Copyright

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

Copyright © 2018 by the New York State Education Department.

Secure Materials.
All rights reserved. No part of this document may be reproduced or transmitted by any means. Use of these materials is expressly limited to the New York State Education Department.

Contents

COPYRIGHT II

CHAPTER 1: INTRODUCTION 1

1.1 INTRODUCTION 1
1.2 PURPOSES OF THE EXAM 1
1.3 TARGET POPULATIONS (STANDARD 7.2) 2

CHAPTER 2: CLASSICAL ITEM STATISTICS (STANDARD 4.10) 3

2.1 ITEM DIFFICULTY 3
2.2 ITEM DISCRIMINATION 4
2.3

5.4 EVIDENCE BASED ON RELATIONS TO OTHER VARIABLES 37
5.5 EVIDENCE BASED ON TESTING CONSEQUENCES 37
REFERENCES 39
APPENDIX A: OPERATIONAL TEST MAPS 43
APPENDIX B: RAW-TO-THETA-TO-SCALE SCORE CONVERSION TABLES 52
APPENDIX C: ITEM WRITING GUIDELINES 55
APPENDIX D: TABLES A

List of Tables

Table 1 Total Examined Population: Regents Examination in Chemistry 2
Table 2 Multiple-Choice Item Analysis Summary: Regents Examination in Chemistry 4
Table 3 Constructed-Response Item Analysis Summary: Regents Examination in Chemistry 6
Table 4 Descriptive Statistics in Value and Point-Biserial Correlation: Regents Examination in Chemistry 7
Table 5 Summary of Item Residual Correlations: Regents Examination in Chemistry 14
Table 6 Summary of INFIT Mean Square Statistics: Regents Examination in Chemistry 15
Table 7 Reliabilities and Standard Errors of Measurement: Regents Examination in Chemistry 20
Table 8 Decision Consistency and Accuracy Results: Regents Examination in Chemistry 25
Table 9 Group Means: Regents Examination in Chemistry 26
Table 10 State Percentile Ranking for Raw Score: Regents Examination in Chemistry 27
Table 11 Test Blueprint: Regents Examination in Chemistry 29

List of Figures

Figure 1 Scatterplot: Regents Examination in Chemistry 7
Figure 2 Student Performance Map: Regents Examination in Chemistry 10
Figure 3 Scree Plot: Regents Examination in Chemistry 12
Figure 4 Conditional Standard Error of Measurement: Regents Examination in Chemistry 16

Chapter 1: Introduction

1.1 INTRODUCTION

**Note: One student was not reported in the Ethnicity and Gender group, but that student is Students.

Chapter 2: Classical Item Statistics (Standard 4.10)

This chapter provides an overview of the two most familiar item-level statistics obtained from classical item analysis: item difficulty and item discrimination. The following results pertain only to the operational Regents Examination in Chemistry items.

2.1 ITEM DIFFICULTY

me

specified group (e.g., grade level).

$$\bar{x} = \frac{1}{n} \quad {}_{i=1}^{n} x_i$$

In the mean score formula above, the ol 18 C /l7 18 CclTQq0f347.35 [illegible] 29 [illegible]

Table 3 Constructed-Response Item Analysis Summary: Regents Examination in Chemistry

2.3 DISCRIMINATION ON DIFFICULTY SCATTER PLOT

Figure 1 shows a scatter plot of item discrimination values (*y*-axis) and item difficulty values (*x*-axis). The distributions of *p*-value and point-biserials, including mean, minimum, Q1, median, Q3, and maximum, are also presented in Table 4.

Figure 1 Scatter Plot: Regents Examination in Chemistry

Table 4 Descriptive Statistics in *p*-value and Point-Biserial Correlation: Regents Examination in Chemistry

Statistics	N	Mean	Min	Q1	Median	Q3	Max
-value	85	0.71	0.31	0.63	0.72	0.83	0.97
Point-Biserial	85	0.43	− 0.03	0.37	0.43	0.51	0.61

2.4 OBSERVATIONS AND INTERPRETATIONS

The *p*-values for the MC items ranged from about 0.37 to 0.97, while the mean proportion-correct values for the CR items (Table 3) ranged from about 0.31 to 0.91. From the difficulty distributions illustrated in the plot, a wide range of item difficulties appeared on each exam, which was one test development goal.

