

New York State Regents Examination in Chemistry

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	2
<i>Item Difficulty</i>	3
<i>Item Discrimination</i>	3
<i>Test Reliability</i>	4
<i>Scoring Reliability</i>	4
<i>Inter-Rater Agreement</i>	5
<i>Constructed-Response Item Means and Standard Deviations</i>	5
<i>Intraclass Correlation</i>	6
<i>Weighted Kappa</i>	7
ITEM RESPONSE THEORY (IRT) AND THE CALIBRATION AND EQUATING OF THE FIELD TEST	
ITEMS	7
<i>Item Calibration</i>	9
<i>Item Fit Evaluation</i>	9
DIFFERENTIAL ITEM FUNCTIONING	11
<i>The Mantel Chi-Square and Standardized Mean Difference</i>	11
<i>Multiple-Choice Items</i>	13
<i>The Odds Ratio</i>	13
<i>The Delta Scale</i>	13
<i>DIF Classification for MC Items</i>	13
<i>DIF Classification for CR Items</i>	14
Section III: Equating Procedure	14
RANDOMLY EQUIVALENT GROUP EQUATING DESIGN	15
Section IV: Scaling of Operational Test Forms	17
References	20
Appendix A: Classical Item Analysis	22
Appendix B: Inter-Rater Consistency – Point Differences Between First and Second Reads	31
Appendix C: Additional Measures of Inter-Rater Reliability and Agreement	34

Appendix D: Partial-Credit Model Item Analysis	38
Appendix E: DIF Statistics	44
Appendix F: Operational Test Maps	50
Appendix G: Scoring Tables.....	56

List of Tables

Table 1. Need/Resource Capacity Category Definitions	1
Table 2. Classical Item Analysis Summary	4
Table 3. Test and Scoring Reliability	5
Table 4. Criteria to Evaluate Mean-Square Fit Statistics	10
Table 5. Partial Credit Model Item Analysis Summary	11
Table 6. DIF Classification for MC Items	14
Table 7. DIF Classification for CR Items	14
Table 8. Initial Mean Abilities and Equating Constants.....	17

List of Figures

Figure 1. $2 \times t$ Contingency Table at the k^{th} of K Levels.....	11
--	----

Section I: Introduction

PURPOSE

The purpose of this report is to document the psychometric properties of the New York State Regents Examination in Chemistry. 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.

Section II: Field Test Analysis

In May 2013, prospective items for the New York State Regents Examination in Chemistry 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 only took a small subset of the items being field tested. The NYSED handled all scanning of the MC responses and scoring of the CR responses along with the composition of the student data file in-house and with other external vendors. After all scoring and scanning activities had been completed and the student data file built, it was supplied to Pearson and contained student MC responses and CR scores. In addition, the NYSED also created and supplied a test map file that documented the items on each of the FT forms and a student data file layout that contained the position of every field within the student data file. Upon receipt of these files, Pearson staff checked the data, test map, and layout 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 then took place. 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 10,454 records; the final field test data file contained 10,401 records.

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:

$$x = t + e$$

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.

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.

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. Two items had difficulties that were greater than 0.90 and three 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
811	1152	24	12	11	1	0	22	2
812	1141	25	10	15	0	0	19	6
813	1163	25	7	18	0	0	19	6
814	1155	25	12	13	0	1	18	6
815	1163	25	15	10	0	1	19	5
816	1161	25	8	17	0	1	17	7
817	1158	25	12	12	1	0	22	3
818	1148	25	12	13	0	0	23	2
819	1160	19	4	15	0	0	12	7

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.

Test Reliability

Reliability is the consistency of the results obtained from a measurement with respect to time or between 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 0.763 to 0.827. 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 0.866 to 0.953, indicating a high degree of reliability.

Table 3. Test and Scoring Reliability

Form Number	Test Reliability	Scoring Reliability
811	0.786	0.901
812	0.805	0.953
813	0.827	0.866
814	0.818	0.939
815	0.808	0.923
816	0.825	0.948
817	0.778	0.889
818	0.811	0.870
819	0.763	0.930

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. Although most items had a maximum point value of one, one item had a maximum point value of two, and one item had a maximum point value of three. Only the three-point item had any ratings that differed by more than one point and this only occurred for 1.5% of the sample that received dual reads. Appendix C contains additional summary information regarding the first and second reads, including the percentage of the first and second scores that were exact or adjacent matches. These were 100% for the two-point item and 98.5% for the three-point item. Nonadjacent scores were not possible for the remaining one-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 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. Only two items had 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 all but 10 items, and good reproducibility for those 10. With the lowest kappa being equal to 0.47, there were no items displaying marginal reproducibility. The 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 they 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 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, are more susceptible to misclassification), but 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 +2.00 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
811	1148	24	1	21	2	0	24	0	0
812	1135	25	0	24	1	0	25	0	0
813	1153	25	0	24	1	0	25	0	0
814	1152	25	0	23	2	0	25	0	0
815	1163	25	0	25	0	0	25	0	0
816	1158	25	1	23	1	0	25	0	0
817	1153	25	1	23	1	0	25	0	0
818	1142	25	0	23	2	0	25	0	0
819	1156	19	1	18	0	0	19	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 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 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:

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

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

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

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 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	Non-significant 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 Chemistry received one of nine test forms (numbered 811–819). Form 819 was the anchor form for the equating and was an intact form that had been administered in the prior year. Because the form had been previously administered, its items had known parameters on the operational scale. The remaining test forms contained 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 nine forms, 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.

RANDOMLY EQUIVALENT GROUP EQUATING DESIGN

The equating analyses were based on the assumption that the groups taking the different forms had equivalent ability distributions and means. Given the random assignment of forms to examinees, this was a reasonable assumption. The initial step in the analyses was to calibrate all forms, both the anchor form and the remaining field test forms. All forms were calibrated using Winsteps, version 3.60 (Linacre, 2005).

The anchor form calibration began with all anchor item difficulty parameters fixed to their known values from the previous year. Because it is possible for item parameters to “drift” (shift their difficulties relative to one another), 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. Five items were identified as having drifted for the 2013 analyses.

After a stable anchor item set had been identified, the mean of the ability estimates of the students who took the anchor form was computed². This mean ability was then

² Because under Rasch IRT the ability of students with extreme scores (either zero or perfect scores) cannot be exactly computed (they are equal to $-\infty$ and $+\infty$, respectively), they were excluded from this and all other analyses for both the anchor and other field test forms.

used as the target ability for the forms with the field test items. Because the groups taking the different forms were randomly equivalent and thus had the same mean ability, adjustment of the parameters of the field test items on any form, with values that produced an ability distribution for students who had taken the form with a mean equal to the target ability from the anchor form, would result in the parameters for the field test items on that form being equated to the scale of the anchor form, which was also the operational scale.

The equated mean ability estimate for Form 819 was 0.27. This value became the target mean ability estimate for the field test forms.

At this point in the analyses, the calibration of the anchor form was complete. The next step was the initial calibration of the field test forms. This was a “free” calibration, meaning that the item parameters were not constrained in any way. This initial calibration produced a set of Rasch difficulty parameters for the items on each form. Also produced as a part of the Winsteps calibration was a set of person ability estimates for each form.

The next step was the computation of an equating constant for each form. Under Rasch IRT, if all of the difficulty parameters on a form have a constant added to them, the ability estimates for examinees will also be changed from their previous values by the amount represented by that constant. Therefore, to adjust the item difficulty parameters such that the mean of the ability distribution is set equal to the target mean ability from the anchor form, an equating constant was calculated for each field test form by subtracting the field test form mean ability from the target mean ability. This value was then added to the Rasch difficulty parameter of all items on the field test form. These adjusted values were then used as anchors for a second Winsteps calibration of the field test form. The mean of the person ability values from this second calibration was computed and compared to the target mean. If the anchored field test mean ability differed from the target mean ability by 0.005 or more, then an additional equating constant was computed using the difference between the mean ability from the field test form anchored run and the target mean ability, and another anchored run was completed. This process continued until all adjusted field test form mean abilities were within the 0.005 tolerance limit around the targeted mean ability. The final equating constant for any field test form was the sum of the constants from each anchored round for that form. At this point, with the adjusted mean abilities for the field test forms all equal (within the specified limits) to the target abilities, all of the adjusted field test item parameters and the anchor item parameters were on the common operational scale, and thus could be used in any subsequent operational administration. The initial mean abilities and final equating constants for the field test forms can be found in Table 8.

Table 8. Initial Mean Abilities and Equating Constants

Form Number	Mean Ability	Constant
811	0.08	0.18
812	0.24	0.03
813	0.31	-0.04
814	-0.08	0.33
815	-0.04	0.30
816	0.45	-0.17
817	0.01	0.25
818	-0.09	0.34

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 and June 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 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 -0.088 . 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 and June 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, the examinations 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 June 2004 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 June 2004 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, because 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 form. 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 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 “M4” 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 “M4” are the proportions of students who received scores of 0 through 4. “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	Mean	Point-Biserial
2013_Chem	811	MC	01	1	1152	0.79	0.00		0.86	0.05	0.01	0.09	0.86	0.34
2013_Chem	811	MC	02	1	1152	0.79	0.00		0.54	0.29	0.11	0.06	0.29	0.46
2013_Chem	811	MC	03	1	1152	0.79	0.01		0.76	0.11	0.10	0.03	0.76	0.39
2013_Chem	811	MC	04	1	1152	0.79	0.01		0.13	0.22	0.14	0.51	0.51	0.27
2013_Chem	811	MC	05	1	1152	0.79	0.01		0.09	0.03	0.20	0.68	0.68	0.43
2013_Chem	811	MC	06	1	1152	0.79	0.00		0.00	0.02	0.94	0.04	0.94	0.28
2013_Chem	811	MC	07	1	1152	0.79	0.00		0.16	0.18	0.60	0.06	0.60	0.32
2013_Chem	811	MC	08	1	1152	0.79	0.00		0.09	0.29	0.61	0.01	0.61	0.42
2013_Chem	811	MC	09	1	1152	0.79	0.01		0.05	0.15	0.57	0.21	0.57	0.40
2013_Chem	811	MC	10	1	1152	0.79	0.00		0.35	0.28	0.06	0.31	0.35	0.40
2013_Chem	811	MC	11	1	1152	0.79	0.00		0.66	0.10	0.12	0.12	0.66	0.38
2013_Chem	811	MC	12	1	1152	0.79	0.01		0.13	0.40	0.28	0.18	0.40	0.38
2013_Chem	811	MC	13	1	1152	0.79	0.01		0.72	0.13	0.06	0.08	0.72	0.42
2013_Chem	811	MC	14	1	1152	0.79	0.01		0.12	0.08	0.65	0.13	0.65	0.50
2013_Chem	811	MC	15	1	1152	0.79	0.03		0.44	0.14	0.15	0.24	0.44	0.38
2013_Chem	811	CR	41	1	1152	0.79	0.05	0.43	0.52				0.52	0.49
2013_Chem	811	CR	42	1	1152	0.79	0.06	0.78	0.16				0.16	0.33

