New York State Regents Examination in Chemistry

2018 Technical Report

Prepared for the New York State Education Department
by Pearson

March 2019

Copyright

Contents

COPYRIGHT II

CHAPTER 1: INTRODUCTION 1

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

CHAPTER 2: CLASSICAL ITEM STATISTICS (STANDARD 4.10) 4

2.1 ITEM DIFFICULTY 4
2.2 ITEM DISCRIMINATION 4
2.3 DISCRIMINATION ON DIFFICULTY SCATTER PLOT 8
2.4 OBSERVATIONS AND INTERPRETATIONS 8

CHAPTER 3: IRT CALIBRATIONS, EQUATING, AND SCALING (STANDARDS 2, AND 4.10) 9

3.1 DESCRIPTION OF THE RASCH MODEL 9
3.2 SOFTWARE AND ESTIMATION ALGORITHM 10
3.3 CHARACTERISTICS OF THE TESTING P

List of Tables

TABLE 1 TOTAL EXAMINEE POPULATION: REGENTS EXAMINATION IN CHEMISTRY 2
TABLE 2 MULTIPLE-CHOICE ITEM ANALYSIS SUMMARY: REGENTS EXAMINATION IN CHEMISTRY 5
TABLE 3 CONSTRUCTED-RESPONSE ITEM ANALYSIS SUMMARY: REGENTS EXAMINATION IN CHEMISTRY 7
TABLE 4 DESCRIPTIVE STATISTICS IN P-VALUE AND POINT-BISERIAL CORRELATION: REGENTS EXAMINATION IN CHEMISTRY 8
TABLE 5 SUMMARY OF ITEM RESIDUAL CORRELATIONS: REGENTS EXAMINATION IN CHEMISTRY 15
TABLE 6 SUMMARY OF INFIT MEAN SQUARE STATISTICS: REGENTS EXAMINATION IN CHEMISTRY 16
TABLE

Chapter 1: Introduction

1.1 INTRODUCTION

This technical report for the Regents Examination in Chemistry will provide New York State with documentation on the purpose of the Regents Examination, scoring information, evidence of both reliability and validity of the exams, scaling information, and guidelines and reporting informnfdu004 Tc8e a-nfo

1.3 TARGET POPULATION (STANDARD 7.2)

The examinee

**Note: One student was not reported in the Ethnicity and Gender group, but that student is reflected in “All 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

At the most general level, an item's difficulty is indicated by its mean score in some specified group (e.g., grade level).

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

In the mean score formula above, the individual item scores (x_i) are summed and then divided by the total number of students (*n*). For multiple-choice (MC) items, student scores are represented by 0s and 1s (0 = wrong, 1 = right). With 0–1 scoring, the equation above also represents the number of students correctly answering the item divided by the total number of students. Therefore, this is also the proportion correct for the item, or the *p*-value. In theory, *p*-values can range from 0.00 to 1.00 on the proportion-correct scale.[2] For example, if an MC item has a *p*-value of 0.89, it means that 89 percent of the students answered the item correctly. Additionally, this value might also suggest that the item was relatively easy and/or that the students who attempted the item were relatively high achievers. For constructed-response (CR) items, mean scores can range from the minimum possible score (usually zero) to the maximum possible score. To facilitate average score comparability across MC and CR items, mean item performance for CR items is divided by the maximum score possible so that the *p*-values for all items are reported as a ratio from 0.0 to 1.0.

Although the *p*-value statistic does not consider individual student ability in its computation, it provides a useful view of overall item difficulty, and can provide an early and simple indication of items that are too difficult for the population of students taking the examination. Items with very high or very low *p*-values receive added scrutiny during all follow-up analyses, including item response theory analyses that factor student ability into estimates of item difficulty. Such items may be removed from the item pool during the test development process, as field testing typically reveals that they add very little measurement information. Items for the June 2018 Regents Examination in Chemistry show a range of *p*-values consistent with the targeted exam difficulty. Item *p*-values, presented in Table 2 and Table 3 for multiple-choice and constructed-response items, respectively, range from 0.32 to 0.96, with a mean of 0.70. Table 2 and Table 3

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

Chapter 3: IRT Calibrations, Equating, and Scaling (Standards 2, and 4.10)

The item response theory (IRT) model used for the Regents Examination in Chemistry is based on the work of Georg Rasch (Rasch, 1960). The Rasch model has a long-standing presence in applied testing programs. IRT has several advantages over classical test theory, and it has become the standard procedure for analyzing item response data in large-scale assessments. According to van der Linden and Hambleton (1997), "The central feature of IRT is the specification of a mathematical function relating the probability of an examinee's response on a test item to an underlying ability." Ability, in this sense, can be thought of as performance on the test and is defined as "the expected value of observed performance on the test of interest" (Hambleton, Swaminathan, and Roger, 1991). This performance value is often referred to as . Performance and   will be used interchangeably throughout the remainder of this report.