Figure 2 Student Performance Map: Regents Examination in Chemistry

3.5 CHECKING RASCH ASSUMPTIONS

Since the Rasch model was the basis of all calibration, scoring, and scaling analyses associated with the Regents Examination in Chemistry, the validity of the inferences from these results depends on the degree to which the assumptions of the model were met and how well the model fits the test data. Therefore, it is important to check these assumptions. This section evaluates the dimensionality of the data, local item independence, and item fit. It should be noted that only operational items were analyzed, since they are the basis of student scores.

Unidimensionality

Rasch models assume that one dominant dimension determines the differences in student performances. Principal Components Analysis (PCA) can be used to assess the unidimensionality assumption. The purpose of the analysis is to verify if any other dominant components exist among the items. If any other dimensions are found, the unidimensionoth00224 Tc 4

A parallel analysis (Horn, 1965) was conducted to help distinguish

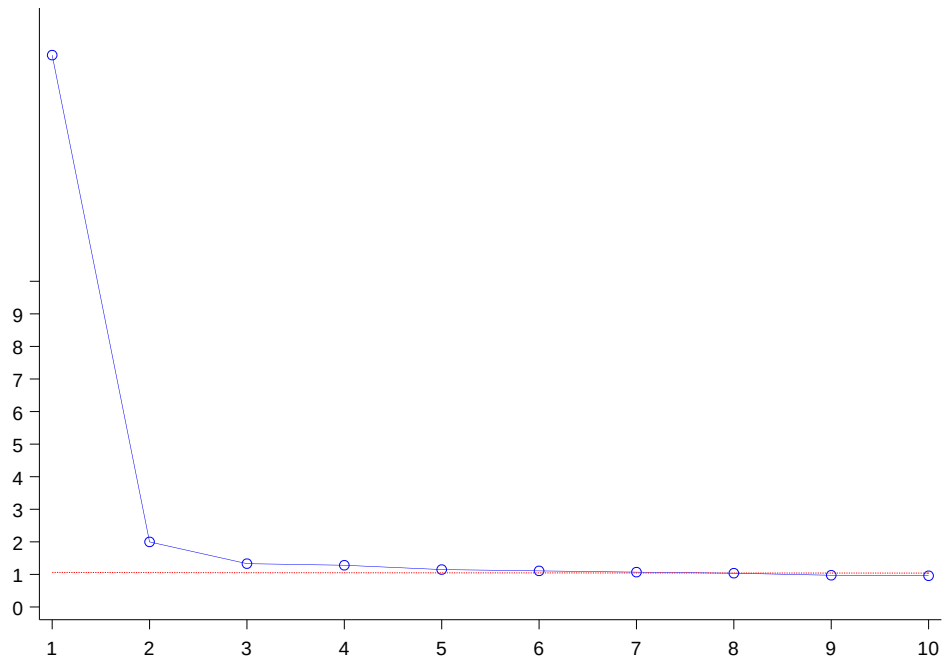

Figure 3 Scree Plot: Regents Examination in Chemistry

Local Independence

Local independence (LI) is a fundamental assumption of IRT. This means that, for statistical purposes, an examinee response to any one item should not depend on the response to any other item on the test. In formal statistical terms, a test X that is comprised of items X_1, X_2 X_n is locally independent with respect to the latent variable θ if, for all $x = (x_1, x_2 \quad x_n)$ and θ,

$$P(\mathbf{X} = \mathbf{x} \mid \quad) = \quad {}_{i=1}^{I} \; P(X_i = x_i \mid \quad).$$

This formula essentially states that the probability of any pattern of responses across all items (**x**) as measured by the test, should be equal to the product of the conditional probabilities across each item (i.e., the multiplication rule for independent events where the joint probabilities are equal to the product of the associated marginal probabilities).

