New York State Regents Examination in Algebra I

2022 Technical Report

Prepared for the New York State Education Department
by Pearson

June 2024

Copyright

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

Copyright © 2024 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

CHAPTER 1: INTRODUCTION 5

1.1. Introduction 5
1.2. Purposes of the Exam 5
1.3. Target Population (Standard 7.2) 6

CHAPTER 2: CLASSICAL ITEM STATISTICS (STANDARD 4.10) 7

2.1. Item Difficulty 7
2.2. Item Discrimination 7
2.3. Discrimination on Difficulty ScatterPlot 9
2.4. Observations and Interpretations 10

CHAPTER 3: IRT CALIBRATIONS, EQUATING, AND SCALING (STANDARDS 2 AND 4.10) 11

3.1. Description of the Rasch Model 11
3.2. Software and Estimation Algorithm 12
3.3. Item Difficulty–Student Performance Map 12
3.4. Checking Rasch Assumptions 12
3.4.1. Unidimensionality 13
3.4.2. Local Independence 14
3.4.3. Item Fit 16
3.5. Scaling of Operational Test Forms 16

CHAPTER 4: RELIABILITY (STANDARD 2) 19

4.1. Reliability Indices (Standard 2.20) 19
4.2. Standard Error of Measurement (Standards 2.13, 2.14, 2.15) 20
4.2.1. Traditional Standard Error of Measurement 20
4.2.2. Traditional Standard Error of Measurement Confidence Intervals 20
4.2.3. Conditional Standard Error of Measurement 21
4.2.4. Conditional Standard Error of Measurement Confidence Intervals 22
4.2.5. Conditional Standard Error of Measurement Characteristics 22
4.2.6. Results and Observations 22
4.3. Decision Consistency and Accuracy (Standard 2.16) 23
4.4. Group Means (Standard 2.17) 25
4.5. State Percentile Rankings 25

CHAPTER 5: VALIDITY (STANDARD 1) 27

5.1. Evidence Based on Test Content 27
5.1.1. Content Validity 28
5.1.2. Item Development Process 28
5.1.3. Item Review Process 29
5.2. Evidence Based on Response Processes 30
5.2.1. Administration and Scoring 30
5.2.2. Statistical Analysis 32
5.3. Evidence Based on Internal Structure 32
5.3.1. Item Difficulty 33
5.3.2. Item Discrimination 33
5.3.3. Differential Item Functioning 33
5.3.4. IRT Model Fit 33
5.3.5. Test Reliability 34
5.3.6. Classification Consistency and Accuracy 34
5.3.7. Test Dimensionality 34
5.4. Evidence Based on Relations to Other Variables 34

5.5. Evidence Based on Testing Consequences 35
5.6. Summary 35
REFERENCES 36
APPENDIX A: OPERATIONAL TEST MAP 40
APPENDIX B: RAW-TO-THETA-TO-SCALE SCORE CONVERSION TABLE 41
APPENDIX C: ITEM WRITING GUIDELINES 42

List of Tables

Table 1.1. Total Student Population in June 2022: Regents Examination in Algebra I 6
Table 2.1. MC Item Analysis Summary: Regents Examination in Algebra I 8
Table 2.2. CR Item Analysis Summary: Regents Examination in Algebra I 9
Table 2.3. Descriptive Statistics in *p*-value and Point-Biserial Correlation: Regents Examination in Algebra I 10
Table 3.1. Summary of Item Residual Correlations: Regents Examination in Algebra I 15
Table 3.2. Summary of INFIT Mean Square Statistics: Regents Examination in Algebra I 16
Table 4.1. Coefficient Alpha and SEM: Regents Examination in Algebra I 21
Table 4.2. Pseudo-Decision Table for Two Hypothetical Categories 24
Table 4.3. Pseudo-Decision Table for Four Hypothetical Categories 24
Table 4.4. Decision Consistency and Accuracy Results: Regents Examination in Algebra I 25
Table 4.5. Group Means: Regents Examination in Algebra I 25
Table 4.6. State Percentile Ranking for Scale Score: Regents Examination in Algebra I 26
Table 5.1. Test Blueprint: Regents Examination in Algebra I 28
Table A.1. June 2022 Test Map 40
Table B.1. June 2022 Score Table 41

List of Figures

Figure 2.1. Scatterplot: Regents Examination in Algebra I 9
Figure 3.1. Student Performance Map: Regents Examination in Algebra I 12
Figure 3.2. Scree Plot: Regents Examination in Algebra I 14
Figure 4.1. Conditional Standard Error Plot: Regents Examination in Algebra I 23
Figure 5.1. NYSED Test Development Process 29

Chapter 1: Introduction

1.1. INTRODUCTION

This technical report for the Regents Examination in Algebra I, based on the New York State (NYS) Learning Standards, provides New York State with documentation of the purposes of the Regents Examination, scoring information, evidence of both the reliability and validity of the exams, scaling information, and guidelines for score reporting for the June 2022 administration. As the *Standards for Education and Psychological Testing* discusses in Standard 7, "The objective of the documentation is to provide test users with the information needed to help them assess the nature and quality of the test, the resulting scores, and the interpretations based on the test scores" (AERA et al., 2014, p. 123).[1] **Please note:** A technical report, by design, addresses the technical documentation of a testing program. Other aspects of a testing program (e.g., content standards, scoring guides, the guide to test interpretation) are thoroughly addressed and referenced in supporting documents. All analyses in this report were conducted using the operational items only.

Typically, the Regents Examinations are administered each year in August, January, and June to students enrolled in New York State schools. However, testing in 2020 and 2021 was disrupted by the COVID-19 pandemic, starting with the cancellation of the June 2020 administration. In 2021, the New York State Education Department (NYSED) cancelled all but four of the Regents Examinations scheduled to be administered in June 2021, and all the Regents Examinations scheduled to be administered in August 2021. The four examinations administered in June 2021 were Algebra 1, Earth Science, English Language Arts and Living Environment, as required to comply with the federal Every Student Succeeds Act (ESSA). NYSED also cancelled the January 2022 administration of the Regents Examination Program in response to the ongoing impact of COVID-19, as described online at https://www.nysed.gov/news/2022/january-2022-regents-examinations-cancelled-due-ongoing-pandemic. Regular testing activities resumed in June 2022.

1.2. PURPOSES OF THE EXAM

The Regents Examination in Algebra I measures student achievement against the NYS P–12 Learning Standards. The exam is prepared by teacher examination committees and New York State Education Department (NYSED) subject matter and testing specialists. Further, 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 to guide classroom teaching and learning. The exam also provides 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 Algebra I is intended for use in satisfying state testing requirements for students who have finished a course in Algebra I. A passing score on the exam counts toward requirements for a high school diploma, as described in the New York State diploma requirements.[2] Results of the Regents Examinations may also be used to satisfy various locally established requirements throughout the state.

[1] References to specific *Standards* will be placed in parentheses throughout the technical report to provide further context for each section.

[2] The New York State diploma requirements are located online at https://www.nysed.gov/common/nysed/files/programs/curriculum-instruction/currentdiplomarequirements2.pdf.

1.3. TARGET POPULATION (STANDARD 7.2)

The student population for the Regents Examination in Algebra I is composed of students who have completed a course in Algebra I. Table 1.1 presents a demographic breakdown of all students who took the June 2022 Regents Examination in Algebra I. All analyses in this report are based on the population described in this table. Annual Regents Examination results reported in the New York State Report Cards are those reported in the Student Information Repository System (SIRS) as of the reporting deadline. The results include the exams administered in June 2022 (see https://data.nysed.gov/). If a student takes the same exam multiple times in the year, only the highest score is included in these results. However, as previously mentioned, this exam was only administered in June 2022 during the 2021–2022 school year. Item-level data used for the analyses in this report are reported by districts on a similar timeline but through a different collection system. These data include all student results for each administration. Therefore, the sample sizes (n-counts) in this technical report will differ from publicly reported counts of student test takers.

Table 1.1. Total Student Population in June 2022: Regents Examination in Algebra I

Demographics	N	%
All Students	211,056	100.00
Race/Ethnicity		
American Indian/Alaska Native	1,429	0.68
Asian/Native Hawaiian/Other Pacific Islander	22,144	10.49
Black/African American	34,168	16.19
Hispanic/Latino	61,361	29.08
Multiracial	5,567	2.64
White	86,332	40.92
English Language Learner		
No	192,798	91.35
Yes	18,258	8.65
Economically Disadvantaged		
No	101,526	48.10
Yes	109,530	51.90
Gender		
Female	103,662	49.13
Male	107,269	50.84
Nonbinary	70	0.03
Student with a Disability		
No	178,996	84.81
Yes	32,060	15.19

Note. 55 students were not reported in the race/ethnicity and gender groups, but they are 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 to the operational Regents Examination in Algebra I 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), calculated as follows:

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

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 answer, 1 = correct answer). 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. For example, if an MC item has a *p*-value of 0.89, it means 89% of the students answered the item correctly. 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 0) to the maximum possible score (e.g., 6 points for some mathematics items). 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 exam. Items with very high or very low *p*-values receive added scrutiny during all follow-up analyses, including item response theory (IRT) 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 2022 Regents Examination in Algebra I show a range of *p*-values consistent with the targeted exam difficulty. Item *p*-values, presented in Table 2.1 and Table 2.2 for MC and CR items, respectively, range from 0.20 to 0.77, with a mean of 0.48. The tables also show a standard deviation (SD) of item score and item mean (Table 2.2 only).

2.2. ITEM DISCRIMINATION

At the most general level, estimates of item discrimination indicate an item's ability to differentiate between high and low performance on an exam. It is expected that students who perform well on the exam would be more likely to answer any given item correctly, while low-performing students (i.e., those who perform poorly on the exam overall) would be more likely to answer the same item incorrectly. Pearson's product-moment correlation coefficient (also commonly referred to as a point-biserial correlation) between item scores and test scores is used to indicate discrimination (Pearson, 1896). The correlation coefficient can range from −1.0 to +1.0. If high-scoring students

tend to get the item correct while low-scoring students do not, the correlation between the item score and the total test score will be both positive and noticeably large in its magnitude (i.e., above 0), indicating that the item is likely discriminating well between high- and low-performing students. Point-biserial correlations are computed for each answer option, including correct and incorrect options (commonly referred to as distractors). Point-biserial values for each distractor are typically negative and an important part of the analysis. Positive values can indicate that higher-performing students are selecting an incorrect answer or that the item key for the correct answer should be checked.

Table 2.1 and Table 2.2 present the point-biserial values on the correct response and three distractors (Table 2.1 only) for the June 2022 Regents Examination in Algebra I. The values for correct answers are 0.18 or higher for all items with all but three items exceeding 0.35, indicating that the items are generally discriminating well between high- and low-performing students. Point-biserial values for all distractors are negative, indicating that students are generally responding to the items as expected during item and rubric development.