A fundamental advantage of IRT is that it links examinee performance and item difficulty estimates and places them on the same scale, allowing for an evaluation of examinee performance that considers the difficulty of the test. This is particularly valuable for final test construction and test form equating, as it facilitates a fundamental attention to fairness for all examinees across items and test forms.

This chapter outlines the procedures used for calibrating the operational Regents Examination in Chemistry items. Generally,

$$P_{ni}(X = 1) = \frac{\exp(\theta_n - D_{ij})}{1 + \exp(\theta_n - D_{ij})}.$$

The Rasch model places both performance and item difficulty (estimated in terms of log-odds or logits) on the same continuum. When the model assumptions are met, the Rasch model provides estimates of examinee performance and item difficulty that are theoretically invariant across random samples of the same examinee population.

3.2 SOFTWARE AND ESTIMATION ALGORITHM

Item calibration was implemented via the WINSTEPS 3.60 computer program (Wright and Linacre, 2015), which employs unconditional (UCON), joint maximum 0 Tw 26.19 0 Tdw 26.2 26.2E>-4

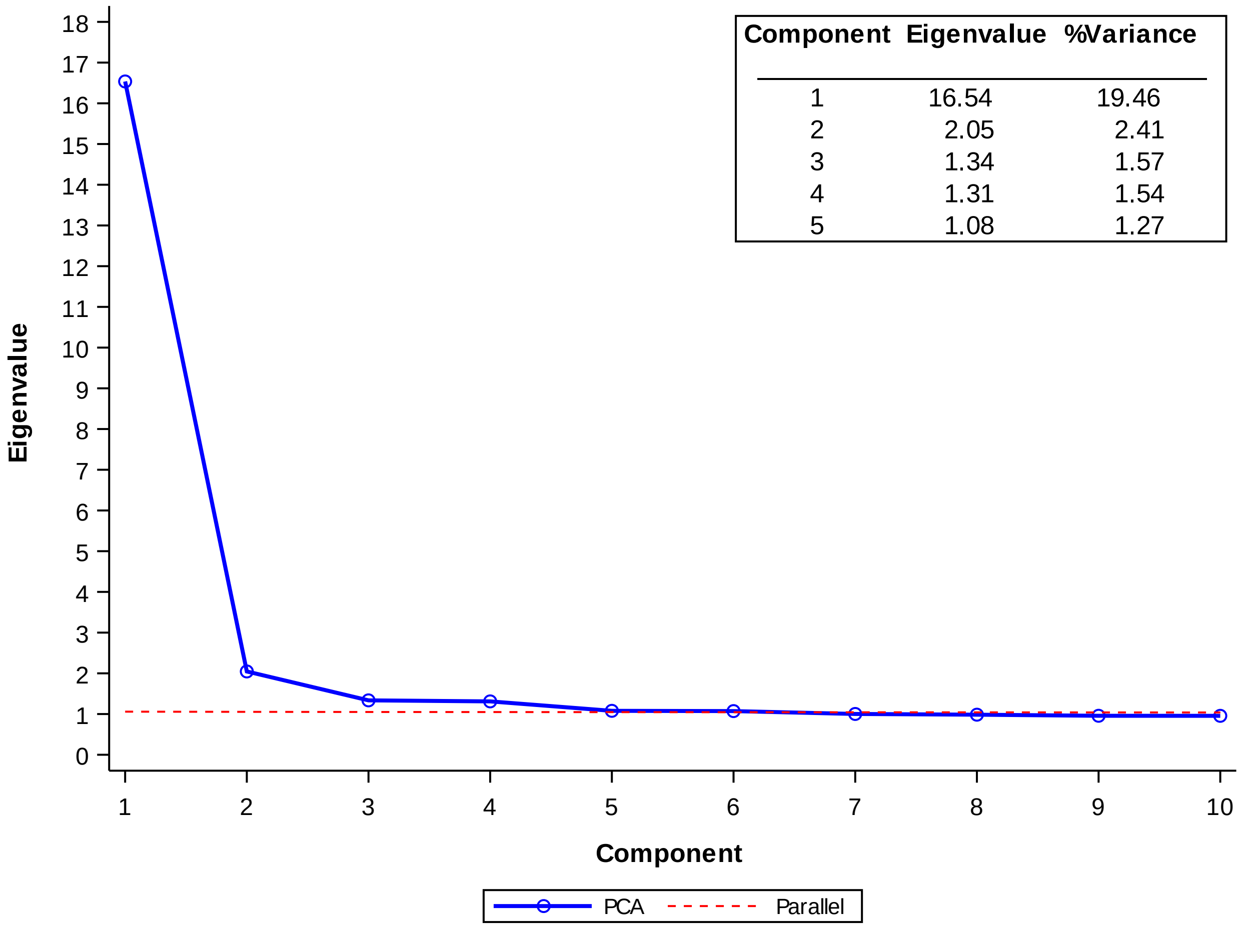


Figure 3 Scree Plot: Regents Examination in Chemiste

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 assumption is met, the joint probability of responses to an item pair, conditioned on the abilities, is the product of the probabilities of responses to these two items, as shown below. Based on the WLI, the following expression can be derived:

$$P\ X_i\quad x_i, X_j\quad x_j\,|\qquad P\ X_i\quad x_i\,|\quad P\ X_j\quad x_j\,|\quad .$$