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	Mean	Point-Biserial
2013_Chem	811	CR	43	1	1152	0.79	0.13	0.45	0.43				0.43	0.47
2013_Chem	811	CR	44	1	1152	0.79	0.10	0.53	0.37				0.37	0.44
2013_Chem	811	CR	45	1	1152	0.79	0.14	0.60	0.25				0.25	0.51
2013_Chem	811	CR	46	1	1152	0.79	0.22	0.37	0.41				0.41	0.42
2013_Chem	811	CR	47	1	1152	0.79	0.31	0.22	0.47				0.47	0.47
2013_Chem	811	CR	48	1	1152	0.79	0.25	0.35	0.40				0.40	0.53
2013_Chem	811	CR	49	1	1152	0.79	0.29	0.58	0.13				0.13	0.43
2013_Chem	812	MC	01	1	1141	0.80	0.00		0.15	0.04	0.16	0.65	0.65	0.48
2013_Chem	812	MC	02	1	1141	0.80	0.01		0.77	0.10	0.05	0.08	0.77	0.48
2013_Chem	812	MC	03	1	1141	0.80	0.01		0.04	0.10	0.21	0.64	0.64	0.46
2013_Chem	812	MC	04	1	1141	0.80	0.01		0.40	0.38	0.10	0.11	0.40	0.36
2013_Chem	812	MC	05	1	1141	0.80	0.01		0.27	0.07	0.27	0.39	0.39	0.29
2013_Chem	812	MC	06	1	1141	0.80	0.01		0.08	0.23	0.57	0.11	0.57	0.31
2013_Chem	812	MC	07	1	1141	0.80	0.01		0.16	0.11	0.45	0.27	0.45	0.36
2013_Chem	812	MC	08	1	1141	0.80	0.01		0.12	0.10	0.20	0.58	0.58	0.40
2013_Chem	812	MC	09	1	1141	0.80	0.00		0.71	0.16	0.06	0.07	0.71	0.36
2013_Chem	812	MC	10	1	1141	0.80	0.00		0.07	0.05	0.75	0.13	0.75	0.44
2013_Chem	812	MC	11	1	1141	0.80	0.01		0.25	0.27	0.31	0.16	0.31	0.28
2013_Chem	812	MC	12	1	1141	0.80	0.01		0.06	0.68	0.15	0.09	0.68	0.41
2013_Chem	812	MC	13	1	1141	0.80	0.01		0.18	0.09	0.59	0.11	0.59	0.37
2013_Chem	812	MC	14	1	1141	0.80	0.02		0.62	0.21	0.08	0.07	0.62	0.47
2013_Chem	812	MC	15	1	1141	0.80	0.04		0.13	0.20	0.16	0.48	0.48	0.40
2013_Chem	812	CR	41	1	1141	0.80	0.00	0.16	0.84				0.84	0.29
2013_Chem	812	CR	42	1	1141	0.80	0.06	0.38	0.56				0.56	0.50
2013_Chem	812	CR	43	1	1141	0.80	0.06	0.37	0.57				0.57	0.40
2013_Chem	812	CR	44	1	1141	0.80	0.12	0.43	0.45				0.45	0.52
2013_Chem	812	CR	45	1	1141	0.80	0.08	0.23	0.70				0.70	0.51

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	Mean	Point-Biserial
2013_Chem	812	CR	46	1	1141	0.80	0.14	0.33	0.53				0.53	0.53
2013_Chem	812	CR	47	1	1141	0.80	0.15	0.42	0.43				0.43	0.42
2013_Chem	812	CR	48	1	1141	0.80	0.14	0.36	0.49				0.49	0.52
2013_Chem	812	CR	49	1	1141	0.80	0.23	0.39	0.38				0.38	0.59
2013_Chem	812	CR	50	1	1141	0.80	0.16	0.69	0.15				0.15	0.31
2013_Chem	813	MC	01	1	1163	0.83	0.01		0.14	0.04	0.66	0.14	0.66	0.43
2013_Chem	813	MC	02	1	1163	0.83	0.00		0.06	0.54	0.05	0.35	0.54	0.35
2013_Chem	813	MC	03	1	1163	0.83	0.00		0.03	0.83	0.07	0.06	0.83	0.39
2013_Chem	813	MC	04	1	1163	0.83	0.00		0.09	0.10	0.75	0.06	0.75	0.51
2013_Chem	813	MC	05	1	1163	0.83	0.01		0.07	0.08	0.79	0.05	0.79	0.49
2013_Chem	813	MC	06	1	1163	0.83	0.00		0.69	0.06	0.14	0.10	0.69	0.41
2013_Chem	813	MC	07	1	1163	0.83	0.00		0.11	0.06	0.62	0.21	0.62	0.37
2013_Chem	813	MC	08	1	1163	0.83	0.00		0.80	0.10	0.05	0.05	0.80	0.39
2013_Chem	813	MC	09	1	1163	0.83	0.01		0.27	0.09	0.59	0.05	0.59	0.44
2013_Chem	813	MC	10	1	1163	0.83	0.01		0.50	0.25	0.18	0.06	0.50	0.53
2013_Chem	813	MC	11	1	1163	0.83	0.00		0.04	0.07	0.83	0.06	0.83	0.49
2013_Chem	813	MC	12	1	1163	0.83	0.01		0.09	0.65	0.10	0.15	0.65	0.42
2013_Chem	813	MC	13	1	1163	0.83	0.01		0.18	0.10	0.51	0.19	0.51	0.25
2013_Chem	813	MC	14	1	1163	0.83	0.04		0.09	0.27	0.16	0.44	0.44	0.42
2013_Chem	813	CR	41	1	1163	0.83	0.01	0.36	0.63				0.63	0.39
2013_Chem	813	CR	42	1	1163	0.83	0.09	0.48	0.43				0.43	0.35
2013_Chem	813	CR	43	1	1163	0.83	0.09	0.52	0.38				0.38	0.52
2013_Chem	813	CR	44	1	1163	0.83	0.09	0.35	0.56				0.56	0.51
2013_Chem	813	CR	45	1	1163	0.83	0.07	0.27	0.66				0.66	0.50
2013_Chem	813	CR	46	1	1163	0.83	0.12	0.38	0.51				0.51	0.56
2013_Chem	813	CR	47	1	1163	0.83	0.24	0.37	0.39				0.39	0.51
2013_Chem	813	CR	48	1	1163	0.83	0.18	0.51	0.31				0.31	0.47

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	Mean	Point-Biserial
2013_Chem	813	CR	49	1	1163	0.83	0.19	0.31	0.50				0.50	0.47
2013_Chem	813	CR	50	1	1163	0.83	0.22	0.57	0.21				0.21	0.46
2013_Chem	813	CR	51	1	1163	0.83	0.26	0.60	0.15				0.15	0.44
2013_Chem	814	MC	01	1	1155	0.82	0.00		0.14	0.10	0.57	0.19	0.57	0.39
2013_Chem	814	MC	02	1	1155	0.82	0.01		0.04	0.28	0.23	0.45	0.45	0.40
2013_Chem	814	MC	03	1	1155	0.82	0.01		0.25	0.07	0.63	0.05	0.63	0.44
2013_Chem	814	MC	04	1	1155	0.82	0.00		0.19	0.68	0.05	0.07	0.68	0.42
2013_Chem	814	MC	05	1	1155	0.82	0.00		0.81	0.07	0.08	0.04	0.81	0.42
2013_Chem	814	MC	06	1	1155	0.82	0.04		0.15	0.26	0.20	0.35	0.35	0.32
2013_Chem	814	MC	07	1	1155	0.82	0.01		0.21	0.13	0.56	0.10	0.56	0.39
2013_Chem	814	MC	08	1	1155	0.82	0.00		0.57	0.25	0.08	0.10	0.57	0.36
2013_Chem	814	MC	09	1	1155	0.82	0.01		0.48	0.32	0.11	0.08	0.48	0.25
2013_Chem	814	MC	10	1	1155	0.82	0.00		0.16	0.75	0.04	0.05	0.75	0.48
2013_Chem	814	MC	11	1	1155	0.82	0.01		0.04	0.29	0.47	0.19	0.47	0.28
2013_Chem	814	MC	12	1	1155	0.82	0.01		0.15	0.08	0.60	0.16	0.60	0.48
2013_Chem	814	MC	13	1	1155	0.82	0.02		0.14	0.71	0.06	0.07	0.71	0.45
2013_Chem	814	CR	41	1	1155	0.82	0.17	0.51	0.32				0.32	0.53
2013_Chem	814	CR	42	1	1155	0.82	0.18	0.42	0.40				0.40	0.46
2013_Chem	814	CR	43	1	1155	0.82	0.16	0.33	0.51				0.51	0.57
2013_Chem	814	CR	44	1	1155	0.82	0.09	0.27	0.64				0.64	0.42
2013_Chem	814	CR	45	1	1155	0.82	0.07	0.27	0.66				0.66	0.42
2013_Chem	814	CR	46	1	1155	0.82	0.12	0.79	0.09				0.09	0.33
2013_Chem	814	CR	47	1	1155	0.82	0.18	0.51	0.31				0.31	0.51
2013_Chem	814	CR	48	1	1155	0.82	0.14	0.53	0.33				0.33	0.44
2013_Chem	814	CR	49	1	1155	0.82	0.11	0.20	0.68				0.68	0.52
2013_Chem	814	CR	50	1	1155	0.82	0.27	0.47	0.26				0.26	0.50
2013_Chem	814	CR	51	1	1155	0.82	0.19	0.51	0.30				0.30	0.59

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	Mean	Point-Biserial
2013_Chem	814	CR	52	1	1155	0.82	0.28	0.57	0.16				0.16	0.48
2013_Chem	815	MC	01	1	1163	0.81	0.00		0.10	0.19	0.08	0.63	0.63	0.30
2013_Chem	815	MC	02	1	1163	0.81	0.01		0.15	0.74	0.06	0.04	0.74	0.43
2013_Chem	815	MC	03	1	1163	0.81	0.01		0.50	0.26	0.07	0.15	0.50	0.44
2013_Chem	815	MC	04	1	1163	0.81	0.01		0.19	0.07	0.25	0.48	0.19	0.15
2013_Chem	815	MC	05	1	1163	0.81	0.01		0.24	0.36	0.20	0.20	0.36	0.33
2013_Chem	815	MC	06	1	1163	0.81	0.01		0.10	0.71	0.09	0.10	0.71	0.39
2013_Chem	815	MC	07	1	1163	0.81	0.01		0.08	0.07	0.35	0.49	0.49	0.46
2013_Chem	815	MC	08	1	1163	0.81	0.02		0.08	0.10	0.69	0.11	0.69	0.43
2013_Chem	815	MC	09	1	1163	0.81	0.00		0.28	0.09	0.22	0.41	0.41	0.35
2013_Chem	815	MC	10	1	1163	0.81	0.00		0.02	0.68	0.15	0.15	0.68	0.45
2013_Chem	815	MC	11	1	1163	0.81	0.01		0.24	0.09	0.27	0.38	0.38	0.39
2013_Chem	815	MC	12	1	1163	0.81	0.01		0.10	0.18	0.10	0.61	0.61	0.51
2013_Chem	815	MC	13	1	1163	0.81	0.01		0.32	0.15	0.38	0.14	0.38	0.36
2013_Chem	815	MC	14	1	1163	0.81	0.01		0.06	0.53	0.09	0.32	0.53	0.43
2013_Chem	815	MC	15	1	1163	0.81	0.01		0.16	0.48	0.14	0.21	0.48	0.48
2013_Chem	815	CR	41	1	1163	0.81	0.04	0.24	0.72				0.72	0.34
2013_Chem	815	CR	42	1	1163	0.81	0.03	0.37	0.60				0.60	0.44
2013_Chem	815	CR	43	1	1163	0.81	0.08	0.48	0.45				0.45	0.40
2013_Chem	815	CR	44	1	1163	0.81	0.08	0.51	0.41				0.41	0.48
2013_Chem	815	CR	45	1	1163	0.81	0.08	0.47	0.45				0.45	0.50
2013_Chem	815	CR	46	1	1163	0.81	0.11	0.49	0.40				0.40	0.30
2013_Chem	815	CR	47	1	1163	0.81	0.23	0.31	0.46				0.46	0.62
2013_Chem	815	CR	48	1	1163	0.81	0.22	0.54	0.24				0.24	0.52
2013_Chem	815	CR	49	1	1163	0.81	0.21	0.38	0.41				0.41	0.53
2013_Chem	815	CR	50	1	1163	0.81	0.23	0.40	0.37				0.37	0.47