Table 2.1. MC Item Analysis Summary: Regents Examination in Algebra I

Item	#Students	*p*-Value	SD	Point-Biserial	Point-Biserial Distractor 1	Point-Biserial Distractor 2	Point-Biserial Distractor 3
1	211,056	0.77	0.42	0.18	-0.14	-0.14	-0.01
2	211,056	0.76	0.43	0.48	-0.20	-0.27	-0.28
3	211,056	0.69	0.46	0.55	-0.47	-0.19	-0.11
4	211,056	0.73	0.44	0.50	-0.23	-0.26	-0.28
5	211,056	0.75	0.43	0.39	-0.16	-0.19	-0.26
6	211,056	0.65	0.48	0.42	-0.14	-0.29	-0.22
7	211,056	0.65	0.48	0.49	-0.25	-0.26	-0.22
8	211,056	0.68	0.47	0.54	-0.38	-0.22	-0.19
9	211,056	0.61	0.49	0.34	-0.17	-0.16	-0.18
10	211,056	0.65	0.48	0.43	-0.24	-0.18	-0.24
11	211,056	0.57	0.50	0.54	-0.20	-0.29	-0.28
12	211,056	0.53	0.50	0.58	-0.26	-0.31	-0.21
13	211,056	0.53	0.50	0.40	-0.25	-0.15	-0.17
14	211,056	0.42	0.49	0.42	-0.15	-0.22	-0.20
15	211,056	0.52	0.50	0.48	-0.12	-0.38	-0.12
16	211,056	0.51	0.50	0.46	-0.27	-0.17	-0.19
17	211,056	0.54	0.50	0.47	-0.18	-0.30	-0.15
18	211,056	0.44	0.50	0.52	-0.17	-0.18	-0.31
19	211,056	0.49	0.50	0.43	-0.15	-0.27	-0.15
20	211,056	0.47	0.50	0.51	-0.22	-0.25	-0.20
21	211,056	0.40	0.49	0.45	-0.21	-0.27	-0.08
22	211,056	0.34	0.47	0.28	-0.09	-0.05	-0.18
23	211,056	0.29	0.45	0.38	-0.20	-0.08	-0.13
24	211,056	0.41	0.49	0.45	-0.13	-0.20	-0.23

Table 2.2. CR Item Analysis Summary: Regents Examination in Algebra I

Item	#Students	Min. Score	Max. Score	Mean	SD	*p*-Value	Point-Biserial
25	211,056	0	2	0.87	0.89	0.43	0.65
26	211,056	0	2	0.82	0.92	0.41	0.74
27	211,056	0	2	0.69	0.86	0.35	0.65
28	211,056	0	2	0.65	0.81	0.32	0.67
29	211,056	0	2	0.40	0.72	0.20	0.67
30	211,056	0	2	0.54	0.82	0.27	0.71
31	211,056	0	2	0.46	0.68	0.23	0.67
32	211,056	0	2	0.41	0.72	0.20	0.69
33	211,056	0	4	2.01	1.29	0.50	0.73
34	211,056	0	4	1.55	1.45	0.39	0.74
35	211,056	0	4	1.27	1.53	0.32	0.77
36	211,056	0	4	1.28	1.49	0.32	0.79
37	211,056	0	6	2.49	2.41	0.41	0.82

2.3. DISCRIMINATION ON DIFFICULTY SCATTERPLOT

Figure 2.1 presents a scatterplot of the item discrimination values (*y*-axis) and item difficulty values (*x*-axis). Table 2.3 presents the descriptive statistics of *p*-value and point-biserial values, including mean, minimum, Q1, median, Q3, and maximum.

Figure 2.1. Scatterplot: Regents Examination in Algebra I

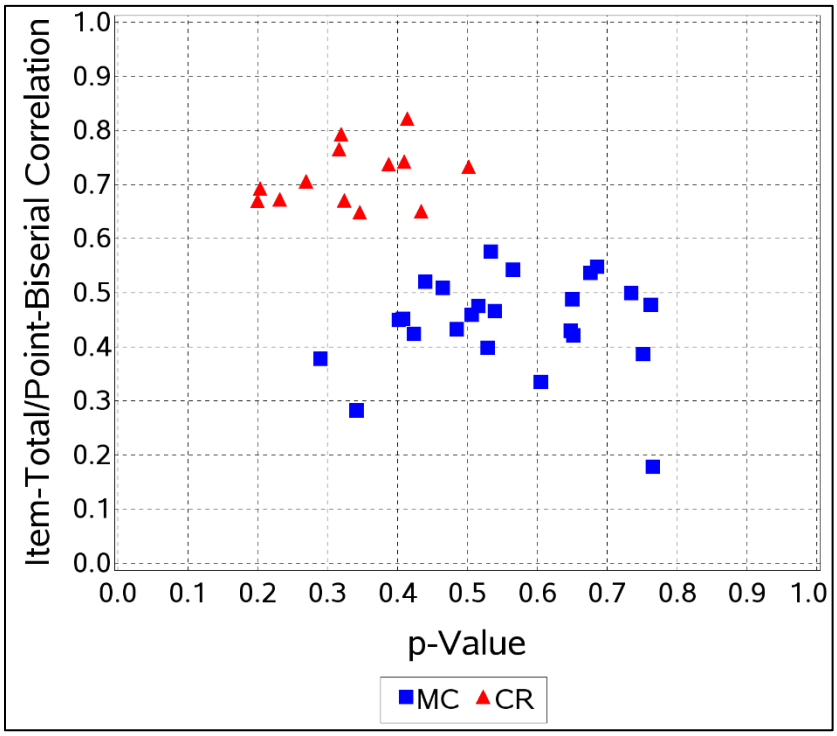


Table 2.3. Descriptive Statistics in *p*-value and Point-Biserial Correlation: Regents Examination in Algebra I

Statistic	N	Mean	Min.	Q1	Median	Q3	Max.
p-value	37	0.48	0.20	0.35	0.47	0.61	0.77
Point-Biserial	37	0.54	0.18	0.43	0.51	0.67	0.82

2.4. OBSERVATIONS AND INTERPRETATIONS

The *p*-values for the MC items ranged from about 0.29 to 0.77, while the *p*-values for the CR items ranged from about 0.20 to 0.50. The difficulty distributions illustrated in the plot indicate that a wide range of item difficulties appeared on each exam, which was one test development goal.

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

The item response theory (IRT) model used for the Regents Examinations 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 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 et al., 1991). This performance value is often referred to as θ. Performance and θ will be used interchangeably throughout the remainder of this report.

A fundamental advantage of IRT is that it links student performance and item difficulty estimates and places them on the same scale, allowing for an evaluation of student 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 students across items and test forms.

This chapter outlines the procedures used for calibrating the operational items, including an introduction to the Rasch model, the results from evaluations of the adequacy of the Rasch assumptions, and the Rasch item statistics. Generally, item calibration is the process of assigning a difficulty, or item "location," estimate to each item on an assessment so that all items are placed onto a common scale.

3.1. DESCRIPTION OF THE RASCH MODEL

The Rasch model (Rasch, 1960) was used to calibrate the MC items, and the partial credit model (PCM; Wright & Masters, 1982) was used to calibrate the CR items. The PCM extends the Rasch model for dichotomous (0, 1) items so that it accommodates the polytomous CR item data. Under the PCM model, for a given item *i* with m_i score categories, the probability of person *n* scoring *x* (x = 0, 1, 2, ... m_i) is given as follows:

$$P_{ni}(X = x) = \frac{\exp \sum_{j=0}^{x} (\theta_n - D_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\theta_n - D_{ij})}$$

where θ_n represents student ability, and D_{ij} is the step difficulty of the j^{th} step on item *i*. D_{ij} can be expressed as $D_{ij} = D_i - F_{ij}$, where D_i is the difficulty for item *i* and F_{ij} is a step deviation value for the j^{th} step. For dichotomous MC items, the PCM reduces to the standard Rasch model, and the single step difficulty is referred to as the item's difficulty. The Rasch model predicts the probability of person *n* getting item *i* correct, as follows:

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

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 student performance and item difficulty that are theoretically invariant across random samples of the same student population.

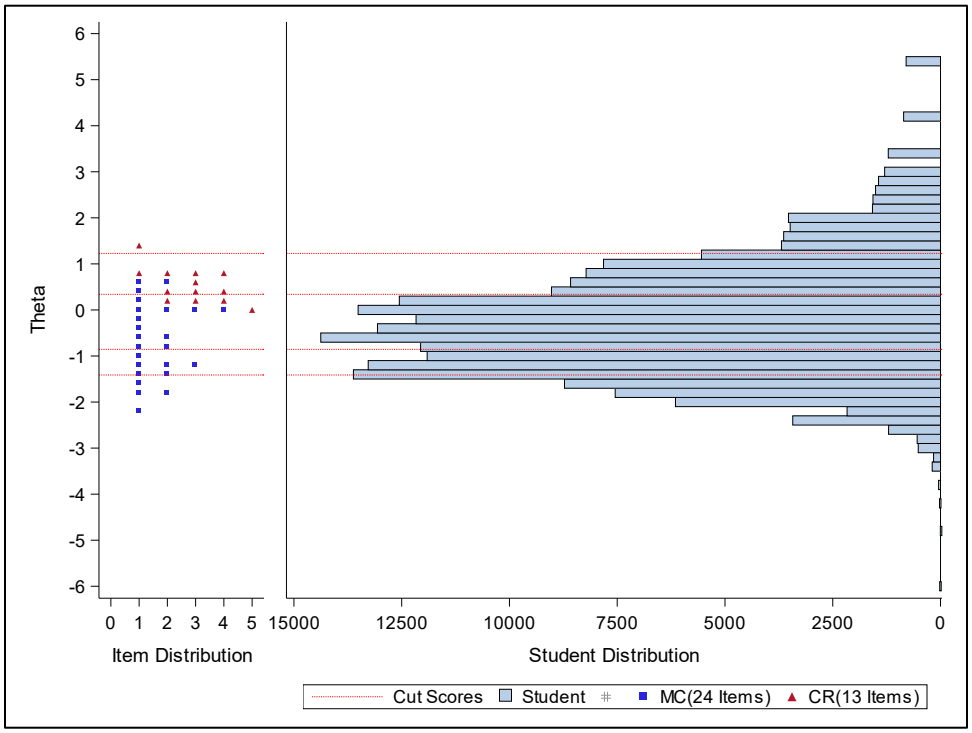
3.2. SOFTWARE AND ESTIMATION ALGORITHM

Item calibration was implemented via the WINSTEPS 3.60 computer program (Linacre, 2005), which employs unconditional (UCON) joint maximum likelihood estimation (JMLE).

3.3. ITEM DIFFICULTY–STUDENT PERFORMANCE MAP

The distributions of the Rasch item logits (item difficulty estimates) and student performance are shown on the item difficulty–student performance map presented in Figure 3.1. This graphic illustrates the location of student performance and item difficulty on the same scale, along with their respective distributions and cut scores (indicated by the horizontal dotted lines). The figure shows more difficult items and higher student performance at the top and lower performance and easier items at the bottom.

Figure 3.1. Student Performance Map: Regents Examination in Algebra I



3.4. CHECKING RASCH ASSUMPTIONS

Because the Rasch model was the basis of all calibration, scoring, and scaling analyses associated with the Regents Examination in Algebra I, 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. Only operational items were analyzed as they are the basis of student scores.

3.4.1. Unidimensionality

Rasch models assume that one dominant dimension determines the differences in students' performances. Principal components analysis (PCA) can be used to assess the unidimensionality assumption. The purpose of the analysis is to verify whether any other dominant components exist among the items. If any other dimensions are found, the unidimensionality assumption would be violated.

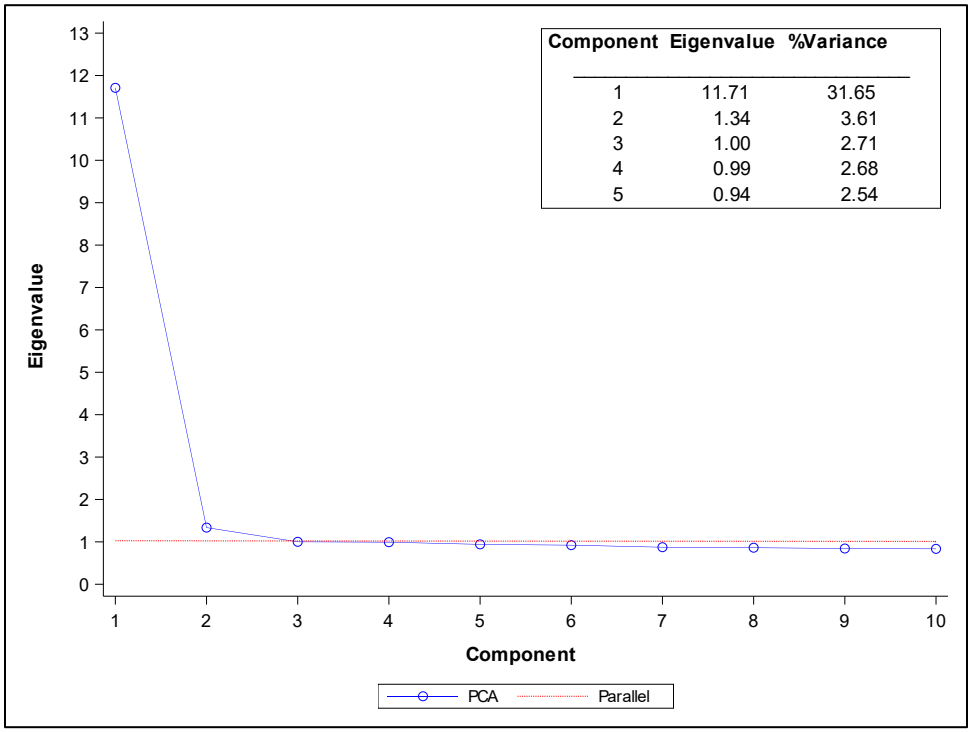
A parallel analysis (Horn, 1965) was conducted to help distinguish components that are real from components that are random. Parallel analysis is a technique used to determine how many factors exist in principal components. For the parallel analysis, 100 random datasets of sizes equal to the original data were created. For each random dataset, a PCA was performed and the resulting eigenvalues stored. For each component, the upper 95th percentile value of the distribution of the 100 eigenvalues from the random data sets was then plotted. Given the size of the data generated for the parallel analysis, the reference line is essentially equivalent to plotting a reference line for an eigenvalue of 1.