Ma & s

Table 5 Summary of Item Residual Correlations: Regents Examination in Chemistry

Statistic Type	Value
N	3,570
Mean	-0.01
SD	0.02
Minimum	-0.09
P_{10}	-0.03
P_{25}	-0.02
P_{50}	-0.01
P_{75}	0.00
P_{90}	0.01
Maximum	0.15
>|0.20|	0

Item Fit

Table 6 Summary of INFIT Mean Square Statistics: Regents Examination in Chemistry

	INFITMean Square					
	N	Mean	SD	Min	Max	[0.7, 1.3]
Chemistry	85	1.00	0.10	0.79	1.24	[85/85]

Items for the Regents Examination in Chemistry were field tested in 2007–2010 and 2012–2017, and a separate technical report was produced for each year to document the full test development, scoring, scaling, and data analysis conducted.

3.6 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 determiermi

í

the August 2017, January 2018, and June 2018 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:

$$SS = m_3 \times RS^3 + m_2 \times RS^2 + m_1 \times RS + m_0,$$

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 applireean cuTw 5.49 0 Td (1 Tm72 ( )]TJ ET EMC BT 088 Tw 22.8(equ)1 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 55, 65, or 85, the lowest raw score's scale score is set equal to 55, 65, or 85 and the scale scores corresponding to the higher raw scores are set to 56, 66, or 86 as appropriate. If no scale score rounds to these 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.

Chapter 4: Reliability (Standard 2)

Test reliability is a measure of the internal consistency of a test (Cronbach, 1951). It is a measure of the extent to which the items on a test provide consistent information about student mastery of a domain. Reliability should ultimately demonstrate that examinee score estimates maximize consistency and therefore minimize error or, theoretically speaking, that examinees who take a test multiple times would get the same score each time.

According to the *Standards for Educational and Psychological Testing*, “A number of factors can have significant effects on reliability/precision, and in some cases, these factors can lead to misinterpretations of test scores, if not taken into account” (AERA et al., 2014, p. 38). First, test length and the (i)6 ()6 (dds)4 (m)-3 ((l)6 (i)6 (t)2 (y)4 (s)4 (h t)12 (f)12 (b)2 (pr)17 v10 (n i)6 d (on r) would be pure random noise (i.e., all measurement error). If the index achieved a value of 1.0,

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 ZKHWKHU WKH &6(0 LV GHULYHG IURP WKH UDZ VFRUHV RU IU The pattern of CSEMs for raw scores and linear transformations of the raw score tend to have a characteristic “inverted-U” shape, with smaller CSEMs at the ends of the score continuum and larger CSEMs towards the middle of the distribution.

Achievable raw score points for these distributions are spaced equally across the score range. Kolen and Brennan (2005, p. 357) state, “When, relative to raw scores, the transformation compresses the scale in the middle and stretches it at the ends, the pattern of the conditional standard errors of measurement will be concave up (U-shaped), even though the pattern for the raw scores was concave down (inverted-U shape).”

Results and Observations

The relationship between raw and scale scores for the Regents Exam

TEST TWO \ TEST ONE	LEVEL I	LEVEL II	MARGINAL
LEVEL I	11	12	í ·
LEVEL II	21	22	î ·
MARGINAL	· í	· î	1

Figure 5 Pseudo-Decision Table for Two Hypothetical Categories

TEST TWO \ TEST ONE	LEVEL I	LEVEL II	LEVEL III	LEVEL IV	MARGINAL
LEVEL I	11	12	13	14	í ·
LEVEL II	21	22	23	24	î ·
LEVEL III	31	32	33	34	ï ·
LEVEL IV	41	42	43	44	ð ·
MARGINAL	· í	· î	· ï	· ð	1

Figure 6 PseudoRGINAL [(6)-421 r2

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 t

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 Psy0 Td [(equi)6 (r. (o*

Content Validity

Content validity is necessarily concerned with the proper definition of the construct and evidence that the test provides an accurate measure of examinee performance within the defined construct. The test blueprint for the Regents Examination in Chemistry is essentially the design document for constructing the exam. It provides an explicit definition of the content domain that is to be represented on the exam

