New York State Regents Examination in Geometry

2018 Technical Report

THE UNIVERSITY OF THE STATE OF NEW YORK
THE STATE EDUCATION DEPARTMENT
EXCELSIOR
1784

Prepared for the New York State Education Department
by Pearson

March 2019

Copyright

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

Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 INTRODUCTION

This technical report for the Regents Examination in Geometry will provide New York State with documentation of the purpose of the Regents Examination, scoring information, evidence of both reliability and validity of the exams, scaling information, and guidelines and reporting information for the August 2017, January 2018, and June 2018 administrations. Chapters 1–5 detail results for the June 2018 administration. Results for the August 2017 and January 2018 administrations are provided in Appendices D and E, respectively. 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" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p.123).[1] Please note that a technical report, by design, addresses technical documentation of a testing program; other aspects of a testing program (content standards, scoring guides, guide to test interpretation, equating, etc.) are thoroughly addressed and referenced in supporting documents.

The Regents Examination in Geometry is given August, January, and June to students enrolled in New York State schools. The examination is based on the Geometry Core Curriculum, which is based on the New York State Learning Standards.

1.2 PURPOSES OF THE EXAM (STANDARD 12.1)

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

1.3 TARGET POPULATION (STANDARD 7.2)

The examinee population for the Regents Examination in Geometry is composed of students who have completed a course of study in Geometry.

Table 1 provides a demographic breakdown of all students who took the August 2017, January 2018, and June 2018 Regents Examination in Geometry. All analyses in this report are based on the population described in Table 1. Annual Regents Examination results in the New York State Report Cards are those reported in the Student Information Repository System (SIRS) as of the reporting deadline (see http://data.nysed.gov/). If a student takes the same exam multiple times in the year, only the highest score is included in these results. 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 n-sizes in this technical report will differ from publicly reported counts of student test-takers.

Table 1 Total Examinee Population: Regents Examination in Geometry

	August Admin*		January Admin**		June Admin***	
Demographics	**Number**	**Percent**	**Number**	**Percent**	**Number**	**Percent**
All Students*	19,481	100.00	14,685	100.00	149,890	100.00
Race/Ethnicity						
American Indian/Alaska Native	129	0.66	121	0.82	891	0.59
Asian/Native Hawaiian/Other Pacific Islander	1,777	9.12	1,631	11.11	17,472	11.66
Black/African American	4,190	21.51	3,844	26.18	20,751	13.84
Hispanic/Latino	4,942	25.37	4,417	30.08	30,506	20.35
Multiracial	312	1.60	240	1.63	2,503	1.67
White	8,127	41.73	4,431	30.18	77,765	51.88
English Language Learner/Multilingual Learner						
No	18,998	97.52	13,816	94.08	144,519	96.42
Yes	483	2.48	869	5.92	5,371	3.58
Economically Disadvantaged						
No	9,776	50.18	5,544	37.75	86,704	57.85
Yes	9,705	49.82	9,141	62.25	63,186	42.15
Gender						
Female	10,367	53.23	7,808	53.17	78,767	52.55
Male	9,110	46.77	6,876	46.83	71,121	47.45
Student with a Disability						
No	17,822	91.48	13,345	90.88	139,526	93.09
Yes	1,659	8.52	1,340	9.12	10,364	6.91

*Note: Four students were not reported in the Ethnicity and Gender group, but they are reflected in “All Students.”

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

***Note: Two students were not reported in the Ethnicity and Gender group, 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 only to the operational Regents Examination in Geometry items.

2.1 ITEM DIFFICULTY

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

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

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

Although the *p*-value statistic does not consider individual student ability in its computation, it provides a useful view of overall item difficulty and can provide an early and simple indication of items that are too difficult for the population of students taking the examination. Items with very high or very low *p*-values receive added scrutiny during all follow-up analyses, including item response theory analyses that factor student ability into estimates of item difficulty. Such items may be removed from the item pool during the test development process, as field testing typically reveals that they add insufficient measurement information. Items for the June 2018 Regents Examination in Geometry show a range of *p*-values consistent with the targeted exam difficulty. Item *p*-values, presented in Table 2 and Table 3 for multiple-choice and constructed-response items, respectively, range from 0.26 to 0.85, with a mean of 0.54. Table 2 and Table 3 also show a standard deviation (SD) of item score and item mean (Table 3, only).