Figure 3.2 presents the PCA and parallel analysis results for the Regents Examination in Algebra I. The results include the eigenvalues and the percentage of variance explained for the first five components, as well as the scree plots. The scree plots show the eigenvalues plotted by component number and the results of a parallel analysis. Although the total number of components in the PCA is the same as the total number of items in a test, Figure 3.2 shows only the first 10 components. This view is sufficient for interpretation because components are listed in descending eigenvalue order. The fact that the eigenvalues for components 2–10 are much lower than the first component demonstrates that there is one dominant component, showing evidence of unidimensionality.

Reckase (1979) proposed that the variance explained by the primary dimension should be greater than 20% to indicate unidimensionality. However, as this rule is not absolute, it is helpful to consider three additional characteristics of the PCA and parallel analysis results: (1) whether the ratio of the first to the second eigenvalue is greater than 3, (2) whether the second value is not much larger than the third value, and (3) whether the second value is not significantly different from those from the parallel analysis.

As shown in Figure 3.2, the primary dimension explained 31.65% of the total variance for the Regents Examination in Algebra I. The eigenvalue of the second dimension is less than one-third of the first at 1.34, and the second value is not significantly different from the parallel analysis. Overall, the PCA suggests that the test is reasonably unidimensional.

Figure 3.2. Scree Plot: Regents Examination in Algebra I



3.4.2. Local Independence

Local independence is a fundamental assumption of IRT. This means that, for statistical purposes, a student's response to any one item should not depend on the student's response to any other item on the test. In formal statistical terms, Test *X*, which comprises items X_1, X_2, …X_n, is locally independent with respect to the latent variable θ if, for all $x = (x_1, x_2, \ldots x_n)$ and θ:

$$P(\mathbf{X} = \mathbf{x} \mid \theta) = \prod_{i=1}^{I} P(X_i = x_i \mid \theta)$$

This formula essentially states that the probability of any pattern of responses across all items (**x**), after conditioning on the student's true score (θ) 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 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 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:

$$P(X_i = x_i, X_j = x_j \mid \theta) = P(X_i = x_i \mid \theta) P(X_j = x_j \mid \theta)$$

Marais and Andrich (2008) point out that local item dependence in the Rasch model can occur in two ways that may be difficult to distinguish. The first way occurs when the assumption of unidimensionality is violated. Here, other nuisance dimensions besides a dominant dimension determine student performance (this can be called "trait dependence"). The second way occurs when responses to an item depend on responses to another item. This is a violation of statistical independence and can be called "response dependence." By distinguishing the two sources of local dependence, one can see that, while local independence can be related to unidimensionality, the two are different assumptions and therefore require different tests.

Residual item correlations, provided in WINSTEPS for each item pair, were used to assess the local dependence between the Regents Examination in Algebra I items. In general, these residuals are computed as follows. First, expected item performance based on the Rasch model is determined using θ and item parameter estimates. Next, deviations (residuals) between the students' expected and observed performance are determined for each item. Finally, for each item pair, a correlation between the respective deviations is computed.

Three types of residual correlations are available in WINSTEPS: raw, standardized, and logit. The raw score residual correlation essentially corresponds to Yen's Q3 index, a popular statistic used to assess local independence. The expected value for the Q3 statistic is approximately -1/(k – 1) when no local dependence exists, where k is test length (Yen, 1993). Thus, the expected Q3 values should be approximately -0.03 for the items on the exam. Index values greater than 0.20 indicate a degree of local dependence that should be examined by test developers (Chen & Thissen, 1997).

Because the three residual correlations are very similar, the default "standardized residual correlation" in WINSTEPS was used for these analyses. Table 3.1 presents the summary statistics (i.e., the mean, standard deviation, minimum, maximum, and several percentiles (P_{10}, P_{25}, P_{50}, P_{75}, P_{90})) for all the residual correlations. The table also presents the total number of item pairs (N) and the number of pairs with residual correlations greater than 0.20. There are no item pairs with residual correlations greater than 0.20. The mean residual correlations are slightly negative, and the values are close to -0.02. Most of the correlations are very small, suggesting that local item independence generally holds for the Regents Examination in Algebra I.

Table 3.1. Summary of Item Residual Correlations: Regents Examination in Algebra I

Statistic	Value
N	666
Mean	-0.02
SD	0.03
Minimum	-0.10
P_{10}	-0.06
P_{25}	-0.04
P_{50}	-0.02
P_{75}	0.00
P_{90}	0.01
Maximum	0.13
>|0.20|	0.00

3.4.3. Item Fit

An important assumption of the Rasch model is that the data for each item fit the model. WINSTEPS provides two item fit statistics (INFIT and OUTFIT) for evaluating the degree to which the Rasch model predicts the observed item responses for a given set of test items. Each fit statistic can be expressed as a mean square (MnSq) statistic or on a standardized metric (Zstd with mean = 0 and variance = 1). MnSq values are more oriented toward practical significance, while Zstd values are more oriented toward statistical significance. INFIT MnSq values are the average of standardized residual variance (the difference between the observed score and the Rasch-estimated score divided by the square root of the Rasch-model variance). The INFIT statistic is weighted by the θ relative to item difficulty and tends to be affected more by unexpected responses close to the person, item, or rating scale category measure (i.e., informative, on-target responses).

The expected MnSq value is 1.0 and can range from 0.0 to infinity. Deviation in excess of the expected value can be interpreted as noise or lack of fit between the items and the model. Values lower than the expected value can be interpreted as item redundancy or overfitting items (too predictable, too much redundancy), and values greater than the expected value indicate underfitting items (too unpredictable, too much noise). Rules of thumb regarding "practically significant" MnSq values vary.

Table 3.2 presents the summary statistics of INFIT mean square statistics for the Regents Examination in Algebra I, including the mean, standard deviation, and minimum and maximum values. The table also presents the number of items within a targeted range of [0.7, 1.3]. The mean INFIT value is 1.00, with all 37 items falling in a targeted range of [0.7, 1.3]. As the range of [0.7, 1.3] is used as a guide for ideal fit, fit values outside of the range are considered individually. Overall, these results indicate that, in general, the Rasch model fits the Regents Examination in Algebra I item data well.

Table 3.2. Summary of INFIT Mean Square Statistics: Regents Examination in Algebra I

N	Mean	SD	Min.	Max.	[0.7, 1.3]
37	1.00	0.12	0.79	1.28	[37/37]

3.5. 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 NYSED 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 A: contains the operational test maps for the June 2022 administration. The statistics presented in the test maps were generated based on the field test data.

All Regents Examinations are pre-equated, meaning that the parameters used to derive the relationship between the raw and scale scores are estimated prior to the construction and administration of the operational form. These field tests are administered to as small a sample of students as possible to minimize the effect on student instructional time throughout the state. The small n-counts associated with such administrations are sufficient for reasonably accurate estimation of most items' parameters. However, the parameters for the 6-point essay item can be unstable when estimated across as small a sample as is typically used. Therefore, a set of

constants is used for these items' parameters on operational examinations. These constants were set by NYSED and are based on the values in the bank for all essay items. The Regents Examination in Algebra I only has one 6-point item with fixed constants as follows:$D = 0.33$, $F_0 = 0.00$, $F_1 = 0.65$, $F_2 = -0.89$, $F_3 = -0.46$, $F_4 = -0.35$, $F_5 = 0.45$, and $F_6 = 0.83$.

The Regents Examination in Algebra I has four cut scores, which are set at the scale scores of 55, 65, 80 (floating), and 85. The third cut point at 80 was determined at the most recent Algebra I standard setting session during which a raw score associated with assessment performance at this performance level was determined. The scale score associated with this raw score was then determined and used for the performance level cut moving forward.

A primary consideration during test construction was to select items that would minimize changes in the raw scores corresponding to these 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 0.133. 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 between the items from administration to administration.

The relationship between raw and scale scores is explicated in the scoring table for the June 2022 administration, presented in Appendix B. This table is the end product of the following scaling procedure.

All Regents Examinations are equated back to a base scale held constant from year to year. Specifically, they are equated to the base scale using 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 exam, the base administration was the June 2014 administration. Scale scores from the June 2022 administration are on the same scale and can be directly compared to scale scores on all previous administrations back to the June 2014 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. 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 55, 65, and 85 were set to correspond to those raw score cuts. A fourth-degree polynomial is required to fit a line exactly to five arbitrary points (e.g., the raw scores corresponding to the five critical scale scores of 0, 55, 65, 85, and 100). The general form of this best-fitting line is as follows:

$$SS = m4 * RS^4 + m3 * RS^3 + m2 * RS^2 + m1 * RS^1 + m0$$

where *SS* is the scale score, *RS* is the raw score, and *m0* through *m4* are the transformation constants that convert the raw score into the scale score (*m0* will always be equal to 0 in this application, as a raw score of 0 corresponds to a scale score of 0). 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 *m4* specific to this subject were then used to determine the scale scores corresponding to the remainder of the raw scores on the exam. 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 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 then 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 55, 65, 80, and 85. The minimum (0) and maximum possible raw scores are assigned scale scores of 0 and 100, respectively. If 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 0 have scale scores that round to 0, their scale scores are instead set equal to 1.

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. This rule does not apply for the third floating cut score. If no scale score rounds to these critical cuts, 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), or 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 student score estimates maximize consistency and therefore minimize error or, theoretically speaking, that students 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 variability of observed scores can influence reliability estimates. Tests with fewer items or with a lack of heterogeneity in scores tend to produce lower reliability estimates. Second, reliability is concerned with random sources of error. Accordingly, the degree of inconsistency due to random error sources is what determines reliability: less consistency is associated with lower reliability, and more consistency is associated with higher reliability. Systematic error sources also exist.

This chapter discusses reliability results for the Regents Examination in Algebra I and three additional statistical measures to address the multiple factors affecting an interpretation of the exam's reliability: standard errors of measurement (SEMs), decision consistency, and group means.

4.1. RELIABILITY INDICES (STANDARD 2.20)

Classical test theory describes reliability as a measure of the internal consistency of test scores. The reliability (ρ_X^2) is defined as the ratio of true score variance (σ_T^2) to the observed score variance (σ_X^2), as presented in the equation below. The total variance contains two components: (1) the variance in true scores and (2) the variance due to the imperfections in the measurement process (σ_E^2). In other words, total variance equals true score variance plus error variance.[3]

$$\rho_X^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

Reliability coefficients indicate the degree to which differences in test scores reflect true differences in the attribute being tested rather than random fluctuations. Total test score variance (i.e., individual differences) is partly due to real differences in the construct (true variance) and partly due to random error in the measurement process (error variance).

