
2013_Engl	601	MC	28	1	747	-0.2711							0.90
2013_Engl	601	MC	29	1	747	0.8698							1.10
2013_Engl	601	MC	30	1	747	0.0903							0.92
2013_Engl	602	MC	01	1	752	0.3207							1.09
2013_Engl	602	MC	02	1	752	-0.1429							1.12
2013_Engl	602	MC	03	1	752	-0.2153							1.00
2013_Engl	602	MC	04	1	752	-1.1961							0.93
2013_Engl	602	MC	05	1	752	0.9781							1.23
2013_Engl	602	MC	06	1	752	0.2617							1.27
2013_Engl	602	MC	07	1	752	0.3798							1.19
2013_Engl	602	MC	08	1	752	-0.3887							1.16
2013_Engl	602	MC	09	1	752	0.4194							1.08
2013_Engl	602	MC	10	1	752	-0.8528							0.88
2013_Engl	602	MC	11	1	752	-0.5526							0.92
2013_Engl	602	MC	12	1	752	1.4928							1.12
2013_Engl	602	MC	13	1	752	-0.9353							0.99
2013_Engl	602	MC	14	1	752	-1.2970							1.00
2013_Engl	602	MC	15	1	752	-0.6788							0.91
2013_Engl	602	MC	16	1	752	-0.4564							0.87
2013_Engl	602	MC	17	1	752	-0.8306							0.90
2013_Engl	602	MC	18	1	752	0.0597							1.06
2013_Engl	602	MC	19	1	752	-0.1298							1.08

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2013_Engl	602	MC	20	1	752	-1.0590							0.89
2013_Engl	602	MC	21	1	752	0.8063							1.11
2013_Engl	602	MC	22	1	752	2.0795							1.08
2013_Engl	602	MC	23	1	752	-0.3618							0.93
2013_Engl	602	MC	24	1	752	0.8699							1.11
2013_Engl	602	MC	25	1	752	0.9490							1.06
2013_Engl	602	MC	26	1	752	0.5325							1.00
2013_Engl	602	MC	27	1	752	0.3207							1.00
2013_Engl	602	MC	28	1	752	0.9128							1.01
2013_Engl	602	MC	29	1	752	0.9635							1.05
2013_Engl	602	MC	30	1	752	0.5527							0.94
2013_Engl	603	MC	01	1	750	-0.9113							1.10
2013_Engl	603	MC	02	1	750	1.0886							1.19
2013_Engl	603	MC	03	1	750	-0.7163							1.16
2013_Engl	603	MC	04	1	750	-0.3962							1.13
2013_Engl	603	MC	05	1	750	-0.8127							1.03
2013_Engl	603	MC	06	1	750	-1.3898							0.91
2013_Engl	603	MC	07	1	750	0.8866							1.12
2013_Engl	603	MC	08	1	750	-0.6436							0.97
2013_Engl	603	MC	09	1	750	0.0053							1.05
2013_Engl	603	MC	10	1	750	-0.2928							0.99
2013_Engl	603	MC	11	1	750	0.4711							1.21
2013_Engl	603	MC	12	1	750	1.6692							1.19
2013_Engl	603	MC	13	1	750	0.3825							1.08
2013_Engl	603	MC	14	1	750	-0.7237							1.09
2013_Engl	603	MC	15	1	750	-0.7828							0.89
2013_Engl	603	MC	16	1	750	1.1819							1.11
2013_Engl	603	MC	17	1	750	0.2001							1.17

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2013_Engl	603	MC	18	1	750	0.4642							0.98
2013_Engl	603	MC	19	1	750	-0.2586							1.12
2013_Engl	603	MC	20	1	750	0.7425							0.98
2013_Engl	603	MC	21	1	750	-0.2245							0.84
2013_Engl	603	MC	22	1	750	-0.3823							0.88
2013_Engl	603	MC	23	1	750	0.1934							0.85
2013_Engl	603	MC	24	1	750	-0.0216							1.04
2013_Engl	603	MC	25	1	750	-0.9266							0.79
2013_Engl	603	MC	26	1	750	1.3684							1.18
2013_Engl	603	MC	27	1	750	-0.5222							0.96
2013_Engl	603	MC	28	1	750	0.0859							1.04
2013_Engl	603	MC	29	1	750	-0.6726							0.79
2013_Engl	603	MC	30	1	750	0.7213							1.15
2013_Engl	604	MC	01	1	752	0.0356							1.12
2013_Engl	604	MC	02	1	752	-0.2610							1.05
2013_Engl	604	MC	03	1	752	-0.3447							1.09
2013_Engl	604	MC	04	1	752	-0.0773							1.15
2013_Engl	604	MC	05	1	752	-0.3124							1.08
2013_Engl	604	MC	06	1	752	0.1105							1.02
2013_Engl	604	MC	07	1	752	-0.1719							0.91
2013_Engl	604	MC	08	1	752	0.3862							1.24
2013_Engl	604	MC	09	1	752	0.7212							0.97
2013_Engl	604	MC	10	1	752	0.5840							1.07
2013_Engl	604	MC	11	1	752	0.3862							1.09
2013_Engl	604	MC	12	1	752	0.1855							1.19
2013_Engl	604	MC	13	1	752	0.1667							1.31
2013_Engl	604	MC	14	1	752	1.1164							1.13
2013_Engl	604	MC	15	1	752	-1.4849							0.99