A

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 set of “anchor” papers representing student responses across the range of possible responses for constructed-response items is selected. The objective of these “range-finding” efforts is to create a training set for scorer training and 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 materials for the vendor’s scorers, who then score the rest of the field test responses to constructed-response items. The final rating guides for the August 2017, January 2018, and June 2018 administrations of the Regents Examination in Chemistry are located at http://www.nysedregents.org/Chemistry

Attention to the rubric design also fundamentally contributes to the validity of examinee response processes. The rubric specifies what the examinee needs to provide as evidence of learning based on the question asked. The more explicit the rubric (and the item), the more clear the response expectations are for examinees. To facilitate the development of constructed-response scoring rubrics, NYSED training for writing items includes specific attention to rubric development as follows:

- The rubric should clearly specify the criteria for awarding each credit.
- The rubric should be aligned to what is asked for in the item and correspond to the knowledge or skill being assessed.
- Whenever possible, the rubric should be written to allow for alternative approaches and other legitimate methods.

In support of the goal of valid score interpretations for each examinee, then, such scoring

test reliability
classification

IRT Model Fit

testing requirement toward graduation for students who have completed a course in Chemistry, the exam is most commonly used to inform admissions and course placement decisions. Such uses can be considered reasonable, assuming that the competencies demonstrated in the Regents Examination in Chemistry are consistent with those required in the courses for which a student is seeking enrollment or placement. Educational institutions (d aB3 (((ng)10 (t)2 (h-10

abengp(r)17

tocdsant th-10 nats

References

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, 37(*2), 163–178.

Cook, D. A., & Beckman, T. J. (2009). Does scale length matter? A comparison of nine- versus five-point rating scales for mini-CEX. *Advances in Health Sciences Education, 14,* 655–684.

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

Cronbach, L. J. (1988). Five Perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 3–17) Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., Linn, R. L., Brennan, R. T., & Haertel, E. (1995, Summer). Generalizability analysis for educational assessments. Los Angeles, CA: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology, 94*(5), 1336–1344.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hambleton, R. K., & Novak, M. R. (1973). Toward an integration of theory and methods for criterion-referenced tests. *Journal of Educational Measurement*, *10*, 159–170.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Item response theory*. Newbury Park, CA: Sage Publications.

Hanson, B. A. (1994). Extension of Lord-Wingersky algorithm to computing test scores for polytomous items. Retrieved February 17, 2016 from http://www.b-a-

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practices*, 10(2), 33–41.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 32, 179–185.

Hoyt, W. T., & Kerns, M. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4,* 403–424.

Huynh, H. (1976). Statistical consideration of mastery scores. *Psychometrika*, *41*, 65–78.

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance.* New York, NY: The Guilford Press.

Kolen, M. J. (2004). POLYCSEM [Computer program]. University of Iowa. Retrieved August 1, 2012, from https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs.

Kolen, M. J., & Brennan, R. L. (2005). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

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 detecting differential accuracy and differential scale use. *Journal of Educational Measurement, 46*(4), 371–389.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Authors.

Patz, R. J., Junk002 Tw 1.5w 0.56 ,ryG [.038 T(s)4 (.)]TJ 0 Tc 0 Tw 12.11 0 Td ( )Tj EMC JuC8t1c

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 2017 Administration

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
37	MC	1	1	4	3	3.1	0.49	0.50	0.2469	0.98
38	MC	1	1	4	3	3.1	0.65	0.42	-0.5058	1.05
39	MC	1	1	4	3					

Table A.2 Test Map for January 2018 Administration

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	

Table A.3 Test Map for June 2018 Administration

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
1	MC	1	1	4	3	3.1	0.84	0.43	-1.7352	0.92

Position

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

Table B.1 Score Table for August 2017 Administration

Raw Score	Ability	Scale Score	Raw Score	Ability	Scale Score	Raw Score	Ability	Scale Score
0	-6.0539	0.000	41	-0.1216	58.588	82	3.7048	94.804
1	-4.8353	3.344	42	-0.0655	59.288	83	4.1379	96.415

Table B.2 Score Table for January 2018 Administration

Raw Score	Ability	Scale Score

Raw Score	Ability	Scale Score

Raw Score	Ability	Scale Score

Table B.3 Score Table for June 2018 Administration

Raw Score	Ability	Scale Score
0	-6.1700	0.000
1	-4.9478	2.980

Raw Score	Ability	Scale Score
41	-0.1047	58.800
42	-0.0473	59.509

Raw Score	Ability	Scale Score
82	3.6829	94.705
83	4.1111	96.324

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.	Does the item have one and only one correct answer?		
5.	Does the item state one single central problem completely in the stem? (See Helpful Hint below.)		
6.	Does the stem include any extraneous material (“window dressing”)?		
7.	Are all responses grammatically consistent with the stem and parallel with one another in form?		
8.	Are all responses plausible (attractive to students who lack the information tested by the item)?		
9.	Are all responses independent and mutually exclusive?		
10.	Are there any extraneous clues due to grammatical inconsistencies, verbal associations, length of response, etc.?		
11.	Were the principles of Universal Design used in constructing the item?		