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	Mean	Point-Biserial
2013_Chem	816	MC	01	1	1161	0.83	0.00		0.88	0.03	0.05	0.04	0.88	0.34
2013_Chem	816	MC	02	1	1161	0.83	0.01		0.15	0.20	0.26	0.38	0.26	0.34
2013_Chem	816	MC	03	1	1161	0.83	0.00		0.04	0.82	0.06	0.07	0.82	0.33
2013_Chem	816	MC	04	1	1161	0.83	0.01		0.03	0.12	0.06	0.79	0.79	0.40
2013_Chem	816	MC	05	1	1161	0.83	0.00		0.11	0.28	0.54	0.07	0.54	0.45
2013_Chem	816	MC	06	1	1161	0.83	0.01		0.61	0.14	0.14	0.11	0.61	0.54
2013_Chem	816	MC	07	1	1161	0.83	0.00		0.82	0.06	0.08	0.04	0.82	0.33
2013_Chem	816	MC	08	1	1161	0.83	0.00		0.47	0.47	0.02	0.04	0.47	0.22
2013_Chem	816	MC	09	1	1161	0.83	0.00		0.12	0.08	0.76	0.03	0.76	0.39
2013_Chem	816	MC	10	1	1161	0.83	0.00		0.70	0.15	0.12	0.03	0.70	0.42
2013_Chem	816	MC	11	1	1161	0.83	0.00		0.61	0.30	0.06	0.03	0.61	0.53
2013_Chem	816	MC	12	1	1161	0.83	0.00		0.83	0.06	0.07	0.04	0.83	0.48
2013_Chem	816	MC	13	1	1161	0.83	0.00		0.17	0.17	0.58	0.07	0.58	0.43
2013_Chem	816	MC	14	1	1161	0.83	0.01		0.08	0.52	0.23	0.16	0.52	0.41
2013_Chem	816	CR	41	1	1161	0.83	0.02	0.24	0.75				0.75	0.35
2013_Chem	816	CR	42	1	1161	0.83	0.02	0.55	0.43				0.43	0.48
2013_Chem	816	CR	43	1	1161	0.83	0.01	0.32	0.66				0.66	0.54
2013_Chem	816	CR	44	1	1161	0.83	0.03	0.32	0.64				0.64	0.52
2013_Chem	816	CR	45	1	1161	0.83	0.12	0.45	0.43				0.43	0.45
2013_Chem	816	CR	46	1	1161	0.83	0.13	0.32	0.55				0.55	0.55
2013_Chem	816	CR	47	1	1161	0.83	0.17	0.40	0.43				0.43	0.59
2013_Chem	816	CR	48	1	1161	0.83	0.14	0.21	0.65				0.65	0.46
2013_Chem	816	CR	49	1	1161	0.83	0.19	0.59	0.22				0.22	0.49
2013_Chem	816	CR	50	1	1161	0.83	0.22	0.52	0.26				0.26	0.55
2013_Chem	816	CR	51	1	1161	0.83	0.21	0.63	0.15				0.15	0.36
2013_Chem	817	MC	01	1	1158	0.78	0.00		0.05	0.13	0.44	0.39	0.39	0.51
2013_Chem	817	MC	02	1	1158	0.78	0.00		0.04	0.07	0.47	0.42	0.47	0.35

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	Mean	Point-Biserial
2013_Chem	817	MC	03	1	1158	0.78	0.01		0.13	0.53	0.17	0.15	0.53	0.31
2013_Chem	817	MC	04	1	1158	0.78	0.02		0.11	0.20	0.51	0.17	0.51	0.34
2013_Chem	817	MC	05	1	1158	0.78	0.00		0.68	0.17	0.07	0.08	0.68	0.43
2013_Chem	817	MC	06	1	1158	0.78	0.01		0.14	0.38	0.20	0.28	0.38	0.32
2013_Chem	817	MC	07	1	1158	0.78	0.00		0.46	0.17	0.32	0.05	0.46	0.41
2013_Chem	817	MC	08	1	1158	0.78	0.00		0.90	0.06	0.03	0.01	0.90	0.34
2013_Chem	817	MC	09	1	1158	0.78	0.01		0.24	0.51	0.12	0.13	0.51	0.40
2013_Chem	817	MC	10	1	1158	0.78	0.00		0.01	0.07	0.28	0.64	0.64	0.47
2013_Chem	817	MC	11	1	1158	0.78	0.00		0.13	0.55	0.13	0.19	0.55	0.39
2013_Chem	817	MC	12	1	1158	0.78	0.01		0.14	0.17	0.61	0.08	0.61	0.45
2013_Chem	817	MC	13	1	1158	0.78	0.01		0.04	0.04	0.08	0.84	0.84	0.37
2013_Chem	817	MC	14	1	1158	0.78	0.01		0.28	0.38	0.18	0.15	0.38	0.27
2013_Chem	817	MC	15	1	1158	0.78	0.01		0.76	0.05	0.10	0.09	0.76	0.41
2013_Chem	817	MC	16	1	1158	0.78	0.03		0.22	0.34	0.23	0.18	0.34	0.31
2013_Chem	817	CR	41	1	1158	0.78	0.15	0.29	0.56				0.56	0.41
2013_Chem	817	CR	42	1	1158	0.78	0.13	0.53	0.34				0.34	0.43
2013_Chem	817	CR	43	1	1158	0.78	0.21	0.42	0.38				0.38	0.54
2013_Chem	817	CR	44	1	1158	0.78	0.28	0.55	0.17				0.17	0.48
2013_Chem	817	CR	45	1	1158	0.78	0.26	0.52	0.22				0.22	0.44
2013_Chem	817	CR	46	1	1158	0.78	0.08	0.28	0.63				0.63	0.35
2013_Chem	817	CR	47	1	1158	0.78	0.21	0.46	0.34				0.34	0.52
2013_Chem	817	CR	48	1	1158	0.78	0.20	0.21	0.59				0.59	0.40
2013_Chem	817	CR	49	1	1158	0.78	0.22	0.50	0.28				0.28	0.36
2013_Chem	818	MC	01	1	1148	0.81	0.00		0.09	0.12	0.05	0.75	0.75	0.32
2013_Chem	818	MC	02	1	1148	0.81	0.01		0.09	0.20	0.09	0.61	0.61	0.37
2013_Chem	818	MC	03	1	1148	0.81	0.01		0.07	0.75	0.13	0.04	0.75	0.47
2013_Chem	818	MC	04	1	1148	0.81	0.00		0.33	0.05	0.58	0.03	0.58	0.30

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	Mean	Point-Biserial
2013_Chem	818	MC	05	1	1148	0.81	0.00		0.54	0.07	0.04	0.34	0.54	0.47
2013_Chem	818	MC	06	1	1148	0.81	0.01		0.07	0.18	0.10	0.64	0.64	0.45
2013_Chem	818	MC	07	1	1148	0.81	0.02		0.44	0.20	0.20	0.14	0.44	0.39
2013_Chem	818	MC	08	1	1148	0.81	0.00		0.20	0.04	0.71	0.05	0.71	0.42
2013_Chem	818	MC	09	1	1148	0.81	0.01		0.23	0.23	0.45	0.08	0.45	0.47
2013_Chem	818	MC	10	1	1148	0.81	0.00		0.21	0.64	0.07	0.07	0.64	0.38
2013_Chem	818	MC	11	1	1148	0.81	0.02		0.30	0.14	0.47	0.07	0.47	0.46
2013_Chem	818	MC	12	1	1148	0.81	0.00		0.10	0.09	0.11	0.70	0.70	0.46
2013_Chem	818	MC	13	1	1148	0.81	0.03		0.13	0.56	0.22	0.06	0.56	0.39
2013_Chem	818	MC	14	1	1148	0.81	0.02		0.10	0.38	0.10	0.41	0.41	0.39
2013_Chem	818	CR	41	1	1148	0.81	0.09	0.55	0.36				0.36	0.38
2013_Chem	818	CR	42	1	1148	0.81	0.10	0.65	0.25				0.25	0.44
2013_Chem	818	CR	43	1	1148	0.81	0.18	0.30	0.52				0.52	0.47
2013_Chem	818	CR	44	1	1148	0.81	0.19	0.66	0.14				0.14	0.38
2013_Chem	818	CR	45	1	1148	0.81	0.05	0.42	0.52				0.52	0.50
2013_Chem	818	CR	46	1	1148	0.81	0.07	0.28	0.65				0.65	0.46
2013_Chem	818	CR	47	1	1148	0.81	0.15	0.48	0.37				0.37	0.53
2013_Chem	818	CR	48	1	1148	0.81	0.20	0.55	0.26				0.26	0.37
2013_Chem	818	CR	49	1	1148	0.81	0.30	0.55	0.14				0.14	0.45
2013_Chem	818	CR	50	1	1148	0.81	0.22	0.35	0.44				0.44	0.45
2013_Chem	818	CR	51	1	1148	0.81	0.32	0.40	0.28				0.28	0.49
2013_Chem	819	MC	01	1	1160	0.76	0.00		0.57	0.18	0.14	0.11	0.57	0.37
2013_Chem	819	MC	02	1	1160	0.76	0.00		0.03	0.15	0.07	0.74	0.74	0.52
2013_Chem	819	MC	03	1	1160	0.76	0.00		0.04	0.07	0.24	0.65	0.65	0.43
2013_Chem	819	MC	04	1	1160	0.76	0.01		0.76	0.14	0.03	0.06	0.76	0.40
2013_Chem	819	MC	05	1	1160	0.76	0.00		0.69	0.14	0.13	0.03	0.69	0.39
2013_Chem	819	MC	06	1	1160	0.76	0.01		0.08	0.79	0.05	0.07	0.79	0.36

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	Mean	Point-Biserial
2013_Chem	819	MC	07	1	1160	0.76	0.00		0.06	0.89	0.04	0.00	0.89	0.37
2013_Chem	819	MC	08	1	1160	0.76	0.01		0.17	0.56	0.26	0.00	0.56	0.34
2013_Chem	819	MC	09	1	1160	0.76	0.02		0.51	0.14	0.16	0.17	0.51	0.49
2013_Chem	819	MC	10	1	1160	0.76	0.00		0.09	0.04	0.82	0.05	0.82	0.39
2013_Chem	819	MC	11	1	1160	0.76	0.01		0.63	0.22	0.06	0.09	0.63	0.51
2013_Chem	819	MC	12	1	1160	0.76	0.01		0.16	0.06	0.75	0.02	0.75	0.49
2013_Chem	819	CR	41	2	1160	0.76	0.05	0.17	0.64	0.14			0.91	0.54
2013_Chem	819	CR	42	1	1160	0.76	0.08	0.25	0.67				0.67	0.53
2013_Chem	819	CR	43	1	1160	0.76	0.16	0.19	0.66				0.66	0.53
2013_Chem	819	CR	44	3	1160	0.76	0.11	0.34	0.16	0.20	0.19		1.13	0.58
2013_Chem	819	CR	45	1	1160	0.76	0.13	0.54	0.32				0.32	0.53
2013_Chem	819	CR	46	1	1160	0.76	0.05	0.34	0.61				0.61	0.29
2013_Chem	819	CR	47	1	1160	0.76	0.17	0.39	0.44				0.44	0.45

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 Points	Difference (First Read Minus Second Read)						
			-3	-2	-1	0	1	2	3
811	16	1			2%	97%	1%		
811	17	1			1%	98%	1%		
811	18	1			4%	92%	3%		
811	19	1			1%	92%	7%		
811	20	1			1%	98%	2%		
811	21	1			0%	99%	1%		
811	22	1			4%	90%	6%		
811	23	1			7%	91%	2%		
811	24	1			0%	98%	2%		
812	16	1			0%	100%	0%		
812	17	1			2%	97%	2%		
812	18	1			1%	99%	0%		
812	19	1			0%	100%	0%		
812	20	1			2%	95%	2%		
812	21	1			0%	98%	2%		
812	22	1			1%	98%	2%		
812	23	1			1%	98%	1%		
812	24	1			1%	98%	1%		
812	25	1			4%	94%	2%		
813	15	1			3%	96%	1%		
813	16	1			2%	97%	1%		
813	17	1			4%	93%	3%		
813	18	1			3%	94%	3%		
813	19	1			7%	88%	5%		
813	20	1			5%	94%	1%		
813	21	1			1%	95%	4%		
813	22	1			5%	91%	4%		
813	23	1			2%	97%	2%		
813	24	1			6%	90%	5%		