Reliability coefficients range from 0.0 to 1.0. The index is 0.0 if none of the test score variances are true. Such scores would be pure random noise (i.e., all measurement error). If all test score variances were true, the index would equal 1.0. 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, larger coefficients are more desirable because they indicate that the test scores are less influenced by random error.

[3] A covariance term is not required, as true scores and error are assumed to be uncorrelated in classical test theory.

Reliability is most often estimated using the formula for coefficient alpha, which provides a practical internal consistency index. Coefficient alpha 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. 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:

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N}\sigma_{Yi}^2}{\sigma_X^2}\right)$$

where N is the number of parts (items), σ_X^2 is the variance of the observed total test scores, and σ_{Yi}^2 is the variance of part *i*.

4.2. STANDARD ERROR OF MEASUREMENT (STANDARDS 2.13, 2.14, 2.15)

Reliability coefficients best reflect the extent to which measurement inconsistencies may be present or absent. The standard error of measurement (SEM) is another indicator of test score precision that is better suited for determining the effect of measurement inconsistencies for the scores obtained by individual examinees. This is particularly so for conditional SEMs (CSEMs).

4.2.1. Traditional Standard Error of Measurement

The SEM is defined as the standard deviation of the distribution of observed scores for students with identical true scores, as shown below. The SEM is an index of the random variability in test scores in test score units and therefore represents important information for test score users.

$$SEM = SD\sqrt{1 - \alpha}$$

This formula indicates that the value of the SEM depends on both the reliability coefficient (the coefficient alpha) and the standard deviation of test scores. If the reliability were equal to 0.00 (the lowest possible value), the SEM would be equal to the standard deviation of the test scores. If test reliability were equal to 1.00 (the highest possible value), the SEM would be 0.0. Therefore, a perfectly reliable test has no measurement error (Harvill, 1991). The value of the SEM also takes the group variation (i.e., score standard deviation) into account. For example, an SEM of 3 on a 10-point test would be very different from an SEM of 3 on a 100-point test.

4.2.2. 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 precision of individual test scores and help place “reasonable limits” (Gulliksen, 1950) 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. For 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%.

[4] Some prefer the following interpretation: If a student were tested an infinite number of times, the ±1 SEM confidence intervals constructed for each score would capture the student’s true score 68% of the time.

Table 4.1 presents the coefficient alpha and associated SEM for the Regents Examination in Algebra I.

Table 4.1. Coefficient Alpha and SEM: Regents Examination in Algebra I

Coefficient Alpha	SEM
0.93	5.66

Assuming normally distributed scores, one would expect about two-thirds of the observations to be within one standard deviation of the mean. An estimate of the standard deviation of the true scores can be computed as follows:

$$\hat{\sigma}_T = \sqrt{\hat{\sigma}_x^2 - \hat{\sigma}_x^2(1-\hat{\rho}_{xx})}$$

4.2.3. Conditional Standard Error of Measurement

Every time an assessment is administered, a student's score contains some error. If the same exam were administered an infinite number of times to the same student, the mean of the distribution of the student's raw scores would be equal to the student's true score (θ) (i.e., the score obtained with no error), and the standard deviation of the distribution of the student's raw scores would be the conditional standard error. Because there is a one-to-one correspondence between the raw score and θ in the Rasch model, this concept can be applied more generally to all students who obtained a particular raw score, and the probability of obtaining each possible raw score can be calculated given the students' estimated θ. The standard deviation of this conditional distribution is defined as the CSEM. The computer program POLYCSEM (Kolen, 2004) was used to carry out the mechanics of this computation.

The relationship between θ and the scale score is not expressible in a simple mathematical form because it is a blend of the fourth-degree polynomial relationship between the raw and scale scores and the nonlinear relationship between the expected raw and θ scores. In addition, as the exam is equated from year to year, the relationship between the raw and scale scores moves away from the original fourth-degree polynomial relationship to one that is also no longer expressible in simple mathematical form. In the absence of a simple mathematical relationship between θ and the scale scores, the CSEMs available for each θ score via Rasch IRT cannot be converted directly to the scale score metric.

The use of Rasch IRT to scale and equate the Regents Examinations does, however, make it possible to calculate CSEMs by using the procedures described by Kolen et al. (1996) for dichotomously scored items and extended by Wang et al. (2000) to polytomously scored items. For tests such as the Regents Examinations that do not have a one-to-one relationship between raw (θ) and scale scores, the CSEM for each achievable scale score can be calculated by using the compound multinomial distribution to represent the conditional distribution of raw scores for each level of θ.

Consider a student with a certain performance level. If it were possible to measure this student's performance perfectly, without any error, this measure could be called the student's true score, which is equal to the expected raw score. However, whenever a student takes a test, the observed test score always includes some level of measurement error. Sometimes this error is positive, and the student achieves a higher score than would be expected given the student's level of θ. Other times it is negative, and the student achieves a lower-than-expected score. If a student

could be given the same test multiple times and their observed test scores recorded, the resulting distribution would be the conditional distribution of raw scores for that student's level of θ with a mean value equal to the student's expected raw (true) score. The CSEM for that level of θ in the raw score metric is the square root of the variance of this conditional distribution.

The conditional distribution of raw scores for any level of θ is the compound multinomial distribution (Wang et al., 2000). An algorithm to compute this can be found in Hanson (1994) and in Thissen et al. (1995) and is also implemented in the computer program POLYCSEM (Kolen, 2004). The compound multinomial distribution yields the probabilities that a student with a given level of θ has of attaining each achievable raw (and accompanying scale) score. The point values associated with each achievable raw or scale score point can be used to calculate the mean and variance of this distribution in the raw or scale score metric, respectively. The square root of the variance is the CSEM of the raw or scale score point associated with the current level of θ.

4.2.4. Conditional Standard Error of Measurement Confidence Intervals

CSEMs allow statements regarding the precision of individual test scores. Like SEMs, they help place reasonable limits around observed scale scores through the construction of an approximate score band. The confidence intervals are constructed by adding and subtracting a multiplicative factor of the CSEM.

4.2.5. Conditional Standard Error of Measurement Characteristics

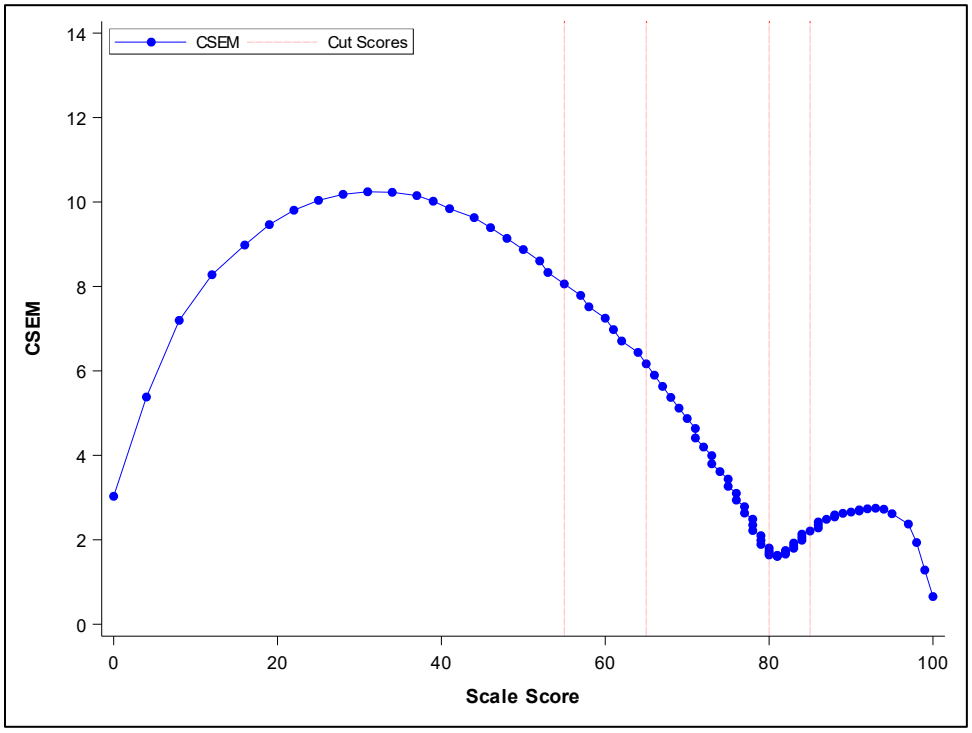
The relationship between the scale score CSEM and θ depends on the nature of the raw-to-scale score transformation (Kolen & Brennan, 2005; Kolen & Lee, 2011) and on whether the CSEM is derived from the raw scores or from θ (Lord, 1980). The pattern of CSEMs for raw scores and linear transformations of the raw score tend to have a characteristic "inverted-U" shape, with smaller CSEMs at the ends of the score continuum and larger CSEMs toward the middle of the distribution. Achievable raw score points for these distributions are spaced equally across the score range. Kolen and Brennan 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)" (2005, p. 357).

4.2.6. Results and Observations

The relationship between raw and scale scores for the Regents Examinations tends to be roughly linear from scale scores of 0 to 65 and then concave down from about 65 to 100. In other words, the scale scores track linearly with the raw scores for the lower two-thirds of the scale score range and are then compressed relative to the raw scores for the remaining one-third of the range, though there are variations. The CSEMs for the Regents Examinations can be expected to have inverted-U shaped patterns, with some variations.

Figure 4.1 shows this type of CSEM variation for the Regents Examination in Algebra I, where the compression of raw score to scale scores around the cut score of 80 changes the shape of the curve noticeably. This type of expansion and compression can be seen in the figure by looking at the changing density of raw score points along the scale score range on the horizontal axis. The largest compression can be seen between about 65 to 90 scale score points.

Figure 4.1. Conditional Standard Error Plot: Regents Examination in Algebra I



4.3. DECISION CONSISTENCY AND ACCURACY (STANDARD 2.16)

In a standards-based testing program, there is interest in knowing how accurately students are classified into performance categories. In contrast to the coefficient alpha that is concerned with the relative rank-ordering of students, it is the absolute values of student scores that are important in decision consistency and accuracy.

Classification consistency refers to the degree to which the performance level for each student can be replicated upon retesting using an equivalent form (Huynh, 1976). Decision consistency answers the question, what is the agreement in classifications between the two nonoverlapping, equally difficult forms of the test? If two parallel forms of the test were given to the same students, the consistency of the measure would be reflected by the extent to which the classification decisions based on the first set of test scores matched the decisions based on the second set of test scores.

Consider Table 4.2 and Table 4.3. If a student is classified as being in one category based on the Test 1 score, how probable would it be that the student would be reclassified as being in the same category if the student took Test 2 (a nonoverlapping, equally difficult form of the test)? This proportion is a measure of decision consistency. The proportions of correct decisions, φ, for two and four categories are computed by the following formulas, respectively:

$$\varphi = \varphi_{11} + \varphi_{22}$$

$$\varphi = \varphi_{11} + \varphi_{22} + \varphi_{33} + \varphi_{44}$$

The sum of the diagonal entries (i.e., the proportion of students classified by the two forms into the same performance level) signifies the overall consistency.

Table 4.2. Pseudo-Decision Table for Two Hypothetical Categories

		Test 1		
		Level 1	Level 2	Marginal
Test 2	Level 1	φ11	φ12	φ1●
	Level 2	φ21	φ22	φ2●
	Marginal	φ●1	φ●2	1

Table 4.3. Pseudo-Decision Table for Four Hypothetical Categories

		Test 1				
		Level 1	Level 2	Level 3	Level 4	Marginal
Test 2	Level 1	φ11	φ12	φ13	φ14	φ1●
	Level 2	φ21	φ22	φ23	φ24	φ2●
	Level 3	φ31	φ32	φ33	φ34	φ3●
	Level 4	φ41	φ42	φ43	φ44	φ4●
	Marginal	φ●1	φ●2	φ●3	φ●4	1

Classification accuracy refers to the agreement of the observed classifications of students with the classifications made based on their true scores. An observed score contains measurement error, while a true score is theoretically free of measurement error. A student's observed score can be formulated by the sum of the true score plus measurement error ($Observed = True + Error$). Decision accuracy is an index to determine the extent to which measurement error causes a classification different than the one expected from the true score. Because true scores are unobserved and decision consistency is computed based on a single administration of the Regents Examination, a statistical model using data only from the available administration is used to estimate the true scores and to project the consistency and accuracy of classifications (Hambleton & Novick, 1973). Although several procedures are available, a well-known method developed by Livingston and Lewis (1995) that implements 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, as classification and accuracy based on 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.