The equation above shows the condition after satisfying the strong form of local independence. A weak form of local independence (WLI) is proposed by McDonald (1979). The distinction is important because many indicators of local dependency are actually framed by WLI. For WLI, the conditional covariances of all pairs of item responses, conditioned on the abilities, are assumed to be equal to zero. When this assump0 0 612 Td[(b)6 (e)6 ( )-31 (e)6 (q)16 (u)6 (a

$$P\left(X_i = x_i, X_j = x_j \mid \quad\right) = P\left(X_i = x_i \mid \quad\right) P\left(X_j = x_j \mid \quad\right).$$

Marais and Andrich (2008) point out that local item dependence in the Rasch model can

Table 5

from the August 2016, January 2017, and June 2017 administrations are on the same scale and can be directly compared to scale scores on all previous administrations back to the June 2004 administration.

When the base administration was concluded, the initial raw score to scale score relationship was established. Three 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 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:

سي 3 طى ١ 2 طى ٠ 1 طى 0,

where SS is the scaled score, RS is the raw score, and m0 through m3 are the transformation constants that convert the raw score into the scale score (please note that m0 will always be equal to zero in this application, since a raw score of zero corresponds to a scale score of zero). A subscript for a person on both dependent and independent variables is not present for simplicity. The above relationship and the values of m1 to m3 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 then used to derive a raw score to Rasch student ability (theta score) relationship. This aseil(to)4e daarelave a

Reliability coefficients range from 0.0 to 1.0. The index will be 0.0 if none of the test score variances is true. If all test score variances were true, the index would equal 1.0. Such scores would be pure random noise (i.e., all measurement error). If the index achieved a value of 1.0, scores would be perfectly consistent (i.e., contain no measurement error). Although values of 1.0 are never achieved in practice, it is clear that larger coefficients are more desirable because they indicate that the test scores are less influenced by random error.

Coefficient Alpha

Reliability is most often estimated using the formula for Coefficient Alpha, which provides a practical internal consistency index. It can be conceptualized as the extent to which an exchangeable set of items from the same domain would result in a similar rank ordering of students. Note that relative error is reflected in this index. Excessive variation in student performance from one sample of items to the next should be of particular concern for any achievement test user.

A general computational formula for Coefficient Alpha is as follows:

$$\text{سي} \quad \frac{\text{إ}}{\text{إ ٩٥}} \quad \text{حي} \quad 1 \quad \text{طي} \quad \sum \quad \text{عك فيتي}$$

standard deviation) into account. Consider that a SEM of 3 on a 10-point test would be very different from a SEM of 3 on a 100-point test.

Traditional Standard Error of Measurement Confidence Intervals

The SEM is an index of the random variability in test scores reported in actual score units, which is why it has such great utility for test score users. SEMs allow statements regarding the
around observed scores through construction of an approximate score band. Often referred to as confidence intervals, these bands are constructed by taking the observed scores, *X*, and adding and subtracting a multiplicative factor of the SEM. As an example, students with a given true score will have observed scores that fall between ±1 SEM about two-thirds of the time.[4] For ±2 SEM confidence intervals, this increases to about 95 percent.

The Coefficient Alpha and associated SEM for the Regents Examination in Chemistry are provided in Table 7.

Table 7 Reliabilities and Standard Errors of Measurement: Regents Examination in Chemistry

Conditional Standard Error of Measurement Characteristics

The relationship between the scale score CSEM and θ depends both on the nature of the raw-to-scale score transformation (Kolen and Brennan, 2005; Kolen and Lee, 2011) and on The pattern of CSEMs for raw scores and linear transformations of the raw score tend to have a - and larger CSEMs towards the middle of the distribution.

Figure 4 Conditional Standard Error Plot: Regents Examination in Chemistry

4.3 DECISION CONSISTENCY AND ACCURACY (STAND

		TEST ONE		
		LEVEL	LEVEL I	MARGINAL
TEST TWO	LEVEL I	11	12	1●
	LEVEL II	21	22	2●
	MARGINAL	●1	●2	1

Figure 5 Pseudo-Decision Table for Two Hypothetical Categories

Since true scores are unobserved and decision consistency is computed based on a single administration of the Regents Examination in Chemistry, a statistical model using solely data from the available administration is used to estimate the true scores and to project the consistency and accuracy of classifications (Hambleton & Novick, 1973). Although a number of procedures are available, a well-known method developed by Livingston and Lewis (1995) that utilizes a specific true score model is used.