HELPFUL HINT

To determine if the stem is complete (meaningful all by itself):

1. Cbemby(t)2 (es)4 (pons)14nielniesad tjsitthe stepl 1. Cbeh h0 Tc 0Tc -0.002 Tw 1.8 0[h(t)2 (t)2

Appendix D: Tables and Figures for August 2017 Administration

Table D.1 Multiple-Choice Item Analysis Summary: Cho As0 Tc 0 Tw ( )Tj 21.4 ET /H1 <fact <</A

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

Item	Min. score	Max. score	Number of Students	Mean	SD	*p*-Value	Point-Biserial
51	0	1	7,281	0.40	0.49	0.40	0.31
52	0	1	7,281	0.51	0.50	0.51	0.43
53	0	1	7,281	0.53	0.50	0.53	0.44
54	0	1	7,281	0.43	0.49	0.43	0.33
55	0	1	7,281	0.66	0.47	0.66	0.25
56	0	1	7,281	0.62	0.49	0.62	0.45
57	0	1	7,281	0.28	0.45	0.28	0.50
58	0	1	7,281	0.31	0.46	0.31	0.46
59	0	1	7,281	0.45	0.50	0.45	0.47
60	0	1	7,281	0.56	0.50	0.56	0.49
61	0	1	7,281	0.77			

70 9.96 252.04.12 6 Tm ( )Tj ET EMC

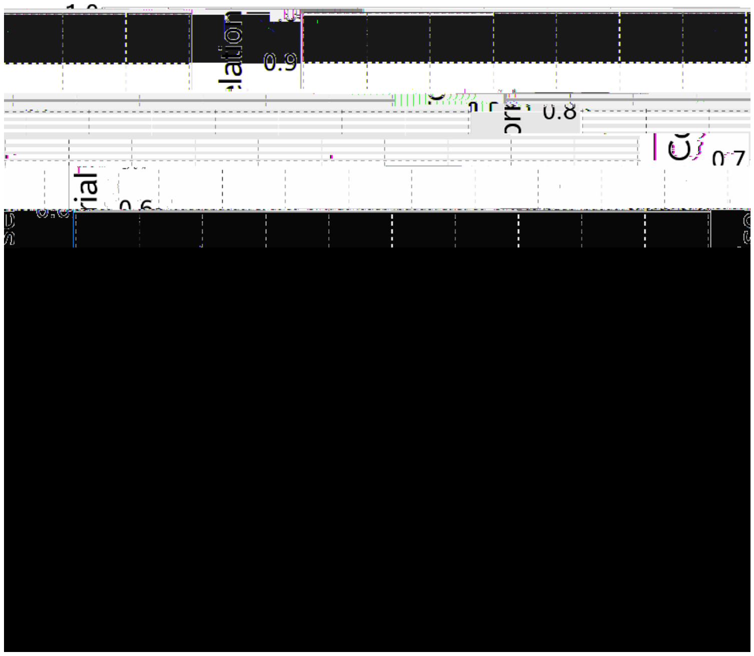


Figure D.1 Scatter Plot: Regents Examination in Chemistry

Table D.3 Descriptive Statistics in *p*-value and Point-Biserial Correlation: Regents Examination in Chemistry

Statistics	N	Mean	Min	Q1	Median	Q3	Max
p-value	85	0.57	0.20	0.43	0.60	0.70	0.85
Point-Biserial	85	0.37	0.17	0.31	0.37	0.44	0.51