Form	Item	Score Points	Difference (First Read Minus Second Read)						
			-3	-2	-1	0	1	2	3
813	25	1			4%	90%	6%		
814	14	1			2%	98%	0%		
814	15	1			2%	93%	5%		
814	16	1			1%	99%	0%		
814	17	1			1%	99%	0%		
814	18	1			2%	97%	1%		
814	19	1			1%	97%	1%		
814	20	1			0%	99%	1%		
814	21	1			4%	89%	8%		
814	22	1			1%	99%	0%		
814	23	1			4%	96%	0%		
814	24	1			1%	98%	1%		
814	25	1			0%	99%	1%		
815	16	1			0%	100%	0%		
815	17	1			2%	96%	2%		
815	18	1			1%	99%	0%		
815	19	1			4%	90%	6%		
815	20	1			0%	99%	1%		
815	21	1			5%	87%	9%		
815	22	1			0%	98%	2%		
815	23	1			0%	99%	1%		
815	24	1			1%	97%	2%		
815	25	1			2%	97%	2%		
816	15	1			1%	99%	0%		
816	16	1			2%	98%	0%		
816	17	1			1%	97%	1%		
816	18	1			1%	99%	1%		
816	19	1			2%	95%	2%		
816	20	1			1%	99%	0%		
816	21	1			5%	91%	4%		
816	22	1			2%	97%	1%		
816	23	1			0%	98%	2%		
816	24	1			1%	99%	0%		
816	25	1			2%	98%	0%		
817	17	1			2%	97%	1%		
817	18	1			0%	99%	1%		
817	19	1			0%	99%	1%		
817	20	1			3%	93%	4%		

Form	Item	Score Points	Difference (First Read Minus Second Read)						
			-3	-2	-1	0	1	2	3
817	21	1			3%	96%	1%		
817	22	1			5%	94%	1%		
817	23	1			3%	97%	1%		
817	24	1			1%	98%	1%		
817	25	1			11%	76%	13%		
818	15	1			1%	97%	2%		
818	16	1			2%	95%	3%		
818	17	1			2%	95%	2%		
818	18	1			3%	92%	5%		
818	19	1			1%	99%	0%		
818	20	1			1%	98%	1%		
818	21	1			7%	92%	1%		
818	22	1			11%	79%	11%		
818	23	1			4%	91%	4%		
818	24	1			2%	98%	0%		
818	25	1			5%	92%	4%		
819	13	2		0%	2%	94%	3%	0%	
819	14	1			3%	96%	1%		
819	15	1			0%	98%	2%		
819	16	3	0%	1%	4%	91%	4%	0%	0%
819	17	1			2%	97%	2%		
819	18	1			12%	83%	5%		
819	19	1			2%	95%	2%		

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 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		Intra-class Corr	Wt Kappa
				Exact	Adj	Total	First Read	Second Read	First Read	Second Read		
811	16	1	194	97.4%	2.6%	100.0%	0.6	0.6	0.50	0.50	0.95	0.95
811	17	1	194	98.5%	1.5%	100.0%	0.2	0.2	0.38	0.39	0.95	0.95
811	18	1	182	92.3%	7.7%	100.0%	0.5	0.5	0.50	0.50	0.85	0.85
811	19	1	184	91.8%	8.2%	100.0%	0.4	0.3	0.49	0.47	0.82	0.82
811	20	1	174	97.7%	2.3%	100.0%	0.3	0.2	0.44	0.43	0.94	0.94
811	21	1	157	99.4%	0.6%	100.0%	0.5	0.5	0.50	0.50	0.99	0.99
811	22	1	140	90.0%	10.0%	100.0%	0.7	0.6	0.48	0.49	0.79	0.78
811	23	1	151	90.7%	9.3%	100.0%	0.5	0.6	0.50	0.49	0.81	0.81
811	24	1	148	98.0%	2.0%	100.0%	0.2	0.2	0.38	0.36	0.93	0.93
812	16	1	190	100.0%	0.0%	100.0%	0.8	0.8	0.40	0.40	1.00	1.00
812	17	1	181	96.7%	3.3%	100.0%	0.6	0.6	0.48	0.48	0.93	0.93
812	18	1	176	99.4%	0.6%	100.0%	0.6	0.6	0.49	0.49	0.99	0.99
812	19	1	166	100.0%	0.0%	100.0%	0.5	0.5	0.50	0.50	1.00	1.00
812	20	1	177	95.5%	4.5%	100.0%	0.7	0.7	0.45	0.45	0.89	0.89
812	21	1	164	98.2%	1.8%	100.0%	0.6	0.6	0.49	0.49	0.96	0.96
812	22	1	167	97.6%	2.4%	100.0%	0.5	0.5	0.50	0.50	0.95	0.95
812	23	1	167	97.6%	2.4%	100.0%	0.6	0.6	0.49	0.49	0.95	0.95
812	24	1	149	98.0%	2.0%	100.0%	0.5	0.5	0.50	0.50	0.96	0.96

Form	Item	Score Points	Total N-Count	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-class Corr	Wt Kappa
				Exact	Adj	Total	First Read	Second Read	First Read	Second Read		
812	25	1	164	93.9%	6.1%	100.0%	0.1	0.1	0.34	0.35	0.74	0.74
813	15	1	235	95.7%	4.3%	100.0%	0.7	0.7	0.48	0.47	0.91	0.91
813	16	1	221	96.8%	3.2%	100.0%	0.5	0.5	0.50	0.50	0.94	0.94
813	17	1	221	93.2%	6.8%	100.0%	0.4	0.4	0.49	0.50	0.86	0.86
813	18	1	220	93.6%	6.4%	100.0%	0.7	0.7	0.48	0.48	0.86	0.86
813	19	1	226	88.5%	11.5%	100.0%	0.7	0.7	0.47	0.46	0.73	0.73
813	20	1	217	94.5%	5.5%	100.0%	0.6	0.6	0.49	0.49	0.88	0.88
813	21	1	184	95.1%	4.9%	100.0%	0.5	0.5	0.50	0.50	0.90	0.90
813	22	1	196	91.3%	8.7%	100.0%	0.4	0.4	0.50	0.50	0.82	0.82
813	23	1	199	97.0%	3.0%	100.0%	0.6	0.6	0.49	0.49	0.94	0.94
813	24	1	193	89.6%	10.4%	100.0%	0.3	0.3	0.44	0.44	0.73	0.73
813	25	1	184	90.2%	9.8%	100.0%	0.2	0.2	0.40	0.38	0.68	0.68
814	14	1	131	97.7%	2.3%	100.0%	0.4	0.4	0.49	0.50	0.95	0.95
814	15	1	132	93.2%	6.8%	100.0%	0.5	0.5	0.50	0.50	0.86	0.86
814	16	1	136	99.3%	0.7%	100.0%	0.6	0.6	0.49	0.49	0.98	0.98
814	17	1	141	99.3%	0.7%	100.0%	0.7	0.7	0.45	0.44	0.98	0.98
814	18	1	152	97.4%	2.6%	100.0%	0.7	0.7	0.46	0.45	0.94	0.94
814	19	1	135	97.0%	3.0%	100.0%	0.1	0.1	0.32	0.32	0.85	0.85
814	20	1	134	98.5%	1.5%	100.0%	0.4	0.4	0.49	0.49	0.97	0.97
814	21	1	132	88.6%	11.4%	100.0%	0.4	0.4	0.50	0.49	0.77	0.76
814	22	1	140	99.3%	0.7%	100.0%	0.7	0.7	0.44	0.44	0.98	0.98
814	23	1	114	95.6%	4.4%	100.0%	0.3	0.4	0.47	0.49	0.91	0.90
814	24	1	129	98.4%	1.6%	100.0%	0.4	0.4	0.49	0.49	0.97	0.97
814	25	1	114	99.1%	0.9%	100.0%	0.2	0.2	0.37	0.37	0.97	0.97

Form	Item	Score Points	Total N-Count	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-class Corr	Wt Kappa
				Exact	Adj	Total	First Read	Second Read	First Read	Second Read		
815	16	1	159	100.0%	0.0%	100.0%	0.8	0.8	0.43	0.43	1.00	1.00
815	17	1	163	96.3%	3.7%	100.0%	0.6	0.6	0.49	0.49	0.92	0.92
815	18	1	150	98.7%	1.3%	100.0%	0.5	0.5	0.50	0.50	0.97	0.97
815	19	1	156	89.7%	10.3%	100.0%	0.5	0.4	0.50	0.50	0.79	0.79
815	20	1	157	99.4%	0.6%	100.0%	0.5	0.5	0.50	0.50	0.99	0.99
815	21	1	149	86.6%	13.4%	100.0%	0.5	0.4	0.50	0.50	0.73	0.73
815	22	1	128	98.4%	1.6%	100.0%	0.6	0.6	0.50	0.50	0.97	0.97
815	23	1	132	99.2%	0.8%	100.0%	0.3	0.3	0.45	0.45	0.98	0.98
815	24	1	128	96.9%	3.1%	100.0%	0.5	0.5	0.50	0.50	0.94	0.94
815	25	1	129	96.9%	3.1%	100.0%	0.5	0.5	0.50	0.50	0.94	0.94
816	15	1	144	98.6%	1.4%	100.0%	0.7	0.8	0.44	0.43	0.96	0.96
816	16	1	144	97.9%	2.1%	100.0%	0.5	0.5	0.50	0.50	0.96	0.96
816	17	1	144	97.2%	2.8%	100.0%	0.7	0.7	0.44	0.44	0.93	0.93
816	18	1	144	98.6%	1.4%	100.0%	0.7	0.7	0.48	0.48	0.97	0.97
816	19	1	130	95.4%	4.6%	100.0%	0.5	0.5	0.50	0.50	0.91	0.91
816	20	1	132	99.2%	0.8%	100.0%	0.6	0.6	0.49	0.48	0.98	0.98
816	21	1	127	91.3%	8.7%	100.0%	0.4	0.5	0.50	0.50	0.83	0.83
816	22	1	131	96.9%	3.1%	100.0%	0.8	0.8	0.41	0.40	0.91	0.91
816	23	1	124	98.4%	1.6%	100.0%	0.3	0.3	0.45	0.44	0.96	0.96
816	24	1	121	99.2%	0.8%	100.0%	0.4	0.4	0.49	0.49	0.98	0.98
816	25	1	123	98.4%	1.6%	100.0%	0.2	0.2	0.38	0.40	0.95	0.95
817	17	1	155	97.4%	2.6%	100.0%	0.6	0.6	0.48	0.48	0.94	0.94
817	18	1	162	99.4%	0.6%	100.0%	0.3	0.3	0.47	0.47	0.99	0.99
817	19	1	147	99.3%	0.7%	100.0%	0.5	0.5	0.50	0.50	0.99	0.99
817	20	1	130	93.1%	6.9%	100.0%	0.2	0.2	0.40	0.39	0.78	0.77

Form	Item	Score Points	Total N-Count	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-class Corr	Wt Kappa
				Exact	Adj	Total	First Read	Second Read	First Read	Second Read		
817	21	1	141	95.7%	4.3%	100.0%	0.3	0.3	0.45	0.46	0.90	0.90
817	22	1	167	94.0%	6.0%	100.0%	0.7	0.7	0.47	0.45	0.86	0.86
817	23	1	144	96.5%	3.5%	100.0%	0.4	0.4	0.48	0.49	0.93	0.93
817	24	1	151	98.0%	2.0%	100.0%	0.8	0.8	0.42	0.42	0.94	0.94
817	25	1	144	75.7%	24.3%	100.0%	0.4	0.3	0.48	0.48	0.47	0.47
818	15	1	146	96.6%	3.4%	100.0%	0.4	0.4	0.49	0.49	0.93	0.93
818	16	1	144	95.1%	4.9%	100.0%	0.3	0.3	0.45	0.44	0.88	0.88
818	17	1	131	95.4%	4.6%	100.0%	0.6	0.6	0.49	0.49	0.90	0.90
818	18	1	130	91.5%	8.5%	100.0%	0.2	0.2	0.41	0.39	0.73	0.73
818	19	1	150	99.3%	0.7%	100.0%	0.5	0.5	0.50	0.50	0.99	0.99
818	20	1	148	98.0%	2.0%	100.0%	0.7	0.7	0.45	0.46	0.95	0.95
818	21	1	141	92.2%	7.8%	100.0%	0.3	0.4	0.48	0.49	0.84	0.84
818	22	1	131	78.6%	21.4%	100.0%	0.3	0.3	0.48	0.48	0.53	0.53
818	23	1	113	91.2%	8.8%	100.0%	0.2	0.2	0.40	0.40	0.73	0.73
818	24	1	123	98.4%	1.6%	100.0%	0.6	0.6	0.50	0.50	0.97	0.97
818	25	1	109	91.7%	8.3%	100.0%	0.3	0.4	0.48	0.48	0.82	0.82
819	13	2	143	94.4%	5.6%	100.0%	0.9	0.9	0.63	0.60	0.93	0.91
819	14	1	140	96.4%	3.6%	100.0%	0.7	0.7	0.47	0.47	0.92	0.92
819	15	1	129	98.4%	1.6%	100.0%	0.8	0.8	0.42	0.43	0.96	0.96
819	16	3	134	91.0%	7.5%	98.5%	1.3	1.3	1.22	1.22	0.95	0.92
819	17	1	131	96.9%	3.1%	100.0%	0.3	0.3	0.46	0.46	0.93	0.93
819	18	1	142	83.1%	16.9%	100.0%	0.6	0.7	0.49	0.47	0.63	0.63
819	19	1	127	95.3%	4.7%	100.0%	0.5	0.5	0.50	0.50	0.91	0.91

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 five columns contain the Rasch Item Difficulty, step difficulties (for multi-point items only), and the INFIT Rasch model fit statistic.