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2013_Engl	604	MC	16	1	752	1.0012							1.12
2013_Engl	604	MC	17	1	752	-0.4624							0.94
2013_Engl	604	MC	18	1	752	1.0727							1.10
2013_Engl	604	MC	19	1	752	-0.3189							1.11
2013_Engl	604	MC	20	1	752	-0.2227							1.03
2013_Engl	604	MC	21	1	752	-0.0270							0.91
2013_Engl	604	MC	22	1	752	-0.1466							0.92
2013_Engl	604	MC	23	1	752	-0.7205							0.88
2013_Engl	604	MC	24	1	752	0.1355							0.99
2013_Engl	604	MC	25	1	752	0.5582							1.01
2013_Engl	604	MC	26	1	752	-0.3902							0.78
2013_Engl	604	MC	27	1	752	-0.5490							0.79
2013_Engl	604	MC	28	1	752	0.4368							0.92
2013_Engl	604	MC	29	1	752	1.2590							1.10
2013_Engl	604	MC	30	1	752	0.0481							0.99
2013_Engl	605	MC	01	1	754	1.7558							1.03
2013_Engl	605	MC	02	1	754	-0.4550							0.92
2013_Engl	605	MC	03	1	754	0.9842							1.08
2013_Engl	605	MC	04	1	754	-1.6914							0.94
2013_Engl	605	MC	05	1	754	-0.3854							1.09
2013_Engl	605	MC	06	1	754	-0.9920							0.98
2013_Engl	605	MC	07	1	754	0.7749							1.04
2013_Engl	605	MC	08	1	754	-1.4763							0.94
2013_Engl	605	MC	09	1	754	0.4394							0.95
2013_Engl	605	MC	10	1	754	1.0057							1.09
2013_Engl	605	MC	11	1	754	0.3012							0.99
2013_Engl	605	MC	12	1	754	0.6796							1.21
2013_Engl	605	MC	13	1	754	-0.7589							0.92

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2013_Engl	605	MC	14	1	754	0.1896							1.30
2013_Engl	605	MC	15	1	754	-1.5057							0.94
2013_Engl	605	MC	16	1	754	1.0488							1.19
2013_Engl	605	MC	17	1	754	0.3077							1.17
2013_Engl	605	MC	18	1	754	2.0400							1.19
2013_Engl	605	MC	19	1	754	-0.4270							1.21
2013_Engl	605	MC	20	1	754	1.7103							1.26
2013_Engl	605	MC	21	1	754	0.0845							1.03
2013_Engl	605	MC	22	1	754	0.0053							0.96
2013_Engl	605	MC	23	1	754	0.0119							0.89
2013_Engl	605	MC	24	1	754	0.5322							0.97
2013_Engl	605	MC	25	1	754	-0.4972							0.83
2013_Engl	605	MC	26	1	754	-0.7290							0.79
2013_Engl	605	MC	27	1	754	0.5855							1.10
2013_Engl	605	MC	28	1	754	0.2749							0.89
2013_Engl	605	MC	29	1	754	-0.6260							0.86
2013_Engl	605	MC	30	1	754	1.3814							1.16
2013_Engl	606	MC	01	1	759	-0.5003							0.97
2013_Engl	606	MC	02	1	759	-0.8625							1.02
2013_Engl	606	MC	03	1	759	-0.2884							1.11
2013_Engl	606	MC	04	1	759	0.1897							1.16
2013_Engl	606	MC	05	1	759	1.1769							1.18
2013_Engl	606	MC	06	1	759	-0.9581							1.00
2013_Engl	606	MC	07	1	759	-0.2884							0.92
2013_Engl	606	MC	08	1	759	0.4174							1.06
2013_Engl	606	MC	09	1	759	0.2156							1.03
2013_Engl	606	MC	10	1	759	-0.9063							0.88
2013_Engl	606	MC	11	1	759	-0.1001							0.89