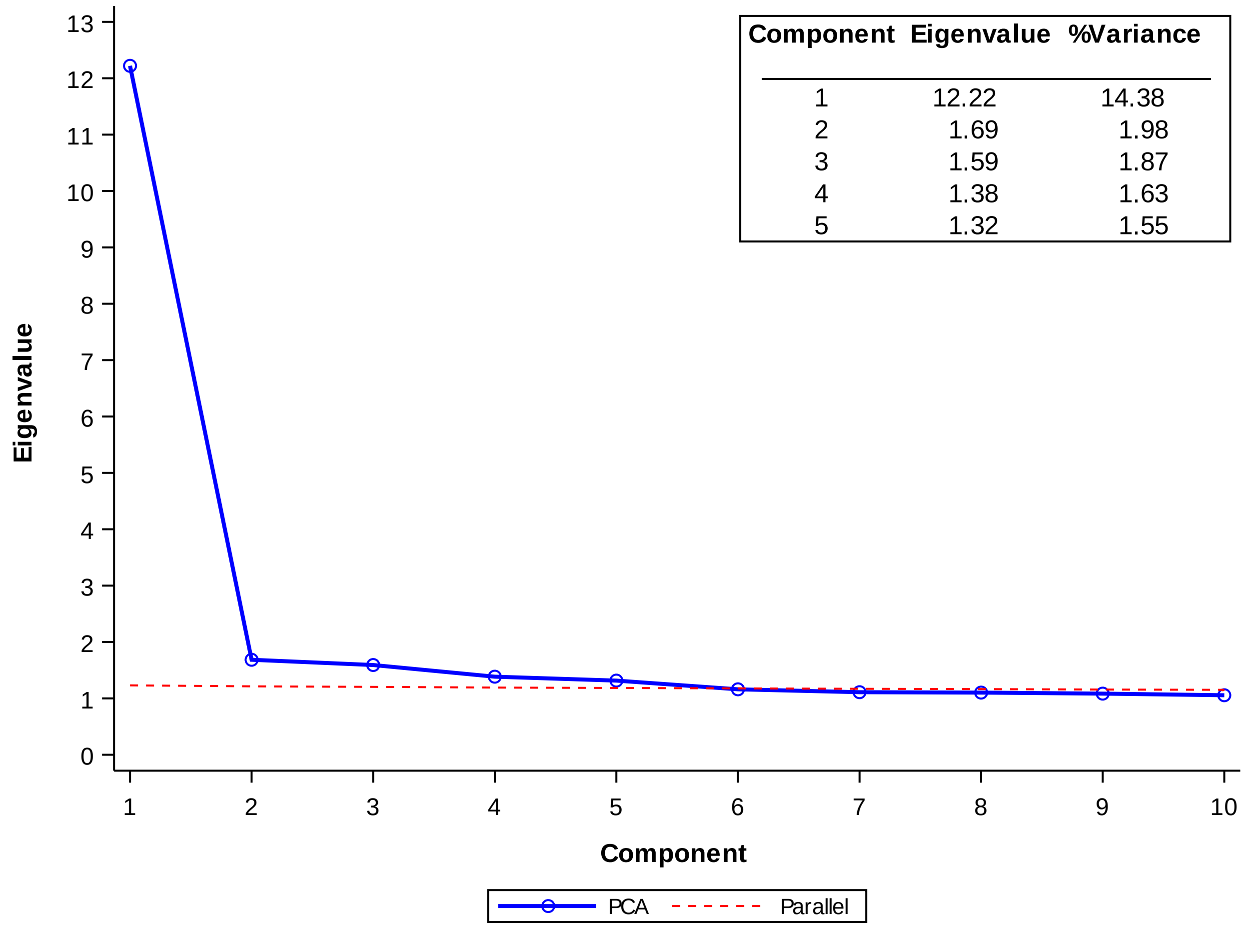


Figure D.3 Scree Plot: Regents Examination in Chemistry

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

Statistic Type	Value
N	3,570
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	0.13
>|0.20|	0

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

	INFIT Mean Square			
	N	Mean	SD	Min

Table D.8 Group Means: Regents Examination in Chemistry

Demographics	Number	Mean Scale Score	SD Scale Score
All Students*	7,281	63.68	11.75
Ethnicity			
American Indian/Alaska Native	50	61.48	9.73
Asian/Native Hawaiian/Other Pacific Islander	873	66.39	13.37
Black/African American	1,174	60.13	10.25
Hispanic/Latino	1,523	59.51	11.55
Multiracial	113	64.75	12.21
White	3,547	65.99	11.06
English Language Learner/Multilingual Learner			
No	7,226	63.79	11.65
Yes	55	49.40	15.77
Economically Disadvantaged			
No	4,295	65.70	11.47
Yes	2,986	60.79	11.53
Gender			
Female			

Item	Number of Students	*p*-Value	SD	Point-Biserial	Point-Biserial Distractor 1	Point-Biserial Distractor 2	Point-Biserial Distractor 3
35	3,210	0.66	0.47	0.45	-0.19	-0.34	-0.13
36	3,210	0.56	0.50	0.42	-0.17	-0.23	-0.19
37	3,210	0.54	0.50	0.30	-0.21	-0.09	-0.11
38	3,210	0.48	0.50	0.17	-0.09	-0.05	-0.09
39	3,210	0.78	0.41	0.35	-0.24	-0.18	-0.13
40	3,210	0.79	0.41	0.33	-0.15	-0.20	-0.19
41	3,210	0.89	0.32	0.31	-0.11	-0.20	-0.20
42	3,210	0.55	0.50	0.31	-0.13	-0.20	-0.12
43	3,210	0.74	0.44	0.49	-0.33	-0.19	-0.23
44	3,210	0.16	0.37	0.25	-0.13	0.00	-0.11
45	3,210	0.59	0.49	0.39	-0.20	-0.22	-0.15
46	3,210	0.46	0.50	0.17	-0.04	-0.10	-0.17
47	3,210	0.46	0.50	0.33	-0.22	-0.10	-0.14
48	3,210	0.41	0.49	0.28	-0.22	-0.01	-0.15
49	3,210	0.54	0.50	0.31	-0.14	-0.12	-0.21
50	3,210	0.60	0.49	0.24	-0.06	-0.19	-0.10