Several factors might affect decision consistency and accuracy. One important factor is the reliability of the scores. All other things being equal, more reliable test scores tend to result in more similar reclassifications and less measurement error. Another factor is the location of the cut score in the score distribution. More consistent and accurate classifications are observed when the cut scores are located away from the mass of the score distribution. The number of performance levels is also a consideration. Consistency and accuracy indices based on four performance levels should be lower than those based on two performance levels. This is not surprising, since classification and accuracy using four performance levels would allow more opportunity to change performance levels. Hence, there would be more classification errors and less accuracy with four performance levels, resulting in lower consistency indices.

Results and Observations

The results for the dichotomies created by the three cut scores are presented in Table 7. The tabled values are derived with the program *BB-Class* (Brennan, 2004) using the Livingston and Lewis method. Decision consistency ranged from 0.87 to 0.96, and the decision accuracy ranged from 0.91

Table 9 Group Means: Regents Examination in Chemistry

Table 10 le

Chapter 5: Validity (Standard 1)

Restating the purpose and uses of the Regents Examination in Chemistry, this exam measures examinee achievement against the New York State learning standards. The exam is prepared by teacher examination committees and New York State Education Department subject matter and testing specialists, and it provides teachers and students with important information about student learning and performance against the established curriculum standards. Results of this exam may be used to identify student strengths and needs in order to guide classroom teaching and learning. The exams also provide students, parents, counselors, administrators, and college admissions officers with objective and easily understood achievement information that may be used to inform empirically based educational and vocational decisions about students. As a state-provided objective benchmark, the Regents Examination in Chemistry is intended for use in satisfying state testing requirements for students who have finished a course in Chemistry. A passing score on the exam counts toward requirements for a high school diploma, as described in the New York State diploma requirements: http://www.nysed.gov/common/nysed/files/programs/curriculum-instruction/currentdiplomarequirements2.pdf. Results of the Regents Examination in Chemistry may also be used to satisfy various locally established requirements throughout the state.

The validity of score interpretations for the Regents Examination in Chemistry is supported by multiple sources of evidence. Chapter 1 of the *Standards for Educational Psychological Testing* (AERA et al., 2014) specifies five sources of validity evidence that are important to gather and document in order to support validity claims for an assessment:

- test content
- response processes
- internal test structure
- relation to other variables
- consequences of testing

It is important to note that these categories are not mutually exclusive. One source of validity evidence often falls into more than one category, as discussed in more detail in this chapter. Nevertheless, these classifications provide a useful framework within the *Standards* (AERA et al., 2014) for the discussion and documentation of validity evidence, so they are used here. The process of gathering evidence of the validity of score interpretations is best characterized as ongoing throughout test development, administration, scoring, reporting, and beyond.

5.1 EVIDENCE BASED ON TEST CONTENT

The validity of test content is fundamental to arguments that test scores are valid for their intended purpose. It demands that a test developer provide evidence that test content is well-aligned with the framework and standards used in curriculum and instruction. Accordingly, detailed attention was given to this correspondence between standards and test content during test design and construction.

5.2 EVIDENCE BASED ON RESPONSE PROCESSES

The second source of validity evidence is based on examinee response processes. This standard requires evidence that examinees are responding in the manner intended by the test items and rubrics and that raters are scoring those responses in a manner that is consistent with the rubrics. Accordingly, it is important to control and monitor whether construct-irrelevant variance in response patterns has been introduced at any point in the test development, administration, or scoring processes.