2.2 ITEM DISCRIMINATION

At the most general level, estimates of item discrimination indicate each item's ability to differentiate between high and low student performance. It is expected that students who perform well on the Regents Examination in Geometry would be more likely to answer any given item correctly, while low-performing students (i.e., those who perform poorly on the exam

[2] For MC items with four response options, pure random guessing would lead to an expected *p*-value of 0.25.

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 right 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 zero), meaning the item is likely discriminating well between high- and low-performing students. Point-biserials are computed for each answer option, including correct and incorrect options (commonly referred to as "distractors"). Finally, point-biserial values for each distractor are an important part of test analysis. The point-biserial values on distractors are typically negative. Positive point-biserial 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 and Table 3 provide the point-biserial values on the correct response and three distractors (Table 2, only) for the June 2018 administration of the Regents Examination in Geometry. The point-biserial values for correct answers are 0.36 or higher, indicating acceptable discrimination between high- and low-performing examinees. Point-biserial values for all are negative or very close to zero. This indicates that examinees are responding to the items as expected during item and rubric development.

Table 2 Multiple-Choice Item Analysis Summary: Regents Examination in Geometry

Item	Number	*p*-Value	SD	Point-Biserial	Point-Biserial Distractor 1	Point-Biserial Distractor 2	Point-Biserial Distractor 3
1	149,890	0.84	0.37	0.39	-0.16	-0.17	-0.30
2	149,890	0.65	0.48	0.51	-0.30	-0.25	-0.23
3	149,890	0.85	0.35	0.36	-0.14	-0.16	-0.28
4	149,890	0.82	0.39	0.40	-0.14	-0.24	-0.26
5	149,890	0.72	0.45	0.54	-0.20	-0.41	-0.20
6	149,890	0.70	0.46	0.54	-0.19	-0.39	-0.22
7	149,890	0.63	0.48	0.42	-0.21	-0.28	-0.17
8	149,890	0.56	0.50	0.42	-0.27	-0.22	-0.17
9	149,890	0.62	0.49	0.40	-0.27	-0.19	-0.14
10	149,890	0.68	0.47	0.51	-0.26	-0.26	-0.28
11	149,890	0.56	0.50	0.47	-0.21	-0.26	-0.22
12	149,890	0.61	0.49	0.52	-0.23	-0.36	-0.20
13	149,890	0.58	0.49	0.52	-0.20	-0.36	-0.18
14	149,890	0.71	0.46	0.50	-0.20	-0.31	-0.26
15	149,890	0.45	0.50	0.38	-0.24	-0.16	-0.13
16	149,890	0.45	0.50	0.49	-0.18	-0.31	-0.19
17	149,890	0.36	0.48	0.36	-0.21	-0.16	-0.17
18	149,890	0.39	0.49	0.38	-0.22	-0.12	-0.15
19	149,890	0.50	0.50	0.45	-0.24	-0.17	-0.19

Item	Number	*p*-Value	SD	Point-Biserial	Point-Biserial Distractor 1	Point-Biserial Distractor 2	Point-Biserial Distractor 3
20	149,890	0.44	0.50	0.48	-0.16	-0.32	-0.12
21	149,890	0.51	0.50	0.40	-0.27	-0.17	-0.20
22	149,890	0.33	0.47	0.45	-0.35	-0.15	-0.01
23	149,890	0.31	0.46	0.46	-0.17	-0.17	-0.17
24	149,890	0.29	0.45	0.45	0.04	-0.40	-0.12

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

Item	Min. score	Max. score	Number of Students	Mean	SD	*p*-Value	Point-Biserial
25	0	2	149,890	1.55	0.69	0.78	0.53
26	0	2	149,890	0.98	0.83	0.49	0.61
27	0	2	149,890	1.41	0.69	0.70	0.47
28	0	2	149,890	0.88	0.84	0.44	0.61
29	0	2	149,890	0.92	0.77	0.46	0.65
30	0	2	149,890	0.90	0.82	0.45	0.65
31	0	2	149,890	0.94	0.80	0.47	0.69
32	0	4	149,890	2.06	1.77	0.52	0.75
33	0	4	149,890	1.79	1.79	0.45	0.78
34	0	4	149,890	1.35	1.23	0.34	0.74
35	0	6	149,890	1.54	1.89	0.26	0.76