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2013_Engl	606	MC	12	1	759	0.8873							1.29
2013_Engl	606	MC	13	1	759	0.2479							1.26
2013_Engl	606	MC	14	1	759	0.5632							1.01
2013_Engl	606	MC	15	1	759	0.2933							0.93
2013_Engl	606	MC	16	1	759	1.1614							1.07
2013_Engl	606	MC	17	1	759	0.5900							1.23
2013_Engl	606	MC	18	1	759	0.4899							1.06
2013_Engl	606	MC	19	1	759	1.5140							0.97
2013_Engl	606	MC	20	1	759	0.0866							0.95
2013_Engl	606	MC	21	1	759	-0.0614							0.87
2013_Engl	606	MC	22	1	759	0.6918							0.99
2013_Engl	606	MC	23	1	759	0.7124							1.09
2013_Engl	606	MC	24	1	759	0.2803							1.04
2013_Engl	606	MC	25	1	759	-0.1648							0.90
2013_Engl	606	MC	26	1	759	0.5632							0.99
2013_Engl	606	MC	27	1	759	1.0328							1.12
2013_Engl	606	MC	28	1	759	0.1962							0.85
2013_Engl	606	MC	29	1	759	1.4705							1.28
2013_Engl	606	MC	30	1	759	1.0254							1.00
2013_Engl	611	CR	Ar	6	641	1.2117	-4.2704	-1.5605	-1.2757	0.4368	2.1445	4.5253	0.89
2013_Engl	612	CR	Ar	6	664	0.9940	-3.7174	-2.5674	-1.3663	0.7313	1.9827	4.9370	0.92
2013_Engl	613	CR	Ar										
2013_Engl	614	CR	Ar	6	668	0.9985	-3.9297	-2.3796	-1.8449	0.6548	2.4980	5.0015	0.99
2013_Engl	615	CR	Ar	6	653	0.8470	-3.9914	-1.8883	-1.2827	0.3710	2.2617	4.5298	1.02
2013_Engl	621	CR	Re	4	436	1.1311	-3.8320	-1.7867	1.7111	3.9076			1.05
2013_Engl	622	CR	Re	4	587	0.8448	-2.9266	-1.3273	1.0127	3.2411			1.00
2013_Engl	623	CR	Re	4	414	1.5023	-4.0808	-2.0258	1.6936	4.4130			1.00
2013_Engl	624	CR	Re	4	600	1.5097	-4.1004	-0.9654	1.5503	3.5155			0.93

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2013_Engl	625	CR	Re	4	611	1.2686	-3.9057	-1.0044	1.7069	3.2032			0.96
2013_Engl	626	CR	Re										
2013_Engl	627	CR	Re	4	533	0.6848	-3.3166	-1.9281	1.5988	3.6459			0.94
2013_Engl	N3	MC	01	1	10321	-0.7100							0.96
2013_Engl	N3	MC	02	1	10321	1.0400							1.12
2013_Engl	N3	MC	03	1	10321	-1.3300							0.89
2013_Engl	N3	MC	04	1	10321	-0.7500							0.95
2013_Engl	N3	MC	05	1	10321	-0.1900							0.99
2013_Engl	N3	MC	06	1	10321	0.7900							0.98
2013_Engl	N3	MC	07	1	10321	-0.2100							0.93
2013_Engl	N3	MC	08	1	10321	0.6900							1.08
2013_Engl	N3	MC	09	1	10321	-0.1200							0.91
2013_Engl	N3	MC	10	1	10321	0.2900							0.86

Appendix E: DIF Statistics

The first four columns from the left contain the test name, form ID, item type, and item sequence number within the form. The next three columns contain the Mantel Haenszel DIF statistical values (note that the MH Delta statistic cannot be calculated for CR items). The final two columns will only have values if the item displays possible moderate or severe DIF; if so, the degree of DIF (B/BB = moderate; C/CC = severe) and the favored group will be shown. Items without statistics are DNS (Do Not Score) status items.

Test	Form	Type	Item	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
2013_Engl	601	MC	01	1.54	10.79	0.23	C	Female
2013_Engl	601	MC	02	0.01	0.00	0.04		
2013_Engl	601	MC	03	-0.23	0.34	-0.04		
2013_Engl	601	MC	04	0.06	0.02	-0.01		
2013_Engl	601	MC	05	-0.54	1.82	-0.08		
2013_Engl	601	MC	06	0.34	0.78	0.09		
2013_Engl	601	MC	07	-0.93	5.47	-0.17		
2013_Engl	601	MC	08	-0.97	6.35	-0.18		
2013_Engl	601	MC	09	0.01	0.00	0.03		
2013_Engl	601	MC	10	-0.38	0.90	-0.05		
2013_Engl	601	MC	11	-0.15	0.13	-0.03		
2013_Engl	601	MC	12	-0.22	0.24	0.01		
2013_Engl	601	MC	13	-0.54	1.81	-0.09		
2013_Engl	601	MC	14	0.89	4.65	0.14		
2013_Engl	601	MC	15	-0.49	1.26	-0.07		
2013_Engl	601	MC	16	-0.52	1.98	-0.07		
2013_Engl	601	MC	17	-0.79	4.12	-0.14		
2013_Engl	601	MC	18	0.00	0.00	0.00		
2013_Engl	601	MC	19	0.13	0.11	0.05		
2013_Engl	601	MC	20	-0.75	3.65	-0.15		
2013_Engl	601	MC	21	-0.04	0.01	-0.05		
2013_Engl	601	MC	22	-0.14	0.10	-0.02		
2013_Engl	601	MC	23	-0.30	0.54	-0.05		
2013_Engl	601	MC	24	-0.08	0.04	-0.03		
2013_Engl	601	MC	25	0.00	0.00	-0.02		
2013_Engl	601	MC	26	0.11	0.08	0.02		
2013_Engl	601	MC	27	0.80	3.57	0.14		
2013_Engl	601	MC	28	0.10	0.05	0.04		
2013_Engl	601	MC	29	0.26	0.41	0.04		
2013_Engl	601	MC	30	-0.49	1.33	-0.06		