The controls and monitoring in place for the Regents Examination in Chemistry include the item development process, with attention paid to mitigating the introduction of construct-irrelevant variance. The development process described in the previous sections details the process and attention given to reducing the potential for construct irrelevance in response

3. Consistency or reliability of ratings (Congdon & McQueen, 2000; Harik, Clauser, Grabovsky, Nungester, Swanson, & Nandakumar, 2009; McQueen & Congdon, 1997; Myford & Wolfe, 2009; Mero & Motowidlo, 1995; Weinrott & Jones, 1984)
4. Rubric designs that facilitate consistency of ratings (Pecheone & Chung, 2006; Wolfe & Gitomer, 2000; Cronbach, Linn, Brennan, & Haertel, 1995; Cook & Beckman, 2009; Penny, Johnson, & Gordon, 2000; Smith, 1993; Leacock, Gonzalez, and Conarroe, 2014)

The distinct steps for operational test scoring include close attention to each of these elements and begin before the operational test is even selected. After the field test process, during which many more items than appear on the operational test are administered to a representative sample of students, a
across the range of possible responses for constructed-response items is selected. The
-
execution, the scores from which are used to generate important statistical information about the item. Training scorers to produce reliable and valid scores is the basis for creating rating guides and scoring ancillaries to be used during operational scoring.

To review and select these anchor papers, NYS educators serve as table leaders during the range-finding session. In the range-finding process, committees of educators receive a set of student papers for each field-tested question. Committee members familiarize themselves with each item type and score a number of responses that are representative of each of the different score points. After the independent scoring is completed, the committee reviews and discusses their results and determines consensus scores for the student responses. During this process, atypical responses are important to identify and annotate for use in training and live scoring. The range-finding results are then used to build training

constructed-response items. The final model response sets for the August 2016, January 2017, and June 2017 administrations of the Regents Examination in Chemistry are located at http://www.nysedregents.org/Chemistry.

During the range-finding and field test scoring processes, it is important to be aware of and control for sources of variation in scoring. One possible source of variation in constructed-response scores is unintended rater bias associated with items and examinee responses. Because the rater is often unaware of such bias, this type of variation may be the most challenging source of variation in scoring to control and measure. Rater biases can appear as severity or leniency in applying the scoring rubric. Bias also includes phenomena such as the halo effect, which occurs when good or poor performance on one element of the rubric encourages inaccurate scoring of other elements. These types of rater bias can be effectively controlled by training practices with a strict focus on rubric requirements.

The training process for operational scoring by state educators begins with a review and discussion of actual student work on constructed-response test items. This helps raters understand the range and characteristics typical of examinee responses, as well as the kinds of mistakes that students commonly make. This information is used to train raters on how to consistently apply key elements of the scoring rubric across the domain of student responses.

Dimensionality

In addition to model fit, a strong assumption of the Rasch model is that the construct measured by a test is unidimensional. Violation of this assumption might suggest that the test is measuring something other than the intended content and indicate that the quality of the test structure is compromised. A principal components analysis was conducted to test the assumption of unidimensionality, and the results provide strong evidence that a single dimension in the Regents Examination in Chemistry is explaining a large portion of the variance in student response data. This analysis does not characterize or explain the dimension, but a reasonable assumption can be made that the test is largely unidimensional.

variables) serve as a consequential validity argument, as well. This evidence supports conclusions based on test scores that social consequences are not likely to be traced to characteristics or qualities of the test itself.

Cronbach (1988), on the other hand, argues that negative consequences could invalidate test use. From this perspective, the test user is obligated to make the case for test use and to ensure appropriate and supported uses. Regardless of perspective on the nature of consequential validity, it is important to caution against uses that are not supported by the validity claims documented for this test. For example, use of this test to predict examinee scores on other tests is not directly supported by either the stated purposes or by the development process and research conducted on examinee d oo tests

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Barkaoui, Khaled. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, 18:3.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement,78 (p)-5 (roB)-3 pnm.01 (&)tT9aE[.)8 (2*

McQueen, J., & Congdon, P. J. (1997, March). *Rater severity in large-scale assessment: Is it invariant?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Mero, N. P., & Motowidlo, S. J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology, 80*(4), 517 524.

Messick, S. (1995). Standards of Validity and the validity of and standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5 8.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for r Td[(M, (S)-2 (zra)6.9 (n))10 (a)1 (W)-fw-81 (o)e( )--3 (ssm)3 (ial6.9 (n)462.7 653.5 Td[(a)-3 (cc3

Wang, X. M., Wong, K. F. E., & Kwong, J. Y. Y. (2010). The roles of rater goals and ratee performance levels in the distortion of performance ratings. *Journal of Applied Psychology, 95*(3), 546 561.

Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing* 15, 263 287.