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	INFIT
2013_Chem	811	MC	01	1	1152	-1.8142				0.98
2013_Chem	811	MC	02	1	1152	1.3195				0.95
2013_Chem	811	MC	03	1	1152	-1.0779				0.98
2013_Chem	811	MC	04	1	1152	0.2078				1.19
2013_Chem	811	MC	05	1	1152	-0.6278				0.98
2013_Chem	811	MC	06	1	1152	-2.8705				0.97
2013_Chem	811	MC	07	1	1152	-0.2475				1.12
2013_Chem	811	MC	08	1	1152	-0.2689				1.01
2013_Chem	811	MC	09	1	1152	-0.0995				1.04
2013_Chem	811	MC	10	1	1152	0.9876				1.03
2013_Chem	811	MC	11	1	1152	-0.5366				1.04
2013_Chem	811	MC	12	1	1152	0.7359				1.07
2013_Chem	811	MC	13	1	1152	-0.8567				0.97
2013_Chem	811	MC	14	1	1152	-0.5006				0.91
2013_Chem	811	MC	15	1	1152	0.5318				1.06
2013_Chem	811	CR	41	1	1152	0.1665				0.94
2013_Chem	811	CR	42	1	1152	2.1530				1.07
2013_Chem	811	CR	43	1	1152	0.5950				0.97
2013_Chem	811	CR	44	1	1152	0.8579				1.00
2013_Chem	811	CR	45	1	1152	1.5208				0.88
2013_Chem	811	CR	46	1	1152	0.6630				1.02
2013_Chem	811	CR	47	1	1152	0.4023				0.97
2013_Chem	811	CR	48	1	1152	0.7273				0.90
2013_Chem	811	CR	49	1	1152	2.4333				0.90
2013_Chem	812	MC	01	1	1141	-0.4588				0.94
2013_Chem	812	MC	02	1	1141	-1.1235				0.89
2013_Chem	812	MC	03	1	1141	-0.4053				0.97
2013_Chem	812	MC	04	1	1141	0.7542				1.08
2013_Chem	812	MC	05	1	1141	0.8067				1.17
2013_Chem	812	MC	06	1	1141	-0.0640				1.13
2013_Chem	812	MC	07	1	1141	0.4846				1.08
2013_Chem	812	MC	08	1	1141	-0.1234				1.03

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	INFIT
2013_Chem	812	MC	09	1	1141	-0.7764				1.04
2013_Chem	812	MC	10	1	1141	-1.0151				0.94
2013_Chem	812	MC	11	1	1141	1.2015				1.12
2013_Chem	812	MC	12	1	1141	-0.6330				1.00
2013_Chem	812	MC	13	1	1141	-0.1917				1.07
2013_Chem	812	MC	14	1	1141	-0.3303				0.96
2013_Chem	812	MC	15	1	1141	0.3545				1.04
2013_Chem	812	CR	41	1	1141	-1.6332				1.03
2013_Chem	812	CR	42	1	1141	-0.0471				0.93
2013_Chem	812	CR	43	1	1141	-0.0514				1.04
2013_Chem	812	CR	44	1	1141	0.5057				0.91
2013_Chem	812	CR	45	1	1141	-0.7135				0.89
2013_Chem	812	CR	46	1	1141	0.1291				0.90
2013_Chem	812	CR	47	1	1141	0.6119				1.02
2013_Chem	812	CR	48	1	1141	0.2960				0.92
2013_Chem	812	CR	49	1	1141	0.8464				0.83
2013_Chem	812	CR	50	1	1141	2.3283				1.04
2013_Chem	813	MC	01	1	1163	-0.5749				1.03
2013_Chem	813	MC	02	1	1163	0.0539				1.14
2013_Chem	813	MC	03	1	1163	-1.6588				0.96
2013_Chem	813	MC	04	1	1163	-1.0834				0.88
2013_Chem	813	MC	05	1	1163	-1.2986				0.90
2013_Chem	813	MC	06	1	1163	-0.7167				1.03
2013_Chem	813	MC	07	1	1163	-0.3217				1.10
2013_Chem	813	MC	08	1	1163	-1.3830				1.01
2013_Chem	813	MC	09	1	1163	-0.1688				1.02
2013_Chem	813	MC	10	1	1163	0.2572				0.92
2013_Chem	813	MC	11	1	1163	-1.6101				0.87
2013_Chem	813	MC	12	1	1163	-0.5058				1.05
2013_Chem	813	MC	13	1	1163	0.1937				1.26
2013_Chem	813	MC	14	1	1163	0.5597				1.05
2013_Chem	813	CR	41	1	1163	-0.4152				1.07
2013_Chem	813	CR	42	1	1163	0.5985				1.15
2013_Chem	813	CR	43	1	1163	0.8583				0.93
2013_Chem	813	CR	44	1	1163	-0.0270				0.94
2013_Chem	813	CR	45	1	1163	-0.5749				0.93
2013_Chem	813	CR	46	1	1163	0.2233				0.89
2013_Chem	813	CR	47	1	1163	0.7912				0.96
2013_Chem	813	CR	48	1	1163	1.2681				0.96

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	INFIT
2013_Chem	813	CR	49	1	1163	0.2360				1.00
2013_Chem	813	CR	50	1	1163	1.8952				0.92
2013_Chem	813	CR	51	1	1163	2.4198				0.92
2013_Chem	814	MC	01	1	1155	-0.1097				1.06
2013_Chem	814	MC	02	1	1155	0.4778				1.07
2013_Chem	814	MC	03	1	1155	-0.3878				1.00
2013_Chem	814	MC	04	1	1155	-0.6712				1.01
2013_Chem	814	MC	05	1	1155	-1.4319				0.95
2013_Chem	814	MC	06	1	1155	0.9924				1.15
2013_Chem	814	MC	07	1	1155	-0.0245				1.07
2013_Chem	814	MC	08	1	1155	-0.0841				1.10
2013_Chem	814	MC	09	1	1155	0.3296				1.25
2013_Chem	814	MC	10	1	1155	-1.0408				0.91
2013_Chem	814	MC	11	1	1155	0.3930				1.20
2013_Chem	814	MC	12	1	1155	-0.2648				0.96
2013_Chem	814	MC	13	1	1155	-0.8159				0.96
2013_Chem	814	CR	41	1	1155	1.1714				0.90
2013_Chem	814	CR	42	1	1155	0.7409				1.00
2013_Chem	814	CR	43	1	1155	0.1865				0.87
2013_Chem	814	CR	44	1	1155	-0.4279				1.02
2013_Chem	814	CR	45	1	1155	-0.5548				0.99
2013_Chem	814	CR	46	1	1155	3.0412				0.96
2013_Chem	814	CR	47	1	1155	1.2298				0.90
2013_Chem	814	CR	48	1	1155	1.1188				1.02
2013_Chem	814	CR	49	1	1155	-0.6807				0.88
2013_Chem	814	CR	50	1	1155	1.5066				0.90
2013_Chem	814	CR	51	1	1155	1.3042				0.81
2013_Chem	814	CR	52	1	1155	2.2691				0.89
2013_Chem	815	MC	01	1	1163	-0.3561				1.13
2013_Chem	815	MC	02	1	1163	-0.9964				0.96
2013_Chem	815	MC	03	1	1163	0.2403				1.00
2013_Chem	815	MC	04	1	1163	1.9638				1.21
2013_Chem	815	MC	05	1	1163	0.9442				1.10
2013_Chem	815	MC	06	1	1163	-0.7933				1.00
2013_Chem	815	MC	07	1	1163	0.2769				0.97
2013_Chem	815	MC	08	1	1163	-0.7129				0.97
2013_Chem	815	MC	09	1	1163	0.6982				1.09
2013_Chem	815	MC	10	1	1163	-0.6299				0.95
2013_Chem	815	MC	11	1	1163	0.8047				1.04

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	INFIT
2013_Chem	815	MC	12	1	1163	-0.2794				0.91
2013_Chem	815	MC	13	1	1163	0.8479				1.07
2013_Chem	815	MC	14	1	1163	0.1180				1.01
2013_Chem	815	MC	15	1	1163	0.3626				0.96
2013_Chem	815	CR	41	1	1163	-0.8562				1.05
2013_Chem	815	CR	42	1	1163	-0.2161				0.98
2013_Chem	815	CR	43	1	1163	0.4981				1.04
2013_Chem	815	CR	44	1	1163	0.6561				0.95
2013_Chem	815	CR	45	1	1163	0.4816				0.93
2013_Chem	815	CR	46	1	1163	0.7236				1.15
2013_Chem	815	CR	47	1	1163	0.4446				0.80
2013_Chem	815	CR	48	1	1163	1.6097				0.86
2013_Chem	815	CR	49	1	1163	0.6729				0.90
2013_Chem	815	CR	50	1	1163	0.8740				0.96
2013_Chem	816	MC	01	1	1161	-2.1613				0.96
2013_Chem	816	MC	02	1	1161	1.5585				1.09
2013_Chem	816	MC	03	1	1161	-1.6059				1.04
2013_Chem	816	MC	04	1	1161	-1.3639				1.00
2013_Chem	816	MC	05	1	1161	0.0419				1.03
2013_Chem	816	MC	06	1	1161	-0.3016				0.91
2013_Chem	816	MC	07	1	1161	-1.5924				1.04
2013_Chem	816	MC	08	1	1161	0.4025				1.31
2013_Chem	816	MC	09	1	1161	-1.1562				1.02
2013_Chem	816	MC	10	1	1161	-0.7878				1.01
2013_Chem	816	MC	11	1	1161	-0.3327				0.92
2013_Chem	816	MC	12	1	1161	-1.6676				0.87
2013_Chem	816	MC	13	1	1161	-0.1569				1.05
2013_Chem	816	MC	14	1	1161	0.1148				1.08
2013_Chem	816	CR	41	1	1161	-1.0798				1.10
2013_Chem	816	CR	42	1	1161	0.5809				1.00
2013_Chem	816	CR	43	1	1161	-0.6059				0.90
2013_Chem	816	CR	44	1	1161	-0.4858				0.93
2013_Chem	816	CR	45	1	1161	0.6029				1.05
2013_Chem	816	CR	46	1	1161	-0.0270				0.91
2013_Chem	816	CR	47	1	1161	0.5721				0.86
2013_Chem	816	CR	48	1	1161	-0.5132				0.99
2013_Chem	816	CR	49	1	1161	1.8044				0.92
2013_Chem	816	CR	50	1	1161	1.5196				0.86
2013_Chem	816	CR	51	1	1161	2.3553				1.05