Table 4.4 presents the results for the dichotomies created by the four corresponding cut scores. For example, the statistics under "2/3" indicate the decision consistency and accuracy when the performance levels are divided into two categories: one for the second and lower performance level, and the other for the third and higher performance levels. The tabled values are derived with the program *BB-Class* (Brennan, 2004) using the Livingston and Lewis method (1995). The decision consistency ranged from 0.86 to 0.95, and the decision accuracy ranged from 0.90 to 0.96. Both decision consistency and accuracy values are high, indicating very good consistency and accuracy of student classifications.

Table 4.4. Decision Consistency and Accuracy Results: Regents Examination in Algebra I

Statistic	1/2	2/3	3/4	4/5
Consistency	0.86	0.88	0.92	0.95
Accuracy	0.90	0.91	0.94	0.96

4.4. GROUP MEANS (STANDARD 2.17)

The student population for the Regents Examination in Algebra I is composed of students who have completed a course in Algebra I. Table 4.5 presents the overall mean scale score that was computed based on all students who took the exam.

Table 4.5. Group Means: Regents Examination in Algebra I

Demographics	#Students	Mean Scale Score	SD Scale Score
All Students	211,056	69.59	15.60
Race/Ethnicity			
American Indian/Alaska Native	1,429	64.37	15.00
Asian/Native Hawaiian/Other Pacific Islander	22,144	78.45	13.79
Black/African American	34,168	61.75	15.03
Hispanic/Latino	61,361	63.45	15.29
Multiracial	5,567	71.42	15.22
White	86,332	74.75	13.18
English Language Learner			
No	192,798	70.89	14.98
Yes	18,258	55.78	15.41
Economically Disadvantaged			
No	101,526	75.00	13.69
Yes	109,530	64.57	15.59
Gender			
Female	103,662	70.51	15.28
Male	107,269	68.69	15.86
Nonbinary	70	74.21	11.71
Student with a Disability			
No	178,996	71.80	14.61
Yes	32,060	57.25	15.23

Note. 55 students were not reported in the race/ethnicity and gender groups, but they are reflected in "All Students."

4.5. STATE PERCENTILE RANKINGS

Table 4.6 presents the state percentile rankings based on scale score distributions of all students taking the Regents Examination in Algebra I for the June 2022 administration. The scale scores for range from 0 to 100, and some scale scores may not be obtainable depending on the raw score-to-scale score relationship for a specific administration. The percentile ranks are computed in the following manner:

- A student's assigned "state percentile rank" will be the cumulative percentage of students scoring at the immediate lower score plus half of the percentage of students obtaining the given score.
- Students who obtain the highest possible score will receive a percentile rank of 99.

Table 4.6. State Percentile Ranking for Scale Score: Regents Examination in Algebra I

Scale Score	Percentile Rank	Scale Score	Percentile Rank	Scale Score	Percentile Rank	Scale Score	Percentile Rank
0	1	26	1	52	16	78	64
1	1	27	1	53	17	79	68
2	1	28	1	54	17	80	72
3	1	29	1	55	18	81	76
4	1	30	1	56	20	82	80
5	1	31	1	57	22	83	82
6	1	32	2	58	23	84	85
7	1	33	2	59	24	85	87
8	1	34	2	60	26	86	89
9	1	35	3	61	28	87	91
10	1	36	3	62	29	88	92
11	1	37	3	63	30	89	93
12	1	38	4	64	32	90	94
13	1	39	4	65	34	91	95
14	1	40	5	66	36	92	96
15	1	41	5	67	38	93	96
16	1	42	6	68	40	94	97
17	1	43	7	69	41	95	98
18	1	44	8	70	43	96	98
19	1	45	9	71	45	97	99
20	1	46	10	72	47	98	99
21	1	47	10	73	49	99	99
22	1	48	11	74	52	100	99
23	1	49	12	75	55	–	–
24	1	50	13	76	57	–	–
25	1	51	15	77	61	–	–

Chapter 5: Validity (Standard 1)

To restate the purposes and uses of the Regents Examination in Algebra I, this exam measures student achievement against the NYS P–12 Learning Standards and was prepared by teacher examination committees and NYSED subject matter and testing specialists. Further, 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, which may then be used to guide classroom teaching and learning. The exam also provides 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 Algebra I is intended for use in satisfying state testing requirements for students who have finished a course in Algebra I. A passing score on the exam counts toward requirements for a high school diploma, as described in the New York State diploma requirements (https://www.nysed.gov/sites/default/files/programs/curriculum-instruction/currentdiplomarequirements2.pdf). Results of the exam may also be used to satisfy various locally established requirements throughout the state.

The validity of score interpretations for this exam is supported by the following sources of evidence, which are important to gather and document to support validity claims for an assessment as provided by the *Standards for Educational Psychological Testing* (AERA et al., 2014):

- Test content
- Response processes
- Internal test structure
- Relation to other variables
- Consequences of testing

While these categories are not mutually exclusive, as one source of validity evidence often falls into more than one category, they provide a useful framework within the *Standards* for the discussion and documentation of validity evidence. 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 within 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. The Regents Examination in Algebra I measures student achievement on the NYS P–12 Learning Standards for Mathematics, located at https://www.nysed.gov/curriculum-instruction/mathematics-learning-standards.

5.1.1. Content Validity

Content validity is concerned with the proper definition of the construct and evidence that the test provides an accurate measure of student performance within the defined construct. The test blueprint is essentially the design document for test construction that provides an explicit definition of the construct domain that is to be represented on the exam. The test development process is in place to ensure, to the extent possible, that the blueprint is met in all operational forms of the exam. Table 5.1 presents a summary of the Algebra I blueprint.

Table 5.1. Test Blueprint: Regents Examination in Algebra I

Conceptual Category	% of Test by Credits	Domains in Algebra I
Number & Quantity	2–8%	The Real Number System (N-RN) Quantities (N-Q)
Algebra	50–56%	Seeing Structure in Expressions (A-SSE) Arithmetic with Polynomials and Rational Expressions (A-APR) Creating Equations (A-CED) Reasoning with Equations and Inequalities (A-REI)
Functions	32–38%	Interpreting Functions (F-IF) Building Functions (F-BF) Linear, Quadratic, and Exponential Models (F-LE)
Statistics & Probability	5–10%	Interpreting Categorical and Quantitative Data (S-ID)

5.1.2. Item Development Process

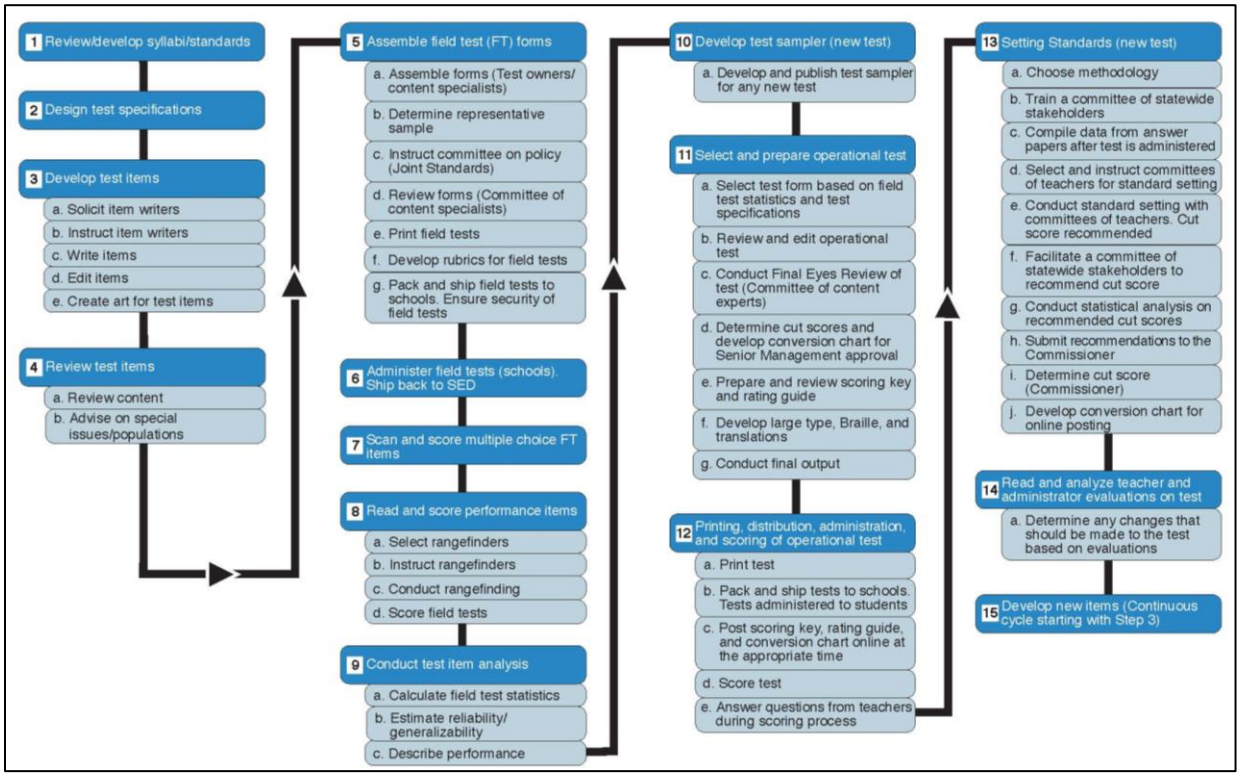
Test development is a detailed, step-by-step process of development and review cycles. An important element of this process is that all test items are developed by NYS educators in a process facilitated by state subject matter and testing experts. Bringing experienced classroom teachers into this central item development role serves to draw a strong connection between classroom and test content.

Only NYS-certified educators may participate in this process. NYSED asks for nominations from districts, and all recruiting is done with diversity of participants in mind, including diversity in gender, ethnicity, geographic region, and teaching experience. Educators with item writing skills from throughout the state are retained to write all items for the Regents Examinations, under strict guidelines that leverage best practices (see Appendix C). State educators also conduct all item quality and bias reviews to ensure that item content is appropriate to the construct being measured and fair for all students. Finally, educators use the defined standards, test blueprint targets, and statistical information generated during field testing to select the highest-quality items for use in the operational test.

Initial item development is conducted under the criteria and guidance provided by multiple documents, including the blueprint, item writing criteria, and a content verification checklist. Both MC and CR items are included on the exam to ensure appropriate coverage of the construct domain. The item writing guidelines in Appendix C provide detailed information about how the items are developed.

Figure 5.1 summarizes the full test development process, with Steps 3 and 4 addressing initial item development and review. This figure also demonstrates the ongoing nature of ensuring the content validity of items through field test trials and final item selection for operational testing.

Figure 5.1. NYSED Test Development Process



5.1.3. Item Review Process

The item review process assists in the consistent application of rigorous item reviews intended to assess the quality of the items developed and identify items that require edits or removal from the pool of items to be field-tested. The following criteria help to ensure that high-quality items are continually developed in a manner that is consistent with the test blueprint.

All reviewers participate in rigorous training designed to assist in a consistent interpretation of the standards throughout the item review process. This is a critical step in item development because consistency between the standards and what the items are asking students is a fundamental form of evidence of the validity of the intended score interpretations. Another integral component of this item review process is to review the scoring rules, or rubrics, for their clarity and consistency in what the student is being asked to demonstrate by responding to each item. Each of these elements is in place to target fairness for all students by targeting consistency in student scores and providing evidence of the validity of their interpretations. Following these reviews, only items that are approved by an assigned educator panel move forward for field testing.