Weinrott, L., & Jones, B. (1984). Overt verses covert assessment of observer reliability. *Child Development, 55,* 1125 1137.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review*. Journal of Occupational and Organizational Psychology, 67,* 189 205.

Wolfe, E. W., & Gitomer, D. H. (2000). *The influence of changes in assessment design on the psychometric quality of scores.* Princeton, NJ: Educational Testing Service.

Appendix A: Operational Test Maps

Table A.1 Test Map for August 2016 Administration

Appendix B: Raw-to-Theta-to-Scale Score Conversion Tables

Table B.1 Score Table for August 2016 Administration

Raw Score	Ability	Scale Score
0	−6.2826	0.000
1	−5.0557	2.631
2	−4.3302	5.120
3	−3.8928	7.525
4	−3.5737	9.856
5	−3.3194	12.117
6	−3.1064	14.297
7	−2.9217	16.410
8	−2.7579	18.456
9	−2.6101	20.433
10	−2.4749	22.347
11	−2.3499	24.192
12	−2.2334	25.972
13	−2.1240	27.705
14	−2.0205	29.369
15	−1.9223	30.981
16	−1.8285	32.525
17	−1.7388	34.040
18	−1.6524	35.488
19	−1.5692	36.878
20	−1.4887	38.243
21	−1.4106	39.543
22	−1.3348	40.803
23	−1.2609	42.021
24	−1.1889	43.193
25	−1.1185	44.333
26	−1.0494	45.434
27	−0.9818	46.496
28	−0.9154	47.530
29	−0.8500	48.535
30	−0.7856	49.498
31	−0.7221	50.440

CHECKLIST OF TEST CONSTRUCTION PRINCIPLES
(Multiple-Choice Items)

	YES	NO
1. Is the item significant?		
2. Does the item have curricular validity?		
3. Is the item presented in clear and simple language, with vocabulary kept as simple as possible?		
4.		