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	INFIT
2013_Chem	817	MC	01	1	1158	0.7955				0.90
2013_Chem	817	MC	02	1	1158	0.3920				1.07
2013_Chem	817	MC	03	1	1158	0.1223				1.11
2013_Chem	817	MC	04	1	1158	0.2309				1.08
2013_Chem	817	MC	05	1	1158	-0.6143				0.97
2013_Chem	817	MC	06	1	1158	0.8466				1.09
2013_Chem	817	MC	07	1	1158	0.4325				1.01
2013_Chem	817	MC	08	1	1158	-2.2356				0.94
2013_Chem	817	MC	09	1	1158	0.2309				1.02
2013_Chem	817	MC	10	1	1158	-0.3939				0.93
2013_Chem	817	MC	11	1	1158	0.0091				1.03
2013_Chem	817	MC	12	1	1158	-0.2373				0.94
2013_Chem	817	MC	13	1	1158	-1.6612				0.94
2013_Chem	817	MC	14	1	1158	0.8125				1.14
2013_Chem	817	MC	15	1	1158	-1.0365				0.95
2013_Chem	817	MC	16	1	1158	1.0344				1.09
2013_Chem	817	CR	41	1	1158	-0.0152				1.00
2013_Chem	817	CR	42	1	1158	1.0077				0.99
2013_Chem	817	CR	43	1	1158	0.8509				0.88
2013_Chem	817	CR	44	1	1158	2.0997				0.87
2013_Chem	817	CR	45	1	1158	1.7115				0.96
2013_Chem	817	CR	46	1	1158	-0.3596				1.04
2013_Chem	817	CR	47	1	1158	1.0568				0.89
2013_Chem	817	CR	48	1	1158	-0.1791				1.01
2013_Chem	817	CR	49	1	1158	1.3525				1.03
2013_Chem	818	MC	01	1	1148	-1.0456				1.07
2013_Chem	818	MC	02	1	1148	-0.2734				1.07
2013_Chem	818	MC	03	1	1148	-1.0563				0.91
2013_Chem	818	MC	04	1	1148	-0.1528				1.15
2013_Chem	818	MC	05	1	1148	0.0460				0.97
2013_Chem	818	MC	06	1	1148	-0.4451				0.97
2013_Chem	818	MC	07	1	1148	0.5201				1.06
2013_Chem	818	MC	08	1	1148	-0.8297				0.98
2013_Chem	818	MC	09	1	1148	0.4862				0.97
2013_Chem	818	MC	10	1	1148	-0.4586				1.05
2013_Chem	818	MC	11	1	1148	0.3809				0.98
2013_Chem	818	MC	12	1	1148	-0.7660				0.95
2013_Chem	818	MC	13	1	1148	-0.0298				1.06
2013_Chem	818	MC	14	1	1148	0.7043				1.06

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	INFIT
2013_Chem	818	CR	41	1	1148	0.9344				1.07
2013_Chem	818	CR	42	1	1148	1.5554				0.98
2013_Chem	818	CR	43	1	1148	0.1424				0.97
2013_Chem	818	CR	44	1	1148	2.3867				0.98
2013_Chem	818	CR	45	1	1148	0.1424				0.93
2013_Chem	818	CR	46	1	1148	-0.4856				0.95
2013_Chem	818	CR	47	1	1148	0.8714				0.90
2013_Chem	818	CR	48	1	1148	1.5282				1.06
2013_Chem	818	CR	49	1	1148	2.3785				0.91
2013_Chem	818	CR	50	1	1148	0.5540				1.01
2013_Chem	818	CR	51	1	1148	1.3651				0.92
2013_Chem	819	MC	01	1	1160	0.0000				1.08
2013_Chem	819	MC	02	1	1160	-0.9300				0.87
2013_Chem	819	MC	03	1	1160	-0.4900				1.01
2013_Chem	819	MC	04	1	1160	-0.9300				0.96
2013_Chem	819	MC	05	1	1160	-0.7000				1.06
2013_Chem	819	MC	06	1	1160	-1.2900				1.01
2013_Chem	819	MC	07	1	1160	-2.0100				0.86
2013_Chem	819	MC	08	1	1160	-0.0100				1.11
2013_Chem	819	MC	09	1	1160	0.1000				0.96
2013_Chem	819	MC	10	1	1160	-1.4800				0.96
2013_Chem	819	MC	11	1	1160	-0.1800				0.90
2013_Chem	819	MC	12	1	1160	-1.0600				0.92
2013_Chem	819	CR	41	2	1160	0.5400	-1.7200	1.7200		0.93
2013_Chem	819	CR	42	1	1160	-0.5756				0.89
2013_Chem	819	CR	43	1	1160	-0.4989				0.89
2013_Chem	819	CR	44	3	1160	0.7625	0.1357	-0.4942	0.3585	1.23
2013_Chem	819	CR	45	1	1160	1.1389				0.86
2013_Chem	819	CR	46	1	1160	-0.2509				1.17
2013_Chem	819	CR	47	1	1160	0.4200				0.99

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.

Test	Form	Type	Item	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
2013_Chem	811	MC	01	-0.41	0.70	-0.07		
2013_Chem	811	MC	02	-0.07	0.03	-0.01		
2013_Chem	811	MC	03	0.32	0.68	0.07		
2013_Chem	811	MC	04	-0.21	0.47	-0.03		
2013_Chem	811	MC	05	0.03	0.01	0.00		
2013_Chem	811	MC	06	0.46	0.43	0.04		
2013_Chem	811	MC	07	0.48	2.08	0.10		
2013_Chem	811	MC	08	-0.18	0.27	-0.04		
2013_Chem	811	MC	09	0.46	1.82	0.09		
2013_Chem	811	MC	10	-0.39	1.25	-0.05		
2013_Chem	811	MC	11	0.48	1.88	0.09		
2013_Chem	811	MC	12	-0.38	1.27	-0.07		
2013_Chem	811	MC	13	-0.33	0.75	-0.06		
2013_Chem	811	MC	14	0.11	0.08	0.02		
2013_Chem	811	MC	15	-0.47	2.07	-0.08		
2013_Chem	811	CR	41		0.64	-0.06		
2013_Chem	811	CR	42		0.85	-0.07		
2013_Chem	811	CR	43		0.88	0.04		
2013_Chem	811	CR	44		9.37	0.19	BB	Females
2013_Chem	811	CR	45		0.07	0.01		
2013_Chem	811	CR	46		5.12	0.13		
2013_Chem	811	CR	47		4.70	-0.13		
2013_Chem	811	CR	48		0.54	-0.04		
2013_Chem	811	CR	49		3.02	-0.09		
2013_Chem	812	MC	01	-0.65	3.13	-0.10		
2013_Chem	812	MC	02	-0.09	0.05	-0.01		
2013_Chem	812	MC	03	0.38	1.11	0.07		
2013_Chem	812	MC	04	0.39	1.32	0.06		
2013_Chem	812	MC	05	0.19	0.36	0.03		
2013_Chem	812	MC	06	-0.52	2.56	-0.09		
2013_Chem	812	MC	07	0.30	0.81	0.05		
2013_Chem	812	MC	08	-0.02	0.00	-0.01		

Test	Form	Type	Item	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
2013_Chem	812	MC	09	-0.82	4.91	-0.15		
2013_Chem	812	MC	10	-0.41	1.00	-0.04		
2013_Chem	812	MC	11	-0.58	2.82	-0.12		
2013_Chem	812	MC	12	1.24	11.47	0.20	B	Females
2013_Chem	812	MC	13	-0.16	0.24	-0.04		
2013_Chem	812	MC	14	-0.09	0.07	-0.03		
2013_Chem	812	MC	15	0.14	0.18	0.03		
2013_Chem	812	CR	41		0.73	-0.04		
2013_Chem	812	CR	42		0.15	0.03		
2013_Chem	812	CR	43		3.28	0.11		
2013_Chem	812	CR	44		0.80	-0.04		
2013_Chem	812	CR	45		1.08	-0.04		
2013_Chem	812	CR	46		0.05	0.02		
2013_Chem	812	CR	47		0.00	-0.01		
2013_Chem	812	CR	48		5.00	0.12		
2013_Chem	812	CR	49		0.02	0.01		
2013_Chem	812	CR	50		1.05	-0.06		
2013_Chem	813	MC	01	0.24	0.46	0.07		
2013_Chem	813	MC	02	-1.01	9.52	-0.21	B	Males
2013_Chem	813	MC	03	0.56	1.53	0.07		
2013_Chem	813	MC	04	0.22	0.29	0.02		
2013_Chem	813	MC	05	-0.47	1.12	-0.06		
2013_Chem	813	MC	06	-1.02	7.64	-0.15	B	Males
2013_Chem	813	MC	07	-0.20	0.35	-0.06		
2013_Chem	813	MC	08	0.11	0.06	0.01		
2013_Chem	813	MC	09	0.39	1.25	0.06		
2013_Chem	813	MC	10	-1.24	11.30	-0.19	B	Males
2013_Chem	813	MC	11	0.11	0.05	0.01		
2013_Chem	813	MC	12	1.06	9.01	0.19	B	Females
2013_Chem	813	MC	13	-0.05	0.02	0.00		
2013_Chem	813	MC	14	-0.31	0.83	-0.03		
2013_Chem	813	CR	41		1.93	-0.10		
2013_Chem	813	CR	42		0.67	-0.05		
2013_Chem	813	CR	43		5.42	0.14		
2013_Chem	813	CR	44		2.02	-0.08		
2013_Chem	813	CR	45		1.22	-0.06		
2013_Chem	813	CR	46		5.84	0.11		
2013_Chem	813	CR	47		13.41	0.21	BB	Females
2013_Chem	813	CR	48		0.17	0.02		

Test	Form	Type	Item	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
2013_Chem	813	CR	49		5.72	0.14		
2013_Chem	813	CR	50		1.02	-0.06		
2013_Chem	813	CR	51		0.06	-0.02		
2013_Chem	814	MC	01	0.40	1.47	0.09		
2013_Chem	814	MC	02	-0.01	0.00	0.00		
2013_Chem	814	MC	03	-0.57	2.66	-0.10		
2013_Chem	814	MC	04	-0.35	0.98	-0.07		
2013_Chem	814	MC	05	-0.23	0.29	-0.01		
2013_Chem	814	MC	06	0.16	0.22	0.01		
2013_Chem	814	MC	07	-0.58	3.10	-0.11		
2013_Chem	814	MC	08	-0.01	0.00	-0.01		
2013_Chem	814	MC	09	0.12	0.16	0.04		
2013_Chem	814	MC	10	0.40	1.00	0.05		
2013_Chem	814	MC	11	-0.46	2.15	-0.10		
2013_Chem	814	MC	12	0.30	0.76	0.06		
2013_Chem	814	MC	13	-0.15	0.16	-0.03		
2013_Chem	814	CR	41		1.70	-0.06		
2013_Chem	814	CR	42		0.08	0.02		
2013_Chem	814	CR	43		0.14	-0.02		
2013_Chem	814	CR	44		7.31	0.16		
2013_Chem	814	CR	45		0.45	0.05		
2013_Chem	814	CR	46		5.66	0.13		
2013_Chem	814	CR	47		2.35	-0.07		
2013_Chem	814	CR	48		0.00	-0.01		
2013_Chem	814	CR	49		0.16	0.01		
2013_Chem	814	CR	50		0.02	0.00		
2013_Chem	814	CR	51		0.47	-0.04		
2013_Chem	814	CR	52		1.57	0.06		
2013_Chem	815	MC	01	-0.28	0.75	-0.07		
2013_Chem	815	MC	02	-0.96	6.16	-0.14		
2013_Chem	815	MC	03	0.30	0.80	0.06		
2013_Chem	815	MC	04	0.19	0.25	0.03		
2013_Chem	815	MC	05	-0.76	5.41	-0.14		
2013_Chem	815	MC	06	0.18	0.25	0.03		
2013_Chem	815	MC	07	0.95	8.20	0.17		
2013_Chem	815	MC	08	0.67	3.33	0.11		
2013_Chem	815	MC	09	-0.80	6.06	-0.14		
2013_Chem	815	MC	10	-0.06	0.02	-0.01		
2013_Chem	815	MC	11	1.37	16.87	0.24	B	Females