Test	Form	Type	Item	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
2013_Engl	602	MC	01	0.49	1.53	0.16		
2013_Engl	602	MC	02	0.40	1.15	0.09		
2013_Engl	602	MC	03	-1.44	12.51	-0.25	B	Male
2013_Engl	602	MC	04	0.06	0.02	0.06		
2013_Engl	602	MC	05	0.52	1.74	0.14		
2013_Engl	602	MC	06	-0.45	1.52	-0.11		
2013_Engl	602	MC	07	0.23	0.38	0.10		
2013_Engl	602	MC	08	0.67	3.16	0.15		
2013_Engl	602	MC	09	-0.05	0.02	0.06		
2013_Engl	602	MC	10	-0.33	0.51	-0.04		
2013_Engl	602	MC	11	0.19	0.19	0.04		
2013_Engl	602	MC	12	-0.08	0.03	0.03		
2013_Engl	602	MC	13	-0.56	1.72	-0.07		
2013_Engl	602	MC	14	-0.44	0.87	-0.04		
2013_Engl	602	MC	15	-0.47	1.11	-0.05		
2013_Engl	602	MC	16	0.10	0.05	0.04		
2013_Engl	602	MC	17	0.18	0.16	0.07		
2013_Engl	602	MC	18	-0.41	1.12	-0.07		
2013_Engl	602	MC	19	0.34	0.82	0.14		
2013_Engl	602	MC	20	0.10	0.04	0.04		
2013_Engl	602	MC	21	-0.11	0.07	-0.01		
2013_Engl	602	MC	22	-0.58	1.30	-0.07		
2013_Engl	602	MC	23	-0.13	0.10	0.01		
2013_Engl	602	MC	24	0.31	0.58	0.08		
2013_Engl	602	MC	25	-0.15	0.13	-0.01		
2013_Engl	602	MC	26	-0.03	0.00	0.03		
2013_Engl	602	MC	27	-0.60	2.20	-0.09		
2013_Engl	602	MC	28	-0.98	5.62	-0.15		
2013_Engl	602	MC	29	-1.15	7.87	-0.14	B	Male
2013_Engl	602	MC	30	-0.59	1.99	-0.08		
2013_Engl	603	MC	01	-1.07	6.29	-0.18	B	Male
2013_Engl	603	MC	02	-0.54	1.63	-0.10		
2013_Engl	603	MC	03	-0.48	1.41	-0.08		
2013_Engl	603	MC	04	0.09	0.05	-0.01		
2013_Engl	603	MC	05	-1.08	6.28	-0.17	B	Male
2013_Engl	603	MC	06	-0.46	0.85	-0.05		
2013_Engl	603	MC	07	-0.66	2.66	-0.05		
2013_Engl	603	MC	08	0.64	2.37	0.10		
2013_Engl	603	MC	09	1.17	8.85	0.24	B	Female