Table E.2 Constructed

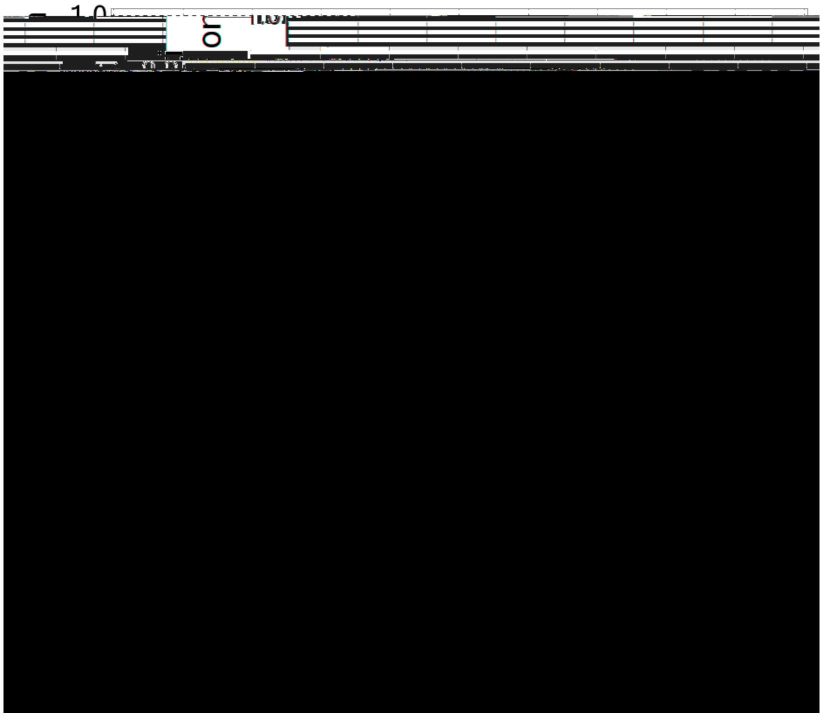
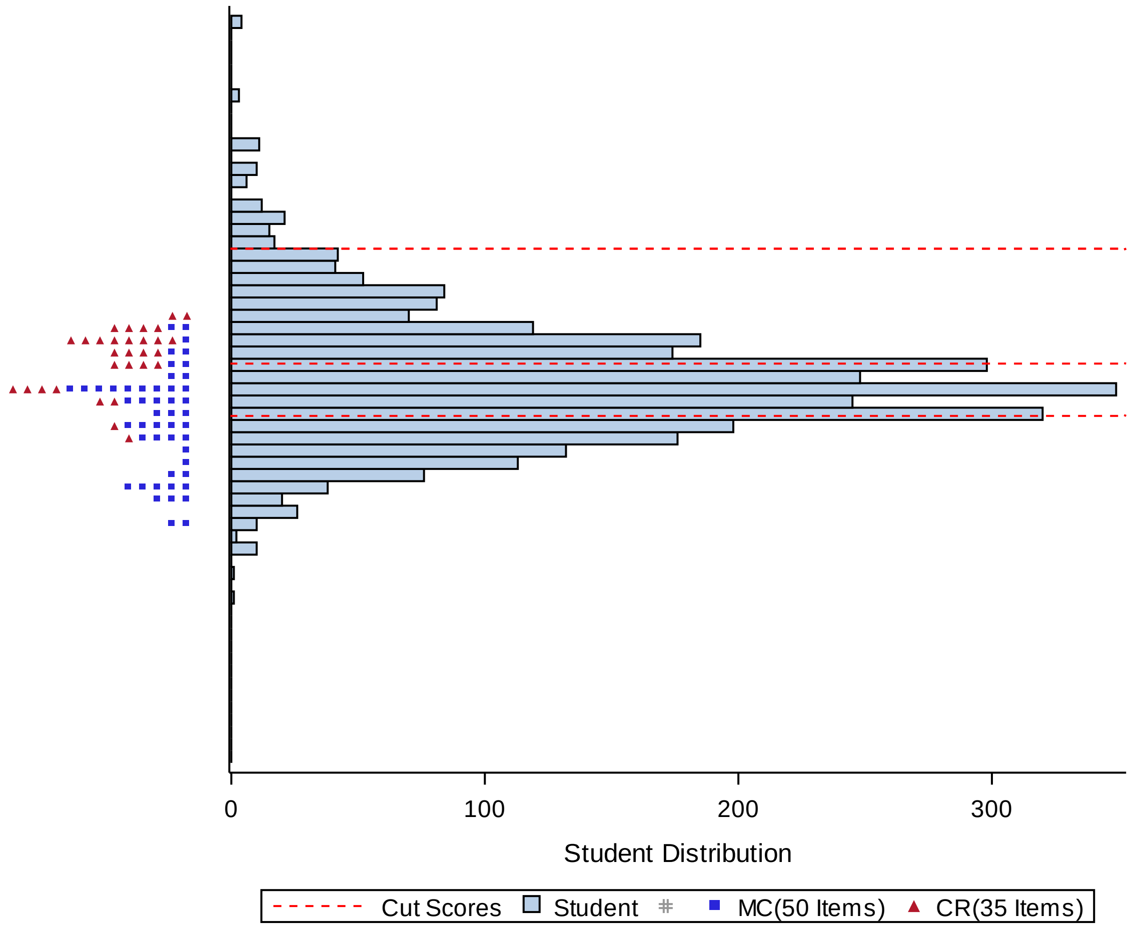