Test	Form	Type	Item	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
2013_Chem	815	MC	12	-0.06	0.03	-0.01		
2013_Chem	815	MC	13	-0.14	0.16	-0.02		
2013_Chem	815	MC	14	-1.55	21.45	-0.28	C	Males
2013_Chem	815	MC	15	-0.17	0.23	-0.02		
2013_Chem	815	CR	41		1.34	-0.07		
2013_Chem	815	CR	42		0.63	-0.03		
2013_Chem	815	CR	43		1.30	0.06		
2013_Chem	815	CR	44		7.19	0.14		
2013_Chem	815	CR	45		0.25	0.03		
2013_Chem	815	CR	46		1.22	0.05		
2013_Chem	815	CR	47		0.45	0.04		
2013_Chem	815	CR	48		3.47	-0.10		
2013_Chem	815	CR	49		0.02	0.01		
2013_Chem	815	CR	50		0.39	0.04		
2013_Chem	816	MC	01	0.85	2.71	0.08		
2013_Chem	816	MC	02	-0.97	7.27	-0.16		
2013_Chem	816	MC	03	0.34	0.68	0.07		
2013_Chem	816	MC	04	-0.36	0.79	-0.05		
2013_Chem	816	MC	05	0.34	0.99	0.06		
2013_Chem	816	MC	06	0.28	0.57	0.04		
2013_Chem	816	MC	07	0.20	0.24	0.02		
2013_Chem	816	MC	08	-0.01	0.00	-0.02		
2013_Chem	816	MC	09	0.59	2.35	0.10		
2013_Chem	816	MC	10	-0.43	1.39	-0.07		
2013_Chem	816	MC	11	0.20	0.31	0.03		
2013_Chem	816	MC	12	0.36	0.56	0.05		
2013_Chem	816	MC	13	-0.30	0.82	-0.06		
2013_Chem	816	MC	14	0.47	2.14	0.09		
2013_Chem	816	CR	41		0.09	-0.04		
2013_Chem	816	CR	42		0.86	-0.06		
2013_Chem	816	CR	43		4.57	-0.11		
2013_Chem	816	CR	44		0.53	-0.05		
2013_Chem	816	CR	45		1.11	-0.06		
2013_Chem	816	CR	46		0.16	-0.03		
2013_Chem	816	CR	47		1.91	0.09		
2013_Chem	816	CR	48		1.70	0.08		
2013_Chem	816	CR	49		0.00	-0.01		
2013_Chem	816	CR	50		0.18	0.00		
2013_Chem	816	CR	51		0.48	0.04		

Test	Form	Type	Item	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
2013_Chem	817	MC	01	-1.40	14.72	-0.23	B	Males
2013_Chem	817	MC	02	0.88	7.28	0.16		
2013_Chem	817	MC	03	0.06	0.03	0.00		
2013_Chem	817	MC	04	-0.18	0.31	-0.04		
2013_Chem	817	MC	05	0.32	0.78	0.05		
2013_Chem	817	MC	06	1.06	9.92	0.20	B	Females
2013_Chem	817	MC	07	0.16	0.24	0.04		
2013_Chem	817	MC	08	0.30	0.27	0.03		
2013_Chem	817	MC	09	-0.50	2.28	-0.07		
2013_Chem	817	MC	10	-0.19	0.28	-0.04		
2013_Chem	817	MC	11	-0.09	0.08	-0.02		
2013_Chem	817	MC	12	0.25	0.49	0.04		
2013_Chem	817	MC	13	0.29	0.39	0.05		
2013_Chem	817	MC	14	-0.24	0.55	-0.05		
2013_Chem	817	MC	15	-1.61	16.00	-0.25	C	Males
2013_Chem	817	MC	16	0.18	0.28	0.04		
2013_Chem	817	CR	41		1.76	0.07		
2013_Chem	817	CR	42		2.77	0.09		
2013_Chem	817	CR	43		1.82	-0.07		
2013_Chem	817	CR	44		3.10	-0.09		
2013_Chem	817	CR	45		1.77	-0.08		
2013_Chem	817	CR	46		0.47	0.05		
2013_Chem	817	CR	47		4.56	0.13		
2013_Chem	817	CR	48		0.22	-0.02		
2013_Chem	817	CR	49		0.10	-0.03		
2013_Chem	818	MC	01	0.03	0.01	0.01		
2013_Chem	818	MC	02	-0.07	0.04	0.01		
2013_Chem	818	MC	03	-0.31	0.54	-0.03		
2013_Chem	818	MC	04	-0.57	3.15	-0.10		
2013_Chem	818	MC	05	0.01	0.00	-0.01		
2013_Chem	818	MC	06	0.47	1.72	0.07		
2013_Chem	818	MC	07	-0.03	0.01	-0.03		
2013_Chem	818	MC	08	0.17	0.22	0.02		
2013_Chem	818	MC	09	-0.34	0.98	-0.07		
2013_Chem	818	MC	10	0.33	0.95	0.06		
2013_Chem	818	MC	11	-0.26	0.57	-0.02		
2013_Chem	818	MC	12	-0.02	0.00	0.01		
2013_Chem	818	MC	13	0.18	0.30	0.03		
2013_Chem	818	MC	14	0.38	1.30	0.04		

Test	Form	Type	Item	MH Delta	MH Chi-Sq	Effect Size	DIF Category	Favored Group
2013_Chem	818	CR	41		0.23	0.02		
2013_Chem	818	CR	42		0.33	0.05		
2013_Chem	818	CR	43		0.58	0.06		
2013_Chem	818	CR	44		0.08	0.03		
2013_Chem	818	CR	45		0.25	0.03		
2013_Chem	818	CR	46		0.47	-0.04		
2013_Chem	818	CR	47		1.75	-0.08		
2013_Chem	818	CR	48		0.00	0.00		
2013_Chem	818	CR	49		2.01	-0.07		
2013_Chem	818	CR	50		3.49	0.10		
2013_Chem	818	CR	51		2.82	-0.10		
2013_Chem	819	MC	01	-0.24	0.54	-0.03		
2013_Chem	819	MC	02	-1.02	6.23	-0.13	B	Males
2013_Chem	819	MC	03	-0.10	0.08	-0.04		
2013_Chem	819	MC	04	0.00	0.00	-0.01		
2013_Chem	819	MC	05	-0.62	3.06	-0.09		
2013_Chem	819	MC	06	0.84	4.35	0.13		
2013_Chem	819	MC	07	-1.25	5.29	-0.14	B	Males
2013_Chem	819	MC	08	1.01	10.25	0.18	B	Females
2013_Chem	819	MC	09	0.21	0.40	0.03		
2013_Chem	819	MC	10	-0.02	0.00	0.00		
2013_Chem	819	MC	11	0.38	1.18	0.06		
2013_Chem	819	MC	12	-0.37	0.83	-0.05		
2013_Chem	819	CR	41		1.32	-0.06		
2013_Chem	819	CR	42		1.35	0.08		
2013_Chem	819	CR	43		1.13	-0.05		
2013_Chem	819	CR	44		0.33	0.03		
2013_Chem	819	CR	45		1.35	-0.06		
2013_Chem	819	CR	46		0.54	0.05		
2013_Chem	819	CR	47		0.29	0.03		

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	4	3	1	0.78	0.31	-1.11	1.05
2	MC	1	1	4	3	1	0.85	0.43	-1.66	0.91
3	MC	1	1	4	3	1	0.87	0.44	-1.87	0.90
4	MC	1	1	4	3	1	0.54	0.52	0.19	0.92
5	MC	1	1	4	3	1	0.75	0.30	-1.10	1.10
6	MC	1	1	4	3	2	0.80	0.48	-1.26	0.90
7	MC	1	1	4	5	2	0.75	0.44	-0.94	0.99
8	MC	1	1	4	3	1	0.82	0.38	-1.57	0.95
9	MC	1	1	4	5	2	0.51	0.24	0.35	1.22
10	MC	1	1	4	5	2	0.46	0.43	0.59	1.05
11	MC	1	1	4	3	4	0.73	0.41	-0.82	1.03
12	MC	1	1	4	3	1	0.65	0.47	-0.54	0.97
13	MC	1	1	4	3	1	0.92	0.24	-2.51	1.00
14	MC	1	1	4	3	1	0.81	0.40	-1.36	0.99
15	MC	1	1	4	4	1	0.65	0.36	-0.39	1.11
16	MC	1	1	4	4	2	0.61	0.42	-0.16	1.01
17	MC	1	1	4	3	4	0.77	0.42	-1.04	0.95
18	MC	1	1	4	3	1	0.75	0.30	-0.95	1.14
19	MC	1	1	4	3	1	0.66	0.45	-0.43	1.01
20	MC	1	1	4	4	2	0.49	0.38	0.40	1.07
21	MC	1	1	4	3	1	0.60	0.50	-0.15	0.95
22	MC	1	1	4	5	2	0.64	0.39	-0.32	1.08
23	MC	1	1	4	3	4	0.54	0.39	0.20	1.04
24	MC	1	1	4	3	1	0.63	0.36	-0.30	1.11
25	MC	1	1	4	3	3	0.75	0.45	-0.95	0.95
26	MC	1	1	4	3	1	0.52	0.43	0.26	1.01
27	MC	1	1	4	3	1	0.40	0.39	0.85	1.04
28	MC	1	1	4	3	1	0.53	0.45	0.24	1.03
29	MC	1	1	4	3	1	0.67	0.31	-0.49	1.16
30	MC	1	1	4	4	4	0.58	0.43	-0.02	1.01
31	MC	1	1	4	3	1	0.62	0.42	-0.21	1.04
32	MC	1	1	6	3	2	0.37	0.39	1.05	1.07
33	MC	1	1	4	3	1	0.56	0.45	0.08	0.98
34	MC	1	1	1	M3	1	0.58	0.47	-0.02	0.95
35	MC	1	1	4	3	3	0.64	0.50	-0.34	0.94

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
36	MC	1	1	1	M3	1	0.65	0.42	-0.56	0.93
37	MC	1	1	1	S3	1	0.75	0.42	-0.94	1.01
38	MC	1	1	4	5	2	0.71	0.51	-0.71	0.88
39	MC	1	1	4	5	2	0.54	0.49	0.20	0.97
40	MC	1	1	1	S3	1	0.45	0.42	0.63	1.01
41	MC	1	1	4	5	3	0.62	0.42	-0.22	1.00
42	MC	1	1	1	M1	1	0.59	0.49	-0.05	0.96
43	MC	1	1	4	4	2	0.63	0.27	-0.50	1.08
44	MC	1	1	4	3	2	0.47	0.44	0.51	1.03
45	MC	1	1	6	2	4	0.58	0.27	0.01	1.24
46	MC	1	1	1	S3	1	0.73	0.33	-0.76	1.06
47	MC	1	1	4	3	3	0.41	0.44	0.82	1.03
48	MC	1	1	4	5	3	0.47	0.41	0.52	1.08
49	MC	1	1	1	S1	1	0.68	0.46	-0.51	0.95
50	MC	1	1	6	2	2	0.60	0.33	-0.11	1.12
51	CR	1	1	4	5	2	0.75	0.39	-0.95	1.01
52	CR	1	1	1	S1	1	0.44	0.47	0.69	1.00
53	CR	1	1	1	M3	1	0.53	0.58	0.22	0.86
54	CR	1	1	4	3	4	0.67	0.45	-0.46	1.00
55	CR	1	1	1	M1	1	0.90	0.27	-2.05	1.01
56	CR	1	1	1	M1	1	0.87	0.31	-1.83	0.97
57	CR	1	1	1	M2	1	0.82	0.31	-1.36	1.01
58	CR	1	1	4	5	2	0.30	0.51	1.36	0.89
59	CR	1	1	1	S1	1	0.38	0.42	0.95	1.01
60	CR	1	1	4	3	3	0.46	0.55	0.54	0.88
61	CR	1	1	4	3	3	0.31	0.46	1.33	0.96
62	CR	1	1	1	M3	1	0.28	0.45	1.51	0.99
63	CR	1	1	4	4	1	0.66	0.48	-0.45	0.96
64	CR	1	1	4	3	1	0.48	0.43	0.48	1.05
65	CR	1	1	1	M1	1	0.47	0.52	0.54	0.94
66	CR	1	1	4	5	2	0.62	0.50	-0.21	0.91
67	CR	1	1	4	3	1	0.45	0.60	0.61	0.82
68	CR	1	1	4	3	1	0.17	0.47	2.21	0.88
69	CR	1	1	4	3	1	0.20	0.42	2.03	0.96
70	CR	1	1	4	3	2	0.44	0.39	0.68	1.09
71	CR	1	1	4	3	1	0.47	0.45	0.50	1.02
72	CR	1	1	4	3	2	0.20	0.49	2.05	0.91
73	CR	1	1	4	3	1	0.34	0.53	1.19	0.90
74	CR	1	1	4	3	1	0.42	0.56	0.79	0.89