Test	Form	Type	Item	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
2013_Engl	603	MC	10	0.44	1.13	0.07		
2013_Engl	603	MC	11	0.62	2.70	0.18		
2013_Engl	603	MC	12	-1.48	10.43	-0.25	B	Male
2013_Engl	603	MC	13	-0.12	0.08	-0.03		
2013_Engl	603	MC	14	0.60	2.27	0.11		
2013_Engl	603	MC	15	0.62	1.83	0.08		
2013_Engl	603	MC	16	-0.22	0.26	-0.03		
2013_Engl	603	MC	17	0.47	1.49	0.13		
2013_Engl	603	MC	18	0.37	0.78	0.09		
2013_Engl	603	MC	19	-0.52	1.89	-0.11		
2013_Engl	603	MC	20	0.97	5.11	0.19		
2013_Engl	603	MC	21	-0.57	1.63	-0.05		
2013_Engl	603	MC	22	-0.71	2.62	-0.09		
2013_Engl	603	MC	23	-1.20	7.05	-0.17	B	Male
2013_Engl	603	MC	24	-0.26	0.45	-0.03		
2013_Engl	603	MC	25	0.54	1.20	0.08		
2013_Engl	603	MC	26	-0.21	0.23	-0.04		
2013_Engl	603	MC	27	-0.05	0.01	-0.01		
2013_Engl	603	MC	28	-0.36	0.76	-0.08		
2013_Engl	603	MC	29	-0.80	2.59	-0.06		
2013_Engl	603	MC	30	-0.08	0.05	-0.03		
2013_Engl	604	MC	01	-0.59	2.53	-0.08		
2013_Engl	604	MC	02	-0.37	0.91	-0.04		
2013_Engl	604	MC	03	0.46	1.49	0.14		
2013_Engl	604	MC	04	0.05	0.02	0.05		
2013_Engl	604	MC	05	-0.94	5.69	-0.15		
2013_Engl	604	MC	06	-1.26	10.40	-0.19	B	Male
2013_Engl	604	MC	07	-0.76	3.06	-0.10		
2013_Engl	604	MC	08	-0.75	4.08	-0.15		
2013_Engl	604	MC	09	-0.73	3.13	-0.11		
2013_Engl	604	MC	10	0.14	0.13	0.05		
2013_Engl	604	MC	11	-0.29	0.58	-0.07		
2013_Engl	604	MC	12	-0.68	3.31	-0.13		
2013_Engl	604	MC	13	0.71	3.92	0.15		
2013_Engl	604	MC	14	0.36	0.79	0.09		
2013_Engl	604	MC	15	0.18	0.13	0.04		
2013_Engl	604	MC	16	-0.09	0.05	0.01		
2013_Engl	604	MC	17	-0.52	1.53	-0.07		
2013_Engl	604	MC	18	-0.15	0.13	-0.01		

Test	Form	Type	Item	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
2013_Engl	604	MC	19	1.30	11.68	0.27	B	Female
2013_Engl	604	MC	20	-0.29	0.52	-0.05		
2013_Engl	604	MC	21	-0.48	1.31	-0.07		
2013_Engl	604	MC	22	0.78	3.70	0.19		
2013_Engl	604	MC	23	0.54	1.46	0.09		
2013_Engl	604	MC	24	-0.25	0.39	-0.04		
2013_Engl	604	MC	25	0.20	0.26	0.06		
2013_Engl	604	MC	26	0.41	0.75	0.08		
2013_Engl	604	MC	27	-0.50	1.06	-0.03		
2013_Engl	604	MC	28	0.19	0.20	0.03		
2013_Engl	604	MC	29	-0.31	0.53	0.00		
2013_Engl	604	MC	30	0.54	1.87	0.11		
2013_Engl	605	MC	01	0.00	0.00	0.01		
2013_Engl	605	MC	02	-0.20	0.23	-0.01		
2013_Engl	605	MC	03	0.23	0.33	0.05		
2013_Engl	605	MC	04	0.89	3.00	0.13		
2013_Engl	605	MC	05	0.01	0.00	0.01		
2013_Engl	605	MC	06	-0.39	0.79	-0.03		
2013_Engl	605	MC	07	0.34	0.70	0.08		
2013_Engl	605	MC	08	-0.14	0.08	-0.02		
2013_Engl	605	MC	09	-0.21	0.27	-0.04		
2013_Engl	605	MC	10	0.48	1.43	0.08		
2013_Engl	605	MC	11	0.96	5.87	0.17		
2013_Engl	605	MC	12	-0.17	0.21	-0.03		
2013_Engl	605	MC	13	-0.82	3.09	-0.14		
2013_Engl	605	MC	14	0.12	0.11	0.01		
2013_Engl	605	MC	15	0.56	1.24	0.05		
2013_Engl	605	MC	16	-0.48	1.46	-0.09		
2013_Engl	605	MC	17	-0.46	1.53	-0.12		
2013_Engl	605	MC	18	0.30	0.37	0.06		
2013_Engl	605	MC	19	-0.36	0.87	-0.08		
2013_Engl	605	MC	20	0.22	0.26	0.06		
2013_Engl	605	MC	21	-0.74	3.46	-0.13		
2013_Engl	605	MC	22	-1.05	6.44	-0.21	B	Male
2013_Engl	605	MC	23	-0.86	4.00	-0.15		
2013_Engl	605	MC	24	-0.53	1.66	-0.09		
2013_Engl	605	MC	25	-0.43	0.89	-0.05		
2013_Engl	605	MC	26	-0.50	1.00	-0.07		
2013_Engl	605	MC	27	0.72	3.35	0.15		