Specifically, the item review process articulates four major item characteristics that NYSED looks for when developing quality items:

1. Language and graphical appropriateness
2. Sensitivity/bias
3. Alignment of measurement to standards
4. Conformity to the expectations for the specific item types and formats

Each section of the criteria includes pertinent questions that help reviewers determine whether an item is of sufficient quality. Within the first two categories, criteria for language appropriateness are used to help ensure that students understand what is asked in each item and that the language in the item does not adversely affect a student's ability to perform the required task. Similarly, the sensitivity/bias criteria are used to evaluate whether items are unbiased, non-offensive, and not disadvantageous to any given subgroup(s). The mathematics art criteria assess the appropriateness and clarity of any graphics that are used within items.

The third category of the item review (alignment) addresses how each item measures a given mathematics standard. This criterion asks the reviewer to comment on key aspects of how the item addresses and calls for the skills demanded by the standards. These criteria also prompt reviewers to comment on how more than one standard is addressed by a given item. The fourth category (criteria framework) addresses the demands for different item types and formats. Reviewers evaluate each item to ensure that it conforms to the given requirements. For example, MC items must have one unambiguously correct answer and several plausible, but incorrect, answer choices.

Ongoing attention is also given to the relevance of the standards used to guide curriculum and assessment. Consistent with a desire to assess this relevance, NYSED is committed to ongoing standards review over time and periodically solicits thoughtful, specific responses from stakeholders about individual standards within the NYS P–12 Learning Standards.

5.2. EVIDENCE BASED ON RESPONSE PROCESSES

The second source of validity evidence is based on student response processes and requires evidence that students 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. 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 Examinations include the item development process, with attention paid to mitigating the introduction of construct-irrelevant variance. The test development process details the methods used and attention given to reducing the potential for construct irrelevance in response processes. This is accomplished by attending to the quality and alignment of test content to the test blueprint and to the item development guidelines (Appendix C). Further evidence is documented in the test administration and scoring procedures and in the results of statistical analyses.

5.2.1. Administration and Scoring

Adherence to standardized administration procedures is fundamental to the validity of test scores and their interpretation, as such procedures allow for adequate and consistently applied conditions for scoring the work of every student who takes the examination. For this reason, guidelines, as contained in the *School Administrator's Manual* (https://www.nysed.gov/state-assessment/test-manuals), have been developed and implemented. All secondary-level Regents Examinations are administered under these standard conditions to support valid inferences for all students. These procedures also cover testing students with disabilities who are provided testing accommodations consistent with their Individualized Education Programs (IEPs) or Section 504 Accommodation Plans (504 Plans). Full test administration procedures are available at https://www.nysed.gov/state-assessment/high-school-regents-examinations.

The implementation of rigorous scoring procedures supports the validity of the scores. Regents test scoring practices therefore focus on producing high-quality scores. MC items are scored via local scanning at testing centers, and trained educators score CR items. Many studies focus on various elements of producing valid and reliable scores for CR items, but generally, attention to the following all contribute to valid and reliable scores for CR items:

1. Quality training (Hoyt & Kerns, 1999; Lumley & McNamara, 1995; Wang et al., 2010; Gorman & Rentsch, 2009; Schleicher et al., 2002; Woehr & Huffcutt, 1994; Johnson et al., 2008; Weigle, 1998)
2. Detection and correction of rating bias (McQueen & Congdon, 1997; Congdon & McQueen, 2000; Myford & Wolfe, 2009; Barkaoui, 2011; Patz et al., 2002)
3. Consistency or reliability of ratings (Congdon & McQueen, 2000; Harik et al., 2009; McQueen & Congdon, 1997; Myford & Wolfe, 2009; Mero & Motowidlo, 1995; Weinrott & Jones, 1984)
4. Rubric designs that facilitate consistency of ratings (Pecheone & Chung, 2007; Wolfe & Gitomer, 2000; Cronbach et al., 1995; Cook & Beckman, 2009; Penny et al., 2000; Smith, 1993; Leacock et al., 2014).

The distinct steps for operational test scoring include close attention to each of these elements and begin before the operational test is selected. After the field test process, a set of anchor papers representing student responses across the range of possible responses for CR items is selected. The objective of these range-finding efforts is to create a training set for scorer training and execution. The scores calculated for this training set 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 item. Committee members familiarize themselves with each item type and score several responses that are representative of each score point. After the independent scoring is completed, the committee reviews and discusses their results and determines consensus scores for the student responses. 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 CR items. The Rating Guide for the June 2022 administration of the Regents Examination in Algebra I are located at https://www.nysedregents.org/algebratwo/.

During the range-finding and field test scoring processes, it is important to control for sources of variation in scoring. One possible source of variation in CR scores is unintended rater bias associated with items and student responses. The rater is often unaware of such bias, so 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 that 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 controlled by training practices with a strict focus on rubric requirements.

The training process for operational scoring by NYS educators begins with a review and discussion of actual student work on CR items. This helps raters understand the range and characteristics typical of student responses and the kinds of mistakes 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.

Raters then receive training consistent with the guidelines and ancillaries produced after field testing and are allowed to practice scoring prior to the start of live scoring. Throughout the scoring process, there are important procedures for correcting inconsistent scoring or the misapplication of scoring rubrics for CR items. When monitoring and correction do not occur during scoring, construct-irrelevant variation may be introduced. Accordingly, a scoring lead may be assigned to review the consistency of scoring for the lead's assigned staff against model responses and to be available for consultation throughout the scoring process.

Attention to the rubric design also contributes to the validity of student response processes. The rubric specifies what the student needs to provide as evidence of learning based on the question asked. The more explicit the rubric (and the item), the clearer the response expectations are for students. To facilitate the development of CR 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 alternate approaches and other legitimate methods.

In support of the goal of valid score interpretations for each student, such scoring training procedures are implemented for the Regents Examinations. Operational raters are selected based on expertise in the exam subject and are assigned a specific set of items to score. No more than one-third of the items on the test are assigned to any one rater. This increases the consistency of scoring across student responses by allowing each rater to focus on a subset of items. It also ensures that no one rater is allowed to score the entire test for any one student. This practice reduces the effect of any potential bias of a single rater on individual students. Raters are also not allowed to score the responses of their own students.

5.2.2. Statistical Analysis

A useful statistic for evaluating the response processes for MC items is an item's point-biserial correlation on the distractors. A high point-biserial on a distractor may indicate that students are not able to identify the correct response for a reason other than the difficulty of the item. A finding of poor model fit for an item may also support a finding that students are not responding in the way in which the item developer intended. As shown in Table 2.1, the point-biserial correlations for distractors in the MC items all appear to be negative or close to zero, indicating that, for the most part, students are not being drawn to an unintended construct.

5.3. EVIDENCE BASED ON INTERNAL STRUCTURE

The third source of validity evidence comes from the internal structure of the test and requires that test developers evaluate the test structure to ensure that it is functioning as intended. Such an evaluation may include attention to item interactions, tests of dimensionality, or indications of test bias for or against one or more student subgroups detected by differential item functioning (DIF) analysis. Evaluation of internal test structure also includes a review of the results of classical item analyses, test reliability, and the IRT scaling and equating. The following analyses were conducted for the Regents Examination in Algebra I:

- Item difficulty
- Item discrimination
- Differential item functioning (DIF)
- IRT model fit
- Test reliability
- Classification consistency
- Test dimensionality

5.3.1. Item Difficulty

Multiple analyses allow for an evaluation of item difficulty. For this exam, *p*-values and Rasch difficulty (item location) estimates were computed for MC and CR items. Items for the June 2022 Regents Examination in Algebra I show a range of *p*-values consistent with the targeted exam difficulty. Item *p*-values range from 0.20 to 0.77, with a mean of 0.48. The difficulty distribution illustrated in Figure 2.1 shows a wide range of item difficulties on the exam. This is consistent with general test development practice, which seeks to measure student ability along a full range of difficulty. Refer to Chapter 2 of this report for additional details.

5.3.2. Item Discrimination

How well the items on a test discriminate between high- and low-performing students is an important measure of the test structure. Items that do not discriminate well generally provide less reliable information about student performance. Table 2.1 and Table 2.2 provide point-biserial values on the correct responses; Table 2.1 also provides point-biserial values on the three distractors. The values for correct answers are 0.18 or higher for all items with all but three items exceeding 0.35, indicating that most items are discriminating well between high- and low-performing students. Point-biserial values for all distractors are negative, indicating that students are responding to the items as expected during item development. Refer to Chapter 2 of this report for additional details.

5.3.3. Differential Item Functioning

Differential Item Functioning (DIF) was conducted for gender, race/ethnicity, needs/resource capacity (NRC) categories, and ELL status based on the data for the June 2022 administration. DIF data are only available after the administration as all Regents exams are pre-equated, meaning that the parameters used to derive the relationship between the raw and scale scores are estimated prior to the construction and administration of the operational form.

The Mantel-Haenszel (Mantel & Haenszel, 1959) and standardized mean difference (Dorans & Schmitt, 1991) methods were used to detect items that may function differently for any subgroup. The Mantel-Haenszelχ^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, a 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 (the total test score in this analysis). The results of these analyses were examined by NYSED content specialists to identify potential systematic issues that could be addressed in future item writing.

5.3.4. IRT Model Fit

Model fit for the Rasch method used to estimate location (difficulty) parameters for the items on the exam provide important evidence that the internal structure of the test is of high technical

quality. The number of items within a targeted range of [0.7, 1.3] is reported in Table 3.2. The mean INFIT value is 1.00, with all items falling in a targeted range of [0.7, 1.3]. As the range of [0.7, 1.3] is used as a guide for ideal fit, fit values outside of the range are considered individually. These results indicate that, for most items, the Rasch model fits the item data well.

5.3.5. Test Reliability

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 the domain. Reliability should ultimately demonstrate that student score estimates maximize consistency and therefore minimize error or, theoretically, that students who take a test multiple times would get the same score each time. The reliability estimate for the Regents Examination in Algebra I is 0.93, showing high reliability of student scores. Refer to Chapter 4 of this report for additional details.

5.3.6. Classification Consistency and Accuracy

Decision consistency measures the agreement between the classifications based on two nonoverlapping, equally difficult forms of the test. If two parallel forms were given to the same students, the consistency of the measure would be reflected by the extent to which the classification decisions based on the first set of test scores matched the decisions based on the second set of scores. Decision accuracy determines the extent to which measurement error causes a classification different from that expected from the true score. High decision consistency and accuracy provide strong evidence that the internal structure of a test is sound.

For the Regents Examination in Algebra I, both decision consistency and accuracy values are high, indicating very good consistency and accuracy of student classifications. The results for the overall consistency across all five performance levels, as well as for the dichotomies created by the four corresponding cut scores, are presented in Table 4.4. The tabled values are derived with the program BB-Class (Brennan, 2004) using the Livingston and Lewis method. The decision consistency ranged from 0.86 to 0.95, and the decision accuracy ranged from 0.90 to 0.96.

5.3.7. Test 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 might suggest that the test is measuring something other than the intended content, indicating that the test structure quality is compromised. The results of a PCA conducted to test the assumption of unidimensionality provide strong evidence that a single dimension in the exam 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 and that the dimension most present is the targeted construct. Refer to Chapter 3 for details of this analysis.

Considering this collection of detailed analyses of the internal structure of the Regents Examination in Algebra I, strong evidence exists that the exam is functioning as intended and is providing valid and reliable information about student performance.

5.4. EVIDENCE BASED ON RELATIONS TO OTHER VARIABLES

Another source of validity evidence is based on the relation of the test to other variables. This source commonly encompasses two validity categories prevalent in the literature and practice: concurrent and predictive validity. To make claims about the validity of a test that is to be used for high-stakes purposes, such as the Regents Examination in Algebra I, these claims could be

supported by providing evidence that performance on the Algebra I test correlates well with other tests that measure the same or similar constructs. Although not absolute in its ability to offer evidence that concurrent test score validity exists, such correlations can be helpful for supporting a claim of concurrent validity if the correlation is high. To conduct such studies, matched student score data for other tests measuring the same content as the Regents Examination in Algebra I is ideal. However, the systematic acquisition of such data is complex and costly.