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
75	CR	1	1	4	3	1	0.46	0.60	0.59	0.85
76	CR	1	1	4	3	3	0.43	0.59	0.76	0.85
77	CR	1	1	4	3	1	0.58	0.53	0.01	0.89
78	CR	1	1	1	M2	1	0.58	0.53	0.00	0.89
79	CR	1	1	6	2	4	0.40	0.44	0.83	1.00
80	CR	1	1	4	4	2	0.68	0.45	-0.52	0.99
81	CR	1	1	4	4	2	0.49	0.57	0.42	0.87
82	CR	1	1	4	4	2	0.19	0.45	2.14	0.94
83	CR	1	1	4	3	2	0.36	0.49	1.11	0.96
84	CR	1	1	1	S1	1	0.44	0.50	0.67	0.97
85	CR	1	1	4	3	1	0.32	0.52	1.33	0.91

June 2013

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
1	MC	1	1	4	3	1	0.88	0.38	-1.89	0.89
2	MC	1	1	4	3	1	0.72	0.43	-0.79	1.01
3	MC	1	1	4	3	1	0.51	0.53	0.32	0.92
4	MC	1	1	4	3	1	0.53	0.29	0.23	1.22
5	MC	1	1	4	3	1	0.54	0.35	0.19	1.06
6	MC	1	1	4	5	2	0.85	0.38	-1.63	0.97
7	MC	1	1	4	3	1	0.81	0.36	-1.32	1.01
8	MC	1	1	4	3	1	0.73	0.40	-0.83	1.04
9	MC	1	1	4	3	1	0.61	0.37	-0.15	1.09
10	MC	1	1	4	5	2	0.75	0.40	-0.93	1.02
11	MC	1	1	4	3	1	0.78	0.30	-1.06	1.02
12	MC	1	1	4	4	2	0.67	0.44	-0.46	1.00
13	MC	1	1	4	4	1	0.51	0.32	0.07	1.12
14	MC	1	1	4	4	2	0.75	0.48	-0.97	0.93
15	MC	1	1	4	3	4	0.68	0.40	-0.48	0.98
16	MC	1	1	4	3	1	0.72	0.38	-0.77	1.05
17	MC	1	1	4	5	2	0.66	0.49	-0.41	0.95
18	MC	1	1	4	3	1	0.37	0.38	0.98	1.03
19	MC	1	1	4	3	4	0.32	0.40	1.02	1.05
20	MC	1	1	4	3	1	0.87	0.41	-1.82	0.93
21	MC	1	1	4	3	1	0.65	0.26	-0.62	1.03
22	MC	1	1	4	3	1	0.74	0.46	-0.87	0.96
23	MC	1	1	4	3	2	0.73	0.40	-0.85	1.05
24	MC	1	1	4	3	2	0.65	0.24	-0.36	1.14
25	MC	1	1	4	3	2	0.75	0.36	-0.97	1.05
26	MC	1	1	4	3	1	0.55	0.48	0.13	0.98
27	MC	1	1	4	3	1	0.66	0.34	-0.41	1.05
28	MC	1	1	4	3	1	0.46	0.45	0.55	0.99
29	MC	1	1	4	4	4	0.85	0.39	-1.67	0.97
30	MC	1	1	4	4	4	0.74	0.37	-0.83	0.98
31	MC	1	1	4	3	1	0.73	0.41	-0.83	1.01
32	MC	1	1	1	M3	1	0.83	0.35	-1.51	1.02
33	MC	1	1	1	S1	1	0.62	0.37	-0.20	1.03
34	MC	1	1	6	2	1	0.75	0.46	-0.97	0.94
35	MC	1	1	4	3	2	0.75	0.50	-0.93	0.90
36	MC	1	1	1	M3	1	0.62	0.51	-0.24	0.93
37	MC	1	1	1	S3	1	0.57	0.40	0.03	1.01
38	MC	1	1	4	5	2	0.60	0.43	-0.11	1.02

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
39	MC	1	1	1	M1	1	0.83	0.37	-1.54	0.96
40	MC	1	1	4	4	2	0.55	0.32	0.13	1.09
41	MC	1	1	4	3	4	0.55	0.37	0.14	1.09
42	MC	1	1	4	3	1	0.87	0.42	-1.88	0.93
43	MC	1	1	4	3	1	0.47	0.39	0.55	1.10
44	MC	1	1	6	3	2	0.52	0.39	0.27	1.10
45	MC	1	1	4	3	1	0.67	0.49	-0.47	0.93
46	MC	1	1	1	M3	1	0.56	0.36	0.07	1.13
47	MC	1	1	2	1	NA	0.81	0.45	-1.37	0.91
48	MC	1	1	4	4	4	0.65	0.37	-0.33	1.02
49	MC	1	1	4	4	4	0.42	0.41	0.76	1.05
50	MC	1	1	4	4	4	0.51	0.43	0.33	0.98
51	CR	1	1	1	M2	1	0.80	0.42	-1.27	0.95
52	CR	1	1	4	5	2	0.55	0.50	0.11	0.96
53	CR	1	1	4	5	2	0.47	0.54	0.53	0.91
54	CR	1	1	1	S1	1	0.51	0.51	0.32	0.95
55	CR	1	1	4	5	2	0.69	0.40	-0.62	1.04
56	CR	1	1	1	S1	1	0.52	0.51	0.28	0.95
57	CR	1	1	4	4	2	0.31	0.38	1.30	1.03
58	CR	1	1	1	M2	1	0.66	0.34	-0.41	1.05
59	CR	1	1	4	4	2	0.46	0.55	0.53	0.87
60	CR	1	1	1	M2	1	0.68	0.27	-0.55	1.17
61	CR	1	1	4	3	4	0.48	0.44	0.45	1.02
62	CR	1	1	4	3	4	0.22	0.47	1.87	0.92
63	CR	1	1	1	S1	1	0.54	0.40	0.20	1.09
64	CR	1	1	4	3	3	0.34	0.46	1.21	1.01
65	CR	1	1	4	5	2	0.33	0.51	1.23	0.94
66	CR	1	1	1	S1	1	0.67	0.34	-0.45	1.03
67	CR	1	1	4	3	1	0.41	0.44	0.80	0.97
68	CR	1	1	4	3	1	0.64	0.37	-0.27	1.02
69	CR	1	1	4	5	2	0.38	0.49	0.94	0.92
70	CR	1	1	4	3	2	0.33	0.58	1.24	0.83
71	CR	1	1	4	3	2	0.43	0.59	0.72	0.85
72	CR	1	1	4	3	4	0.32	0.47	1.31	0.98
73	CR	1	1	4	3	4	0.63	0.45	-0.26	1.01
74	CR	1	1	1	S1	1	0.40	0.58	0.84	0.84
75	CR	1	1	4	3	1	0.36	0.45	1.06	0.99
76	CR	1	1	2	1	NA	0.60	0.53	-0.11	0.90
77	CR	1	1	1	S1	1	0.40	0.44	0.86	1.01

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
78	CR	1	1	4	4	1	0.38	0.59	0.95	0.81
79	CR	1	1	1	M3	1	0.16	0.42	2.28	0.94
80	CR	1	1	4	4	1	0.40	0.47	0.82	0.94
81	CR	1	1	2	2	NA	0.45	0.51	0.62	0.93
82	CR	1	1	4	3	1	0.34	0.47	1.19	0.97
83	CR	1	1	1	S1	1	0.33	0.54	1.23	0.89
84	CR	1	1	1	M3	1	0.35	0.52	1.15	0.94
85	CR	1	1	1	S3	1	0.30	0.49	1.41	0.95

Appendix G: Scoring Tables

January 2013

Raw Score	Ability	Scale Score
0	-6.109	0.000
1	-4.886	3.180
2	-4.166	5.978
3	-3.733	8.633
4	-3.419	11.173
5	-3.169	13.609
6	-2.961	15.941
7	-2.780	18.167
8	-2.621	20.286
9	-2.477	22.313
10	-2.346	24.248
11	-2.225	26.098
12	-2.113	27.886
13	-2.007	29.587
14	-1.907	31.226
15	-1.813	32.789
16	-1.723	34.311
17	-1.636	35.756
18	-1.554	37.144
19	-1.474	38.493
20	-1.397	39.778
21	-1.322	41.019
22	-1.249	42.215

Raw Score	Ability	Scale Score
23	-1.178	43.365
24	-1.109	44.480
25	-1.042	45.553
26	-0.976	46.588
27	-0.911	47.596
28	-0.848	48.571
29	-0.785	49.506
30	-0.723	50.421
31	-0.663	51.306
32	-0.603	52.168
33	-0.543	53.004
34	-0.485	53.822
35	-0.426	54.617
36	-0.369	55.391
37	-0.311	56.157
38	-0.254	56.905
39	-0.198	57.625
40	-0.141	58.344
41	-0.085	59.051
42	-0.028	59.741
43	0.028	60.429
44	0.084	61.107
45	0.141	61.777

Raw Score	Ability	Scale Score
46	0.197	62.445
47	0.254	63.110
48	0.311	63.769
49	0.368	64.432
50	0.426	65.098
51	0.484	65.762
52	0.542	66.424
53	0.601	67.088
54	0.661	67.763
55	0.722	68.446
56	0.783	69.129
57	0.845	69.824
58	0.908	70.530
59	0.973	71.245
60	1.039	71.976
61	1.106	72.724
62	1.174	73.485
63	1.245	74.261
64	1.317	75.055
65	1.391	75.872
66	1.468	76.714
67	1.547	77.578
68	1.630	78.460

Raw Score	Ability	Scale Score
69	1.716	79.378
70	1.805	80.322
71	1.899	81.296
72	1.998	82.307
73	2.104	83.350
74	2.216	84.432
75	2.336	85.557
76	2.466	86.725
77	2.609	87.939
78	2.768	89.207
79	2.948	90.529
80	3.156	91.908
81	3.404	93.349
82	3.718	94.863
83	4.149	96.454
84	4.869	98.130
85	6.091	100.000

June 2013

Raw Score	Ability	Scale Score
0	-6.088	0.000
1	-4.868	3.237
2	-4.153	6.045
3	-3.725	8.693
4	-3.414	11.212
5	-3.168	13.621
6	-2.963	15.913
7	-2.786	18.098
8	-2.629	20.174
9	-2.488	22.159
10	-2.359	24.053
11	-2.240	25.864
12	-2.130	27.611
13	-2.026	29.280
14	-1.928	30.886
15	-1.835	32.417
16	-1.746	33.912
17	-1.661	35.339
18	-1.580	36.700
19	-1.501	38.037
20	-1.425	39.308
21	-1.351	40.535
22	-1.279	41.724

Raw Score	Ability	Scale Score
23	-1.209	42.863
24	-1.141	43.973
25	-1.074	45.043
26	-1.009	46.077
27	-0.945	47.076
28	-0.881	48.061
29	-0.819	48.994
30	-0.758	49.911
31	-0.698	50.798
32	-0.638	51.664
33	-0.579	52.504
34	-0.520	53.325
35	-0.462	54.127
36	-0.405	54.907
37	-0.348	55.673
38	-0.291	56.428
39	-0.234	57.162
40	-0.178	57.879
41	-0.121	58.592
42	-0.065	59.293
43	-0.009	59.977
44	0.047	60.666
45	0.104	61.338

Raw Score	Ability	Scale Score
46	0.160	62.008
47	0.217	62.677
48	0.274	63.339
49	0.331	64.000
50	0.388	64.667
51	0.446	65.334
52	0.505	65.999
53	0.564	66.665
54	0.623	67.336
55	0.684	68.018
56	0.745	68.706
57	0.807	69.392
58	0.870	70.100
59	0.934	70.811
60	0.999	71.536
61	1.066	72.280
62	1.134	73.038
63	1.204	73.810
64	1.275	74.599
65	1.349	75.409
66	1.425	76.244
67	1.504	77.104
68	1.585	77.986

Raw Score	Ability	Scale Score
69	1.670	78.891
70	1.759	79.836
71	1.852	80.805
72	1.950	81.813
73	2.053	82.859
74	2.164	83.938
75	2.283	85.070
76	2.411	86.244
77	2.553	87.471
78	2.709	88.753
79	2.887	90.101
80	3.093	91.510
81	3.339	92.994
82	3.650	94.558
83	4.079	96.216
84	4.796	97.977
85	6.017	100.000