Test	Form	Type	Item	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
2013_Engl	605	MC	28	-0.94	4.96	-0.16		
2013_Engl	605	MC	29	-0.70	2.39	-0.07		
2013_Engl	605	MC	30	-1.15	7.91	-0.23	B	Male
2013_Engl	606	MC	01	0.24	0.37	0.08		
2013_Engl	606	MC	02	0.77	3.46	0.14		
2013_Engl	606	MC	03	-0.02	0.00	-0.01		
2013_Engl	606	MC	04	-0.38	1.00	-0.08		
2013_Engl	606	MC	05	0.43	1.10	0.05		
2013_Engl	606	MC	06	0.46	1.22	0.09		
2013_Engl	606	MC	07	0.15	0.14	0.05		
2013_Engl	606	MC	08	0.27	0.50	0.05		
2013_Engl	606	MC	09	0.04	0.01	0.02		
2013_Engl	606	MC	10	0.22	0.26	0.07		
2013_Engl	606	MC	11	0.49	1.42	0.07		
2013_Engl	606	MC	12	-0.19	0.25	-0.05		
2013_Engl	606	MC	13	0.90	5.77	0.18		
2013_Engl	606	MC	14	-0.46	1.28	-0.09		
2013_Engl	606	MC	15	0.03	0.01	0.03		
2013_Engl	606	MC	16	0.11	0.06	0.03		
2013_Engl	606	MC	17	-0.24	0.42	-0.04		
2013_Engl	606	MC	18	-0.67	2.97	-0.08		
2013_Engl	606	MC	19	-1.02	4.93	-0.16	B	Male
2013_Engl	606	MC	20	-0.44	1.12	-0.07		
2013_Engl	606	MC	21	-0.68	2.44	-0.11		
2013_Engl	606	MC	22	-0.42	1.08	-0.08		
2013_Engl	606	MC	23	-0.58	2.13	-0.09		
2013_Engl	606	MC	24	0.11	0.07	0.03		
2013_Engl	606	MC	25	-0.31	0.52	-0.06		
2013_Engl	606	MC	26	-1.34	9.78	-0.21	B	Male
2013_Engl	606	MC	27	0.97	5.53	0.21		
2013_Engl	606	MC	28	-1.88	17.75	-0.24	C	Male
2013_Engl	606	MC	29	0.79	3.31	0.14		
2013_Engl	606	MC	30	-1.18	7.70	-0.23	B	Male
2013_Engl	611	CR	Ar		17.55	0.22	BB	Female
2013_Engl	612	CR	Ar		8.11	0.15		
2013_Engl	613	CR	Ar					
2013_Engl	614	CR	Ar		4.50	0.16		
2013_Engl	615	CR	Ar		8.08	0.21		
2013_Engl	621	CR	Re		16.03	0.40	CC	Female

Test	Form	Type	Item	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
2013_Engl	622	CR	Re		8.27	0.18		
2013_Engl	623	CR	Re		2.07	0.11		
2013_Engl	624	CR	Re		1.99	0.08		
2013_Engl	625	CR	Re		2.87	0.11		
2013_Engl	626	CR	Re					
2013_Engl	627	CR	Re		3.34	0.12		
2013_Engl	N3	MC	01	0.24	4.97	0.04		
2013_Engl	N3	MC	02	-0.18	3.07	-0.03		
2013_Engl	N3	MC	03	0.87	48.44	0.13		
2013_Engl	N3	MC	04	1.16	12.64	0.21	B	Female
2013_Engl	N3	MC	05	0.17	2.87	0.04		
2013_Engl	N3	MC	06	-1.04	0.61	-0.18	B	Male
2013_Engl	N3	MC	07	0.23	4.90	0.05		
2013_Engl	N3	MC	08	0.65	44.16	0.13		
2013_Engl	N3	MC	09	1.03	97.41	0.19	B	Female
2013_Engl	N3	MC	10	0.18	2.97	0.05		

DIF category meanings: A/AA = negligible, B/BB = moderate, C/CC = severe.

Appendix F: Operational Test Maps

January 2013

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
1	MC	1	1	S2			0.86	0.31	-1.68	1.11
2	MC	1	1	S1			0.86	0.39	-1.74	1.00
3	MC	1	1	CPI			0.82	0.49	-1.32	0.95
4	MC	1	1	S2			0.82	0.53	-1.38	0.88
5	MC	1	1	S1			0.66	0.56	-0.26	0.94
6	MC	1	1	CPI			0.90	0.48	-2.17	0.83
7	MC	1	1	S3			0.51	0.47	0.65	1.06
8	MC	1	1	S3			0.79	0.44	-1.13	1.02
9	MC	1	1	S3			0.64	0.40	0.10	1.08
10	MC	1	1	S3			0.82	0.38	-1.04	1.00
11	MC	1	1	S3			0.68	0.48	-0.15	0.98
12	MC	1	1	S1			0.74	0.56	-0.53	0.87
13	MC	1	1	S2			0.68	0.52	-0.16	0.93
14	MC	1	1	S3			0.51	0.37	0.79	1.13
15	MC	1	1	S2			0.63	0.58	0.17	0.86
16	MC	1	1	S3			0.53	0.41	0.70	1.08
17	MC	1	1	S1			0.83	0.48	-1.20	0.89
18	MC	1	1	S3			0.77	0.56	-0.69	0.86
19	MC	1	1	S2			0.71	0.56	-0.33	0.87
20	MC	1	1	CPI			0.71	0.45	-0.29	1.01
21	MC	1	1	S3			0.83	0.49	-1.49	0.91
22	MC	1	1	S3			0.87	0.51	-1.84	0.86
23	MC	1	1	S1			0.80	0.49	-1.21	0.94
24	MC	1	1	CPI			0.80	0.49	-1.24	0.94
25	MC	1	1	S2			0.75	0.41	-0.83	1.09
26	CR	2	3	CPI,S1,S2,S3			1.46	0.58	-1.11	0.99
27	CR	2	3	CPI,S1,S2			1.31	0.61	-0.30	1.05
28	Essay	6	3	CPI,S1,S2,S3			3.18	0.74	0.99	0.92