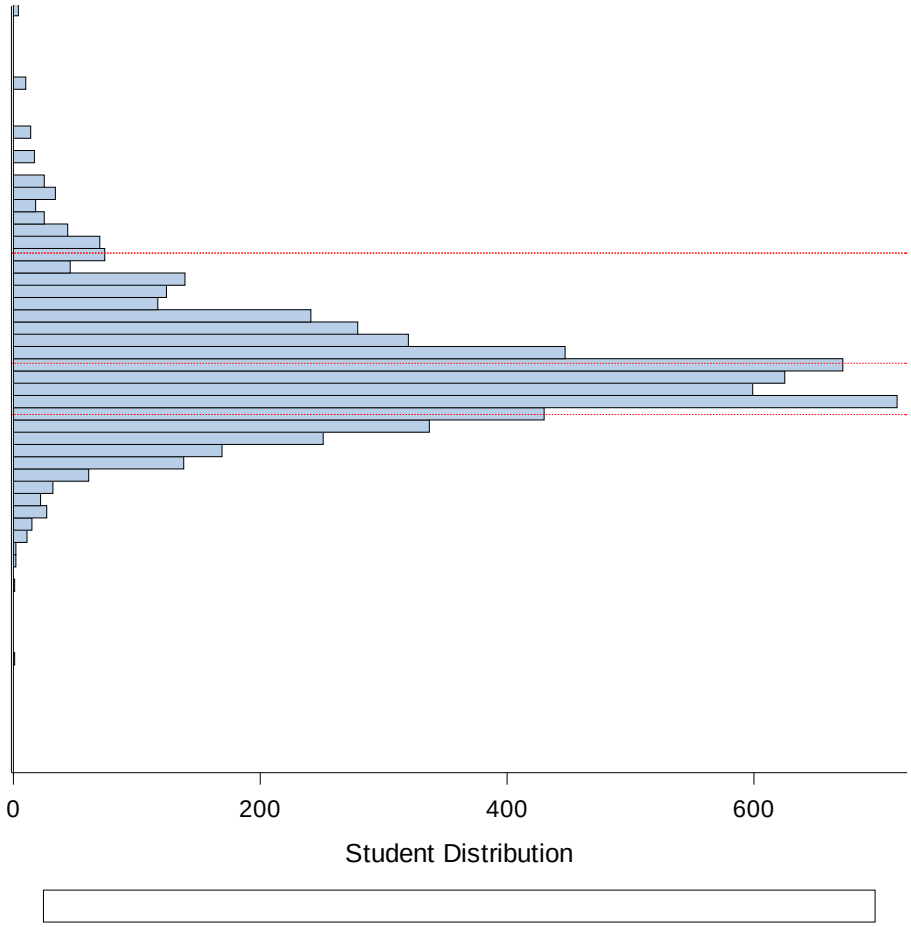

Figure D.2 Student Performance Map: Regents Examination in Chemistry

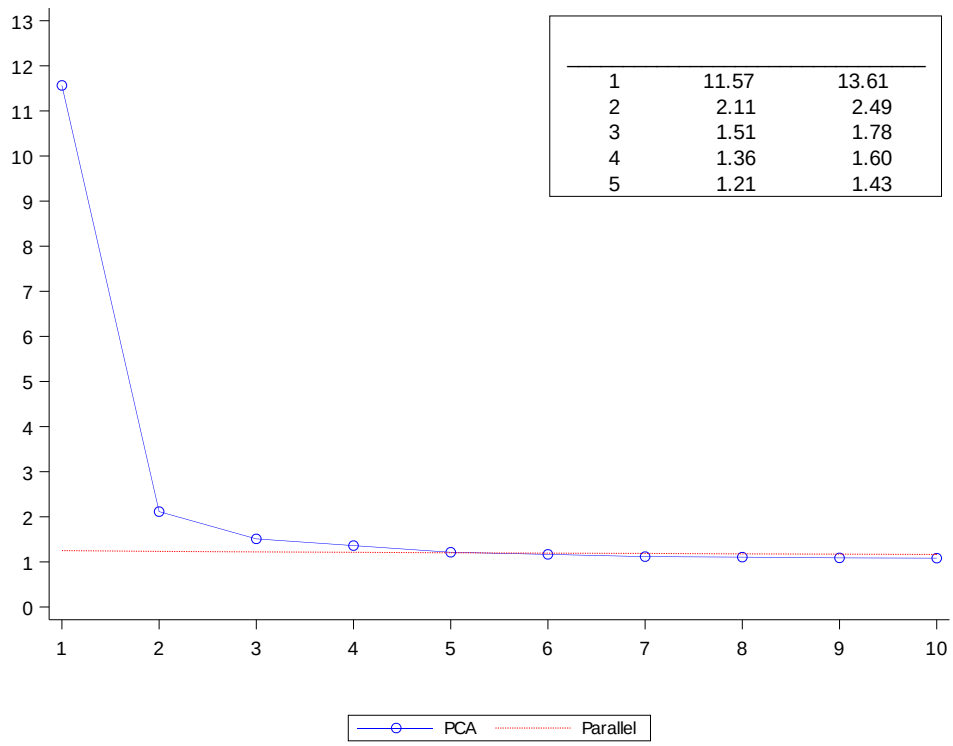

Figure D.3 Scree Plot: Regents Examination in Chemistry

Table D.4 Summary of Item Residual Correlations: Regents Examination in Chemistry

Statistic Type	Value
N	3570
Mean	−0.01
SD	0.02
Minimum	−0.09
P_{10}	−0.04
P_{25}	−0.03
P_{50}	−0.01
P_{75}	0.00
P_{90}	0.02
Maximum	0.14
>|0.20|	0

Table D.5 Summary of INFIT Mean Square Statistics: Regents Examination in Chemistry

	INFIT Mean Square					
	N	Mean	SD	Min	Max	[0.7, 1.3]
Chemistry	85	0.99	0.08	0.83	1.28	[85/85]

Table D.8 Group Means: Regents Examination in Chemistry

Demographics	Number	Mean Scale Score	SD Scale Score
All Students*	6,159	63.34	11.27
Ethnicity			
American Indian/Alaska Native	28	61.93	6.64
Asian/Native Hawaiian/Other Pacific Islander	745	68.24	12.87
Black/African American	1,081	59.48	11.25
Hispanic/Latino	1,160	59.51	10.80
Multiracial	88	63.23	8.42
White	3,055	64.97	10.23
English Language Learner			
No	6,135	63.40	11.21
Yes	24	47.08	14.98
Economically Disadvantaged			
No	3,782	64.80	11.06
Yes	2,377	61.01	11.20
Gender			
Female	3,493	63.62	11.32
Male	2,664	62.96	11.20
Student with a Disability			
No	5,907	63.60	11.13
Yes	252	57.04	12.62

*Note: Two students were not reported in the Ethnicity and Gender group, Students.