Importantly, a strong connection between classroom curriculum and test content may be inferred by the fact that NYS educators, deeply familiar with the curriculum standards and their enactment in the classroom, develop all content for the Regents Examination in Algebra I. In terms of predictive validity, time is a fundamental constraint on gathering evidence. The gold standard for supporting the validity of predictive statements about test scores requires empirical evidence of the relationship between test scores and future performance on a defined characteristic. To the extent that the objective of the standards is to prepare students for meeting graduation requirements, it will be important to gather evidence of this empirical relationship over time.

5.5. EVIDENCE BASED ON TESTING CONSEQUENCES

There are two general approaches in the literature for evaluating consequential validity. Messick (1995) points out that adverse social consequences invalidate test use mainly if they are due to flaws in the test. In this sense, the sources of evidence documented in this report (based on the construct, internal test structure, response processes, and relation to other 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 student scores on other tests is not directly supported by either the stated purposes or by the development process and research conducted on student data. A brief survey of websites for NYS universities and colleges finds that, beyond the explicitly defined use as a testing requirement toward graduation for students who have completed a course in Algebra I, 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 Algebra I are consistent with those required in the courses for which a student is seeking enrollment or placement. Educational institutions using the exam for placement purposes are advised to examine the scoring rules for the Regents Examination in Algebra I and to assess their appropriateness for the inferences being made about course placement.

5.6. SUMMARY

As stated, the nature of validity arguments is not absolute. Rather, it is supported through ongoing processes and studies designed to accumulate support for validity claims. The evidence provided in this report documents the evidence to date that supports the use of the Regents Examination in Algebra I scores for the purposes described.

References

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

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

Brennan, R. L. (2004). *BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy* [Computer software] (Version 1.0). University of Iowa.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289.

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. I. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995). *Generalizability analysis for educational assessments*. University of California, Center for the Study of Evaluation and The National Center for Research on Evaluation, Standards, and Student Testing.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS Research Report No. 91-49). Educational Testing Service.

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*. 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 Sage Publications.

Hanson, B. A. (1994). *Extension of Lord-Wingersky algorithm to computing test scores for polytomous items*. http://www.openirt.com/b-a-h/papers/note9401.pdf

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46(*1), 43–58.

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*(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.* The Guilford Press.

Kolen, M. J., (2004). POLYCSEM [Computer program]. University of Iowa. https://education.uiowa.edu/casma/computer-programs

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

Kolen, M. J. & Lee, W. (2011). Psychometric properties of raw and scale scores on mixed-format tests. *Educational Measurement: Issues and Practice 30*(2), 15–24.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*(2), 129–140.

Leacock, C., Gonzalez, E., & Conarroe, M. (2014). *Developing effective scoring rubrics for AI short answer scoring*. McGraw-Hill Education CTB Innovative Research and Development Grant.

Linacre, J. M. (2005). *WINSTEPS Rasch measurement computer program* (Version 3.60) [Computer software]. Winsteps.com.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*, 54–71.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719–748.

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*(3), 200–215.

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of a general theory. *Multivariate Behavioral Research, 14*, 21–38.

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 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.

Partnership for Assessment of Readiness for College and Careers (PARCC). (2014). *PARCC model content frameworks: Mathematics grades 3–11*.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*(4), 341–384.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. —III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society of London, 187*, 253–318.

Pecheone, R. L., & Chung, R. (2007). *Performance assessment for California teachers: Summary of validity and reliability studies for the 2003−04 pilot year*. Stanford University PACT Consortium.

Penny, J., Johnson, R. L., & Gordon, B. (2000). Using rating augmentation to expand the scale of an analytic rubric. *The Journal of Experimental Education, 68*(3), 269–287.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford. Nielsen & Lydiche.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207–230.

Schleicher, D. J., Day, D. V., Bronston, T., Mayes, B. T., & Riggo, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*(4), 735–746.

Smith, W. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M. M. Williamson & B. A. Hout (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations.* Hampton Press.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*, 39–49.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. Springer-Verlag.

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement*, *37*(2), 141–162.

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*(2), 263–287.

Weinrott, L., & Jones, B. (1984). Overt versus 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*(3)*,* 189–205.

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

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* MESA press.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Appendix A: Operational Test Map

Table A.1. June 2022 Test Map

Position	Item Type	Max. Points	Weight	Cluster	Mean	Point-Biserial	Rasch Difficulty	INFIT
01	MC	1	2	S-ID.C	0.78	0.28	-2.1015	1.14
02	MC	1	2	F-IF.A	0.72	0.44	-1.7237	0.95
03	MC	1	2	F-BF.A	0.71	0.47	-1.7222	0.92
04	MC	1	2	A-SSE.B	0.68	0.47	-1.5776	0.98
05	MC	1	2	A-CED.A	0.66	0.45	-1.4503	0.96
06	MC	1	2	F-IF.B	0.66	0.37	-1.4277	1.07
07	MC	1	2	A-REI.B	0.61	0.44	-1.1402	1.02
08	MC	1	2	F-LE.A	0.65	0.44	-1.2704	1.01
09	MC	1	2	A-SSE.B	0.59	0.35	-1.1005	1.15
10	MC	1	2	F-IF.A	0.60	0.41	-1.0414	1.04
11	MC	1	2	A-CED.A	0.55	0.5	-0.8110	0.98
12	MC	1	2	A-APR.B	0.54	0.5	-0.7305	0.97
13	MC	1	2	A-CED.A	0.51	0.36	-0.6612	1.09
14	MC	1	2	S-ID.A	0.40	0.4	0.0781	1.13
15	MC	1	2	F-IF.A	0.48	0.36	-0.4953	1.15
16	MC	1	2	F-IF.C	0.46	0.45	-0.3932	1.02
17	MC	1	2	A-APR.A	0.42	0.48	-0.2012	0.99
18	MC	1	2	A-REI.D	0.39	0.48	-0.0309	0.98
19	MC	1	2	F-IF.C	0.38	0.35	-0.0042	1.13
20	MC	1	2	A-SSE.A	0.38	0.45	0.0269	1.04
21	MC	1	2	F-LE.A	0.36	0.45	0.2482	1.03
22	MC	1	2	N-Q.A	0.30	0.19	0.4366	1.26
23	MC	1	2	A-REI.B	0.27	0.39	0.5144	1.06
24	MC	1	2	F-IF.A	0.27	0.4	0.5797	1.00
25	CR	2	1	N-RN.B	0.69	0.64	0.0380	0.98
26	CR	2	1	F-BF.B	0.61	0.64	0.1280	0.88
27	CR	2	1	F-IF.B	0.53	0.59	0.3868	1.00
28	CR	2	1	A-APR.A	0.55	0.62	0.4276	0.92
29	CR	2	1	F-IF.A	0.39	0.64	0.7471	0.86
30	CR	2	1	A-REI.B	0.38	0.6	0.8479	0.98
31	CR	2	1	A-SSE.A	0.46	0.57	0.7841	0.96
32	CR	2	1	A-REI.B	0.25	0.56	1.3205	0.92
33	CR	4	1	F-IF.B	1.34	0.75	0.1166	0.93
34	CR	4	1	A-REI.D	1.42	0.78	0.1862	0.89
35	CR	4	1	S-ID.C	1.02	0.73	0.5848	0.96
36	CR	4	1	A-REI.D	0.94	0.76	0.7500	0.92
37	CR	6	1	A-CED.A	1.57	0.84	0.2472	0.69

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

Table B.1. June 2022 Score Table

Raw Score	Ability	Scale Score	Raw Score	Ability	Scale Score	Raw Score	Ability	Scale Score
0	-6.0886	0.000	29	-0.7374	67.208	58	0.6146	81.197
1	-4.8650	4.372	30	-0.6807	68.172	59	0.6632	81.465
2	-4.1438	8.392	31	-0.6256	69.082	60	0.7128	81.740
3	-3.7101	12.216	32	-0.5720	69.938	61	0.7635	82.024
4	-3.3941	15.859	33	-0.5197	70.745	62	0.8154	82.319
5	-3.1425	19.331	34	-0.4687	71.505	63	0.8687	82.629
6	-2.9318	22.640	35	-0.4189	72.221	64	0.9236	82.955
7	-2.7491	25.797	36	-0.3701	72.894	65	0.9802	83.301
8	-2.5870	28.807	37	-0.3224	73.526	66	1.0386	83.669
9	-2.4407	31.678	38	-0.2756	74.121	67	1.0992	84.062
10	-2.3068	34.416	39	-0.2295	74.680	68	1.1621	84.482
11	-2.1830	37.028	40	-0.1842	75.205	69	1.2277	84.933
12	-2.0675	39.518	41	-0.1394	75.699	70	1.2964	85.417
13	-1.9590	41.890	42	-0.0952	76.162	71	1.3686	85.938
14	-1.8567	44.150	43	-0.0513	76.599	72	1.4448	86.498
15	-1.7595	46.303	44	-0.0078	77.010	73	1.5256	87.100
16	-1.6670	48.353	45	0.0355	77.398	74	1.6118	87.747
17	-1.5786	50.303	46	0.0786	77.764	75	1.7045	88.442
18	-1.4938	52.156	47	0.1217	78.110	76	1.8050	89.187
19	-1.4125	53.916	48	0.1649	78.437	77	1.9149	89.986
20	-1.3342	55.591	49	0.2081	78.750	78	2.0366	90.839
21	-1.2588	57.179	50	0.2515	79.049	79	2.1734	91.750
22	-1.1859	58.686	51	0.2952	79.337	80	2.3302	92.724
23	-1.1155	60.113	52	0.3392	79.616	81	2.5144	93.764
24	-1.0474	61.465	53	0.3836	79.886	82	2.7386	94.876
25	-0.9816	62.747	54	0.4285	80.152	83	3.0266	96.062
26	-0.9177	63.958	55	0.4740	80.413	84	3.4314	97.323
27	-0.8558	65.103	56	0.5201	80.673	85	4.1231	98.644
28	-0.7957	66.185	57	0.5670	80.934	86	5.3255	100.000

Appendix C: Item Writing Guidelines

GUIDELINES FOR WRITING MULTIPLE-CHOICE MATH ITEMS

1. **The item measures the knowledge, skills, and proficiencies characterized by the standards within the identified cluster.**

2. **The focus of the problem or topic should be stated clearly and concisely.**
   The stem should be meaningful and convey the central problem. A multiple-choice item functions most effectively when a student is required to compare specific alternatives related to the stem. It should not be necessary for the student to read all the alternatives to understand an item. *(Hint: Cover the alternatives and read the stem on its own. Then ask yourself if the question includes the essential elements or if the essential elements are lost somewhere in the alternatives.)*

3. **Include problems that come from a real-world context or problems that make use of multiple representations.**
   When using real-world problems, use formulas and equations that are real-world *(e.g., the kinetic energy of an object with mass, m, and velocity, v, is k = ½ mv2).* Use real-world statistics whenever possible.

4. **The item should be written in clear and simple language, with vocabulary and sentence structure kept as simple as possible.**
   Each multiple-choice item should be specific and clear. The important elements should generally appear early in the stem of an item, with qualifications and explanations following. Difficult and technical vocabulary should be avoided, unless essential for the purpose of the question.

5. **The stem should be written as a direct question or an incomplete statement.**
   Direct questions are often more straightforward. However, an incomplete statement may be used to achieve simplicity, clarity, and effectiveness. Use whichever format seems more appropriate to present the item effectively.

6. **The stem should not contain irrelevant or unnecessary detail.**
   Be sure that sufficient information is provided to answer the question and avoid excessive detail or "window dressing."