June 2013

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
1	MC	1	1	CPI			0.83	0.39	-1.23	0.98
2	MC	1	1	S2			0.83	0.40	-1.23	0.97
3	MC	1	1	S3			0.59	0.32	0.24	1.15
4	MC	1	1	S3			0.88	0.31	-1.75	1.01
5	MC	1	1	S3			0.79	0.53	-0.96	0.86
6	MC	1	1	S1			0.79	0.29	-0.96	1.10
7	MC	1	1	CPI			0.68	0.43	-0.25	0.99
8	MC	1	1	S2			0.87	0.39	-1.62	0.96
9	MC	1	1	S3			0.63	0.56	0.07	0.92
10	MC	1	1	S1			0.82	0.65	-1.20	0.75
11	MC	1	1	S3			0.76	0.61	-0.74	0.84
12	MC	1	1	S3			0.81	0.65	-1.14	0.76
13	MC	1	1	S2			0.71	0.50	-0.42	1.01
14	MC	1	1	S2			0.85	0.62	-1.53	0.76
15	MC	1	1	S1			0.65	0.53	-0.04	0.97
16	MC	1	1	S1			0.72	0.65	-0.45	0.80
17	MC	1	1	S3			0.76	0.69	-0.77	0.73
18	MC	1	1	S1			0.63	0.57	0.06	0.92
19	MC	1	1	S2			0.68	0.59	-0.24	0.89
20	MC	1	1	S3			0.60	0.63	0.28	0.82
21	MC	1	1	CPI			0.61	0.41	-0.18	1.09
22	MC	1	1	S2			0.77	0.44	-1.14	1.03
23	MC	1	1	S3			0.64	0.38	-0.37	1.13
24	MC	1	1	S1			0.84	0.35	-1.78	1.07
25	MC	1	1	S2			0.82	0.32	-1.56	1.13
26	CR	2	3	CPI,S1,S2,S3			1.59	0.58	-1.34	1.01
27	CR	2	3	CPI,S1,S2			1.40	0.63	-0.59	0.99
28	Essay	6	3	CPI,S1,S2,S3			3.20	0.66	0.83	1.22

August 2013

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
1	MC	1	1	S1			0.81	0.33	-1.13	1.11
2	MC	1	1	S2			0.76	0.44	-0.81	1.02
3	MC	1	1	CPI			0.88	0.26	-1.76	1.13
4	MC	1	1	CPI			0.77	0.34	-0.88	1.13
5	MC	1	1	S2			0.77	0.41	-0.85	1.05
6	MC	1	1	S1			0.89	0.24	-1.91	1.12
7	MC	1	1	S3			0.78	0.49	-0.90	0.95
8	MC	1	1	S3			0.87	0.23	-1.75	1.13
9	MC	1	1	CPI			0.75	0.48	-0.68	0.73
10	MC	1	1	S2			0.60	0.50	0.22	0.58
11	MC	1	1	S3			0.71	0.47	-0.41	0.69
12	MC	1	1	S3			0.69	0.56	-0.28	0.67
13	MC	1	1	S3			0.70	0.49	-0.38	0.69
14	MC	1	1	S3			0.51	0.44	0.70	0.49
15	MC	1	1	S1			0.67	0.56	-0.17	0.65
16	MC	1	1	S1			0.63	0.52	0.02	0.62
17	MC	1	1	S2			0.54	0.57	0.53	0.53
18	MC	1	1	S3			0.63	0.46	0.02	0.62
19	MC	1	1	S1			0.62	0.60	0.12	0.60
20	MC	1	1	S3			0.61	0.62	0.13	0.60
21	MC	1	1	S1			0.90	0.40	-2.21	0.94
22	MC	1	1	S3			0.87	0.39	-1.90	0.99
23	MC	1	1	S3			0.73	0.53	-0.78	0.91
24	MC	1	1	S2			0.78	0.38	-1.12	1.06
25	MC	1	1	S2			0.67	0.28	-0.40	1.24
26	CR	2	3	CPI,S1,S2,S3			1.33	0.64	-0.39	0.92
27	CR	2	3	CPI,S1,S2			1.26	0.64	-0.15	0.95
28	Essay	6	3	CPI,S1,S2,S3			3.13	0.70	1.11	1.12