Appendix E: Tables and Figures for January 2017 Administration

Table E.1 Multiple-Choice Item Analysis Summary: Regents Examination in Chemistry

Table E.2 Constructed-Response Item Analysis Summary: Regents Examination in Chemistry

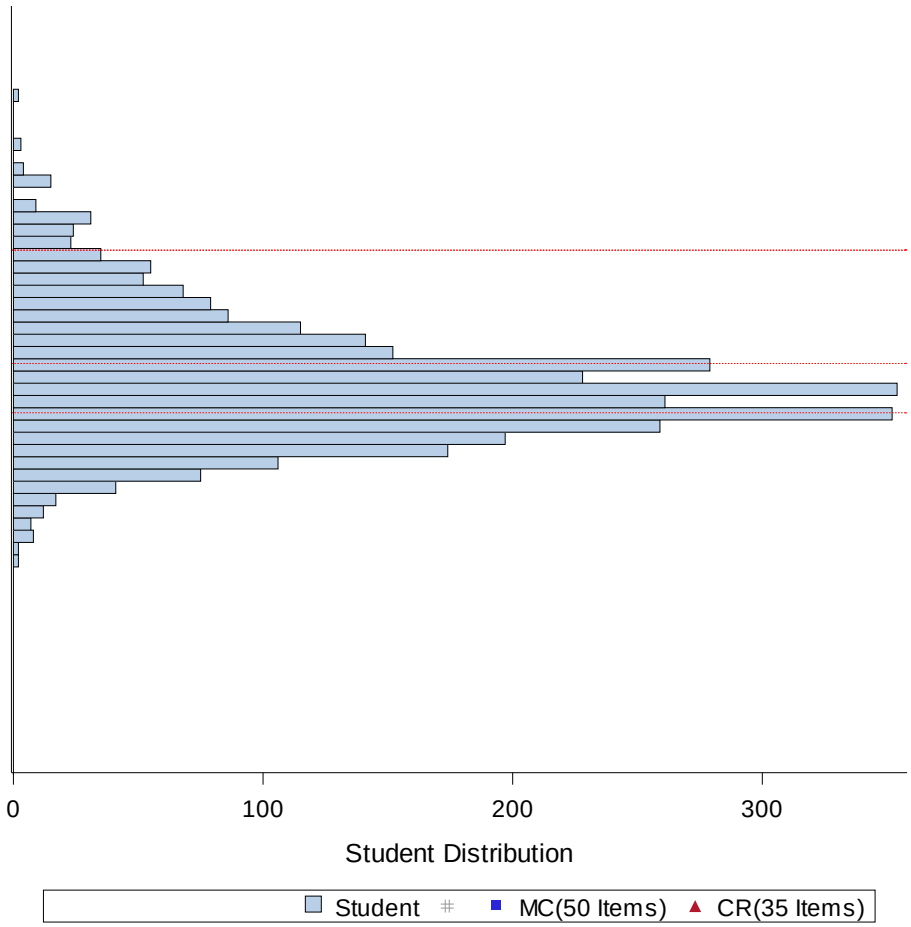
0
100
200
300
Student Distribution
Student
MC(50 Items)
CR(35 Items)

Figure E.2

Table E.5 Summary of INFIT Mean Square Statistics: Regents Examination in Chemistry

	INFIT Mean Square					
	N	Mean	SD	Min	Max	[0.7, 1.3]
Chemistry	85	1.00	0.09	0.82	1.29	[85/85]

Table E.6

Figure E.4 Conditional Standard Error Plot: Regents Examination in Chemistry

Table E.8 Group Means: Regents Examination in Chemistry