7. **The phrase *which of the following* should not be used to refer to the alternatives; instead, use *which followed* by a noun.**
   In the stem, *which of the following* requires the student to read all the alternatives before knowing what is being asked and assessed. Expressions such as *which statement*, *which expression*, *which equation*, and/or *which graph* are acceptable.

8. **The stem should include any words that must otherwise be repeated in each alternative.**
   In general, the stem should contain everything the alternatives have in common or as much as possible of their common content. This practice makes an item concise. Exceptions include alternatives containing units and alternatives stated as complete sentences.

9. **The item should have one and only one correct answer.**
   Items should not have two or more correct alternatives. *All the above* and *none of the above* are not acceptable alternatives.

10. **The distractors should be plausible and attractive to students who lack the knowledge, understanding, or ability assessed by the item.**
    Distractors should be designed to reflect common errors or misconceptions of students.

11. **The alternatives should be grammatically consistent with the stem.**
    Use similar terminology, phrasing or sentence structure in the alternatives. Alternatives must use consistent language, including verb tense, nouns, singular/plurals, and declarative statements. Place a period at the end of an alternative *only* if the alternative by itself is a complete sentence.

12. **The alternatives should be parallel with one another in form.**
    The length, complexity and specificity of the alternatives should be similar. For example, if the stem refers to a process, then all the alternatives must be processes. Avoid the use of absolutes such as *always* and *never* in phrasing alternatives.

13. **The alternatives should be arranged in logical order, when possible.**
    When the alternatives consist of numbers and letters, they should ordinarily be arranged in ascending or descending order. An exception would be when the number of an alternative and the value of that alternative are the same. For example: (1) 1 (2) 2 (3) 0 (4) 4.

14. **The alternatives should be independent and mutually exclusive.**
    Alternatives that are synonymous or overlap in meaning often assist the student in eliminating distractors.

15. **The item should not contain extraneous clues to the correct answer.**
    Any aspect of the item that provides an unintended clue that can be used to select or eliminate an alternative should be avoided. For example, any term that appears in the stem should not appear in only one of the alternatives.

16. **Notation and symbols as presented on examinations should be used consistently.**
    For example, *AB* means the length of line segment *AB*, $\overline{AB}$ means line segment *AB*, m∠*A* means the number of degrees in the measure of angle *A*, etc.

REVIEW CRITERIA CHECKLIST FOR POTENTIAL MATH ITEMS

The following list of criteria will be used to train item writers and then to review items for possible inclusion on test forms.

Language Appropriateness	Yes	No	N/A	Explain or Describe
1. Item: Uses grade-level vocabulary. Uses the simplest terms possible to convey information. Avoids technical terms unrelated to content.				
2. Sentence complexity well within grade expectations.				
3. Avoids ambiguous or double-meaning words.				
4. Pronouns have clear referents.				
5. Item avoids irregularly spelled words. Use most common spelling of words.				
6. Item can be put into Braille. Item can be translated appropriately according to the specific accommodations as outlined in universal design guidelines.				

Sensitivity/Bias	Yes	No	N/A	Explain or Describe
1. The item is free of content that might be deemed offensive to groups of students, based upon culture, religion, race, ethnicity, gender, geographic location, ability, socioeconomic status, etc.				
2. The item is free of content that contains stereotyping.				
3. The item is free of content that might unfairly advantage or disadvantage subgroups of students (ethnicity, gender, geographic location, ability, socioeconomic status, etc.) by containing unfamiliar contexts or examples, unusual names of people or places, or references to local events or issues.				

Math Art	Yes	No	N/A	Explain or Describe
1. The artwork clearly relates to the item and is important as an aspect of the problem-solving experience.				
2. The details in the artwork accurately and appropriately portray numbers/concepts contained in text or in lieu of text. Items should be drawn to scale as much as possible. By default, we do not include the text “Not drawn to scale” on every item; however, if a figure is drawn and there is a distortion in the figure, it should be indicated under the art that the figure is “not drawn to scale.” The degree of distortion should not be actively misleading.				
3. Graphics are clear (symbols are highly distinguished, free from clutter, at a reasonable scale, etc.).				
4. Visual load requirements are reasonable (interpreting graphic does not confuse underlying construct) and as simple as possible to present the prompt. “Visual load” refers to the amount of visual/graphic material included within a contained space. When graphics become overly busy, they break the cognitive process for different people or trip people up.				

Item Alignment	Yes	No	N/A	Explain or Describe
1. Is the item aligned to the standard to which it is written? List the primary standard to which the item is aligned and explain the degree to which there is alignment/lack of alignment.				
2. Is the item aligned to the correct secondary/tertiary standard(s)?				
3. The stem is reflective of the concept embedded within the standard and is representative of the goal of the standard.				

Item Alignment	Yes	No	N/A	Explain or Describe
4. The item requires students to show understanding of key aspects of the standard. If “No,” which aspects are not attended to? For constructed-response items, it is important that the item be solved through an understanding of the key point of the standard. For example, if the language of the standard calls for “prove” or “show,” items should actually involve proof to be aligned, not simply the ability to solve a related problem or perform a related manipulation.				
5. Does the question lend itself to being answered using a below-grade-level standard rather than the skills/concepts references in the on-grade-level standard?				
6. The item requires the student to use skills referenced in the primary standard and any additional standards listed.				
7. The item includes grade/course-appropriate standard numbers/variables (e.g., students are asked to solve questions using numbers/variables that are grade appropriate). Note: This includes the parameters outlined in the PARCC Pathways document for guidance on how some standards are split across A1 and A2.				
8. The item is aligned to the correct primary Multiple Representation(s). If “No,” indicate the correct MR code(s).				

Item Alignment	Yes	No	N/A	Explain or Describe
9. The item expects students to use a formula that is: - from a standard for an earlier grade level (i.e., prior knowledge); - part of the current mathematics curriculum; - not from another content area (e.g., physics). If "No," the formula should be in the item stem. For example, the formula for kinetic energy from physics should be included in the item stem.				

Application/Modeling Items	Yes	No	N/A	Explain or Describe
1. The item is aligned to a standard that requires modeling/application. Note: See starred items in CCSS for high school math. These items are identified as lending themselves to modeling.				
2. Does the language of the item obscure the match concept being assessed? Students should not stumble over irrelevant information.				
3. Modeling/application scenario is realistic and appropriate to the grade level (the situation is one that a reasonable person would encounter in everyday life—no stretching velvet ropes or weighing kittens in milligrams). If "No," explain why it's not.				
4. Standard does not call for modeling/application, but there is a reason for it to be represented as such. Even non-starred standards can and should involve appropriate applications where possible.				
5. Figures/numbers/concepts used in modeling/application as well as in the response are realistic (e.g., downloads cost 99 cents, the side of a house isn't 2x-32 long).				

Application/Modeling Items	Yes	No	N/A	Explain or Describe
6. Modeling scenario is presented in the most realistic and simple manner possible.				
7. Modeling/application scenario does not assume outside knowledge (e.g., approximate weight of paper, definition of a micron).				
8. Modeling/application scenario provides all necessary information for student to apply math concepts.				
9. Item does not clue students to which math strategy is needed to solve, but rather allows the student to choose a strategy to solve the item correctly. For example, we should not tell students to use Pythagorean theorem, but rather allow them to decide which approach to solving is appropriate.				

Mathematic Correctness	Yes	No	N/A	Explain or Describe
1. The stem addresses a central math concept, either implicitly or explicitly.				
2. The math presented in stem is clear, accurate, and conceptually plausible.				
3. At least one strategy exists that is on grade level to solve the problem.				
4. If there is more than one strategy, regardless of the strategy employed, the same correct answer will be achieved.				
5. There is a rationale for the correct response that is aligned to the language of the Standards and that demonstrates knowledge and/or application of the Standards.				
6. For MCQs: Is answer Choice 1 plausible or the correct answer? If not, why?				
7. For MCQs: Is answer Choice 2 plausible or the correct answer? If not, why?				

Mathematic Correctness	Yes	No	N/A	Explain or Describe
8. For MCQs: Is answer Choice 3 plausible or the correct answer? If not, why?				
9. For MCQs: Is answer Choice 4 plausible or the correct answer? If not, why?				

Constructed-Response and All Regents	Yes	No	N/A	Explain or Describe
1. The item involves a multi-step process.				
2. The item requires students to show work. Work referenced in item should not be trivial (e.g., if work was not shown, it would be likely that mistakes would be made).				
3. The item assesses more than computation.				
4. The item asks students to explain a concept or procedure used to solve the problem. Note: Not always applicable.				
5. If students are asked to describe what they did, clear direction is given as to what they should describe (the theory, the rationale for the answer, the reason a strategy is wrong, etc.).				
6. The item explicitly describes what we are trying to elicit from the students.				
7. The item is presented in a manner consistent with the Application MRs.				

Overarching Comments	Yes	No	N/A	Explain or Describe
1. The item is aligned to standard.				
2. The item is rigorous. The math should be sound, tight, challenging, and at the appropriate level of difficulty.				
3. The item is fair.				
4. The item is mathematically correct.				
5. The item is coded correctly for MR.				

Final Recommendation	Yes	No	N/A	Explain or Describe
1. Accept.				
2. Accept with Edits. Are suggested edits minor (will not impact statistics)? Note: Does not apply if at final typesetting phase.				
3. Reject.				

Guidelines for Writing Constructed-Response Math Items

1. **The item measures the knowledge, skills, and proficiencies characterized by the standards within the identified cluster.**

2. **The focus of the problem or topic should be stated clearly and concisely.**
   The item should be meaningful, address important knowledge and skills, and focus on key concepts.

3. **Include problems that come from a real-world context or problems that make use of multiple representations.**
   When using real-world problems, use formulas and equations that are real world *(e.g., the kinetic energy of an object with mass, m, and velocity, V is $k = \frac{1}{2} mv^2$).* Use real-world statistics whenever possible.

4. **The item should be written with terminology, vocabulary and sentence structure kept as simple as possible. The item should be free of irrelevant or unnecessary detail.**
   The important elements should generally appear early in the item, with qualifications and explanations following. Present only the information needed to make the context/scenario clear.

5. **The item should not contain extraneous clues to the correct answer.**
   The item should not provide unintended clues that allow a student to obtain credit without the appropriate knowledge or skill.

6. **The item should require students to demonstrate depth of understanding and higher-order thinking skills through written expression, numerical evidence, and/or diagrams.**
   An open-ended item should require more than an either/or answer or any variation such as yes/no, decrease/increase, and faster/slower. Often either/or items can be improved by asking for an explanation.

7. **The item should require work rather than just recall.**
   Students need to show their mathematical thinking in symbols or words.

8. **The stimulus should provide information/data that is/are mathematically accurate.**
   Examples of stimuli include, but are not limited to, art, data tables, and diagrams. It is best to use actual data whenever possible. Hypothetical data, if used, should be plausible and clearly identified as hypothetical.

9. **The item should be written so that the student does not have to identify units of measurement in the answer, unless the question is testing dimensional analysis.**
   For example, consider the question: “A circle has a radius of length 4 centimeters. Find the number of centimeters in the length of the arc intercepted by a central angle measuring 2 radians.” Students would receive credit for an answer of “8” and would not be penalized for writing “8 cm.”

10. The item should be written to require a specific form of answer.

Phrases like "in terms of π," "*to the nearest tenth*," and "in simplest radical form" may simplify the writing of the rubric for these types of items.

11. Items that require students to explain in words are encouraged.

One of the emphases of the New York learning standards is to foster student ability to communicate mathematical thinking. An example is to have students construct viable arguments such as to make conjectures, analyze situations or justify conclusions. These items would require students to demonstrate precision of knowledge in their responses.

12. Items may be broken into multiple parts that may be labeled a, b, c, etc.

Clear division of the parts of the problems may simplify the writing of the rubric for these types of items.

13. Notation and symbols as presented on examinations should be used consistently.

For example, *AB* means the length of line segment *AB*, $\overline{AB}$ means line segment *AB*, m∠*A* means the number of degrees in the measure of angle A, etc.