Figure E.1 Scatter Plot: Regents Examination in Chemistry

Table E.3 Descriptive Statistics in *p*-value and Point-Biserial Correlation: Regents Examination in Chemistry

Statistics	N	Mean	Min	Q1	Median	


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

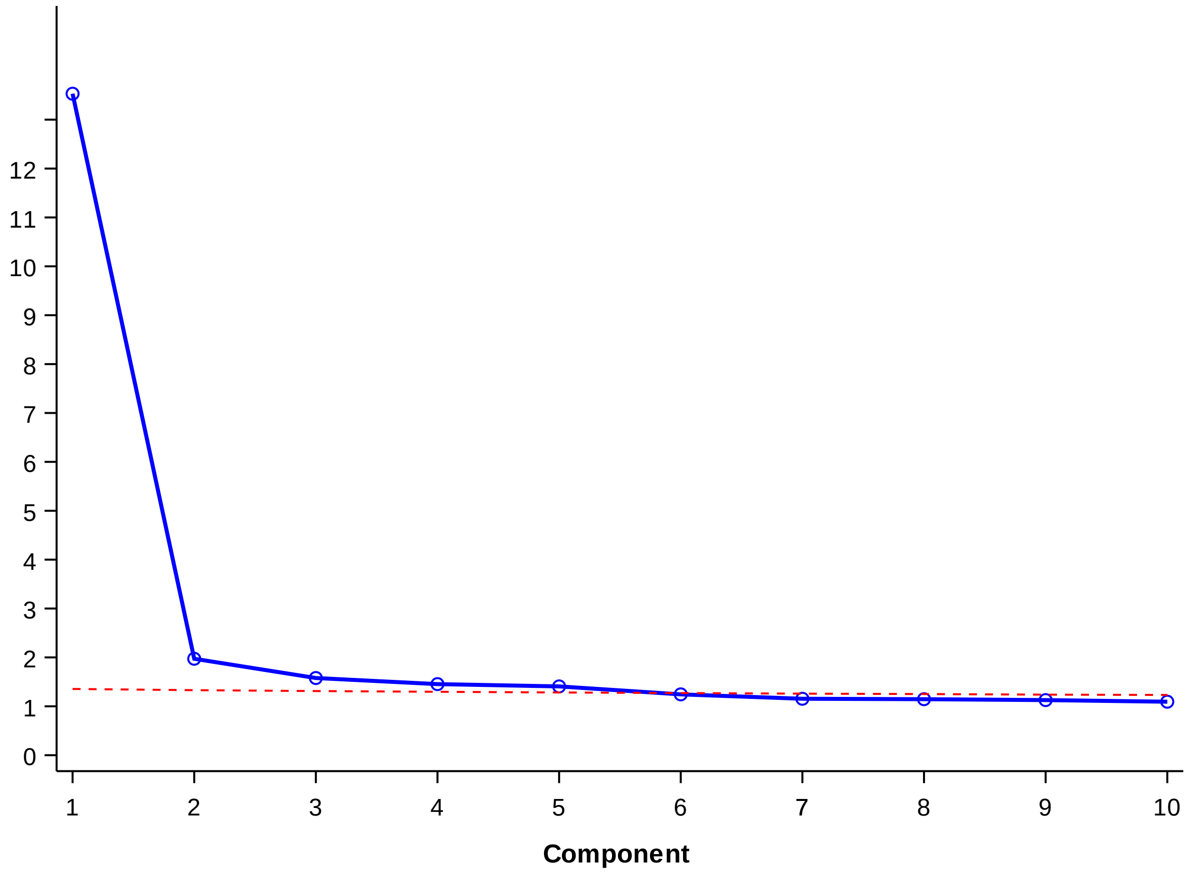

12
11
10
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
Component

Table E.5 Summary of