Appendix G: Scoring Tables

January 2013

Raw Score	Ability	Scale Score
0	-6.039	0.000
1	-4.792	1.313
2	-4.037	2.844
3	-3.570	4.196
4	-3.221	5.637
5	-2.937	7.205
6	-2.694	8.799
7	-2.480	10.420
8	-2.285	12.072
9	-2.106	13.753
10	-1.940	15.460
11	-1.783	17.189
12	-1.633	18.940
13	-1.490	20.709

Raw Score	Ability	Scale Score
14	-1.352	22.493
15	-1.219	24.292
16	-1.089	26.104
17	-0.963	27.925
18	-0.838	29.757
19	-0.717	31.596
20	-0.597	33.441
21	-0.479	35.291
22	-0.362	37.144
23	-0.248	39.000
24	-0.137	40.857
25	-0.028	42.715
26	0.078	44.571
27	0.180	46.426

Raw Score	Ability	Scale Score
28	0.277	48.278
29	0.369	50.126
30	0.456	51.971
31	0.537	53.813
32	0.614	55.652
33	0.686	57.489
34	0.754	59.324
35	0.819	61.162
36	0.881	62.998
37	0.942	64.836
38	1.002	66.678
39	1.061	68.523
40	1.120	70.373
41	1.180	72.227

Raw Score	Ability	Scale Score
42	1.242	74.086
43	1.308	75.952
44	1.377	77.825
45	1.452	79.707
46	1.535	81.602
47	1.628	83.507
48	1.737	85.429
49	1.866	87.371
50	2.024	89.338
51	2.227	91.338
52	2.499	93.380
53	2.896	95.471
54	3.587	97.625
55	4.792	100.000

June 2013

Raw Score	Ability	Scale Score
0	-5.719	0.000
1	-4.492	1.767
2	-3.766	3.528
3	-3.327	5.197
4	-3.004	6.836
5	-2.746	8.447
6	-2.528	10.036
7	-2.337	11.605
8	-2.166	13.158
9	-2.010	14.707
10	-1.866	16.254
11	-1.731	17.788
12	-1.603	19.313
13	-1.481	20.832

Raw Score	Ability	Scale Score
14	-1.363	22.347
15	-1.250	23.863
16	-1.140	25.389
17	-1.032	26.918
18	-0.927	28.453
19	-0.822	29.997
20	-0.719	31.553
21	-0.617	33.125
22	-0.515	34.715
23	-0.413	36.328
24	-0.312	37.966
25	-0.210	39.630
26	-0.109	41.321
27	-0.008	43.044

Raw Score	Ability	Scale Score
28	0.091	44.797
29	0.188	46.578
30	0.283	48.384
31	0.373	50.213
32	0.460	52.060
33	0.542	53.922
34	0.620	55.794
35	0.693	57.674
36	0.763	59.561
37	0.829	61.454
38	0.893	63.356
39	0.956	65.263
40	1.017	67.176
41	1.079	69.094

Raw Score	Ability	Scale Score
42	1.141	71.023
43	1.205	72.963
44	1.271	74.914
45	1.342	76.883
46	1.418	78.873
47	1.503	80.891
48	1.599	82.948
49	1.713	85.050
50	1.851	87.181
51	2.029	89.393
52	2.273	91.731
53	2.638	94.280
54	3.297	96.849
55	4.482	100.000

August 2013

Raw Score	Ability	Scale Score
0	-5.623	0.000
1	-4.386	1.961
2	-3.646	3.884
3	-3.194	5.791
4	-2.859	7.683
5	-2.588	9.559
6	-2.358	11.418
7	-2.156	13.261
8	-1.974	15.087
9	-1.808	16.897
10	-1.653	18.692
11	-1.509	20.472
12	-1.372	22.235
13	-1.241	23.986

Raw Score	Ability	Scale Score
14	-1.116	25.724
15	-0.995	27.452
16	-0.878	29.167
17	-0.764	30.876
18	-0.653	32.574
19	-0.544	34.266
20	-0.437	35.954
21	-0.332	37.638
22	-0.229	39.321
23	-0.128	41.005
24	-0.029	42.692
25	0.068	44.383
26	0.161	46.081
27	0.252	47.786

Raw Score	Ability	Scale Score
28	0.339	49.500
29	0.421	51.222
30	0.500	52.953
31	0.575	54.693
32	0.645	56.439
33	0.713	58.194
34	0.777	59.957
35	0.839	61.728
36	0.898	63.506
37	0.957	65.289
38	1.014	67.079
39	1.072	68.877
40	1.130	70.682
41	1.189	72.495

Raw Score	Ability	Scale Score
42	1.250	74.315
43	1.315	76.141
44	1.383	77.978
45	1.457	79.826
46	1.539	81.686
47	1.631	83.560
48	1.738	85.451
49	1.865	87.363
50	2.021	89.304
51	2.220	91.281
52	2.488	93.307
53	2.879	95.395
54	3.563	97.563
55	4.764	100.000