2.3 DISCRIMINATION ON DIFFICULTY SCATTER PLOT

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

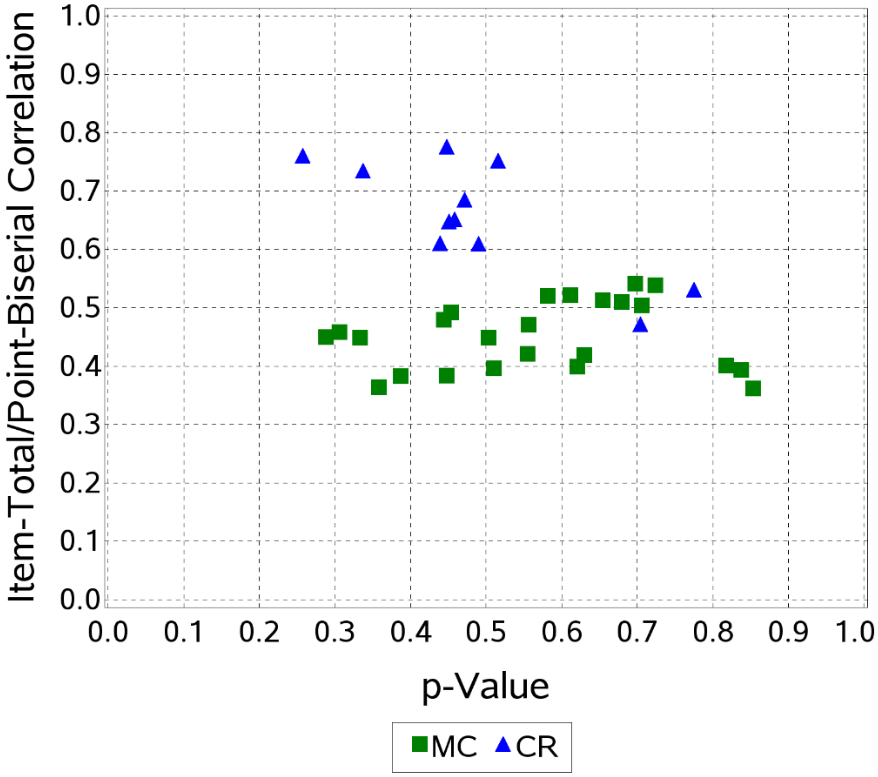


Figure 1 Scatter Plot: Regents Examination in Geometry

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

Statistics	N	Mean	Min	Q1	Median	Q3	Max
p-value	35	0.54	0.26	0.44	0.51	0.68	0.85
Point-Biserial	35	0.52	0.36	0.42	0.49	0.61	0.78

2.4 OBSERVATIONS AND INTERPRETATIONS

The *p*-values for the MC items ranged from about 0.29 to 0.85 while proportion-correct values for the constructed-response items (Table 3) ranged from about 0.26 and 0.78. The difficulty distribution illustrated in Figure 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.

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

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

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

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

3.1 DESCRIPTION OF THE RASCH MODEL

The Rasch model (Rasch, 1960) was used to calibrate multiple-choice items, and the partial credit model, or PCM (Wright and Masters, 1982), was used to calibrate constructed-response 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 by

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

where θ_n represents examinee 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:

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

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

3.2 SOFTWARE AND ESTIMATION ALGORITHM

Item calibration was implemented via the WINSTEPS 3.60 computer program (Wright and Linacre, 2015), which employs unconditional (UCON), joint maximum likelihood estimation (JMLE).

3.3 CHARACTERISTICS OF THE TESTING POPULATION

The data analyses reported here are based on all students who took the Regents Examination in Geometry in June 2018. The characteristics of this population are provided in Table 1.

3.4. ITEM DIFFICULTY-STUDENT PERFORMANCE MAPS

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 2. 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 examinee performance at the top and lower performance and easier items at the bottom.

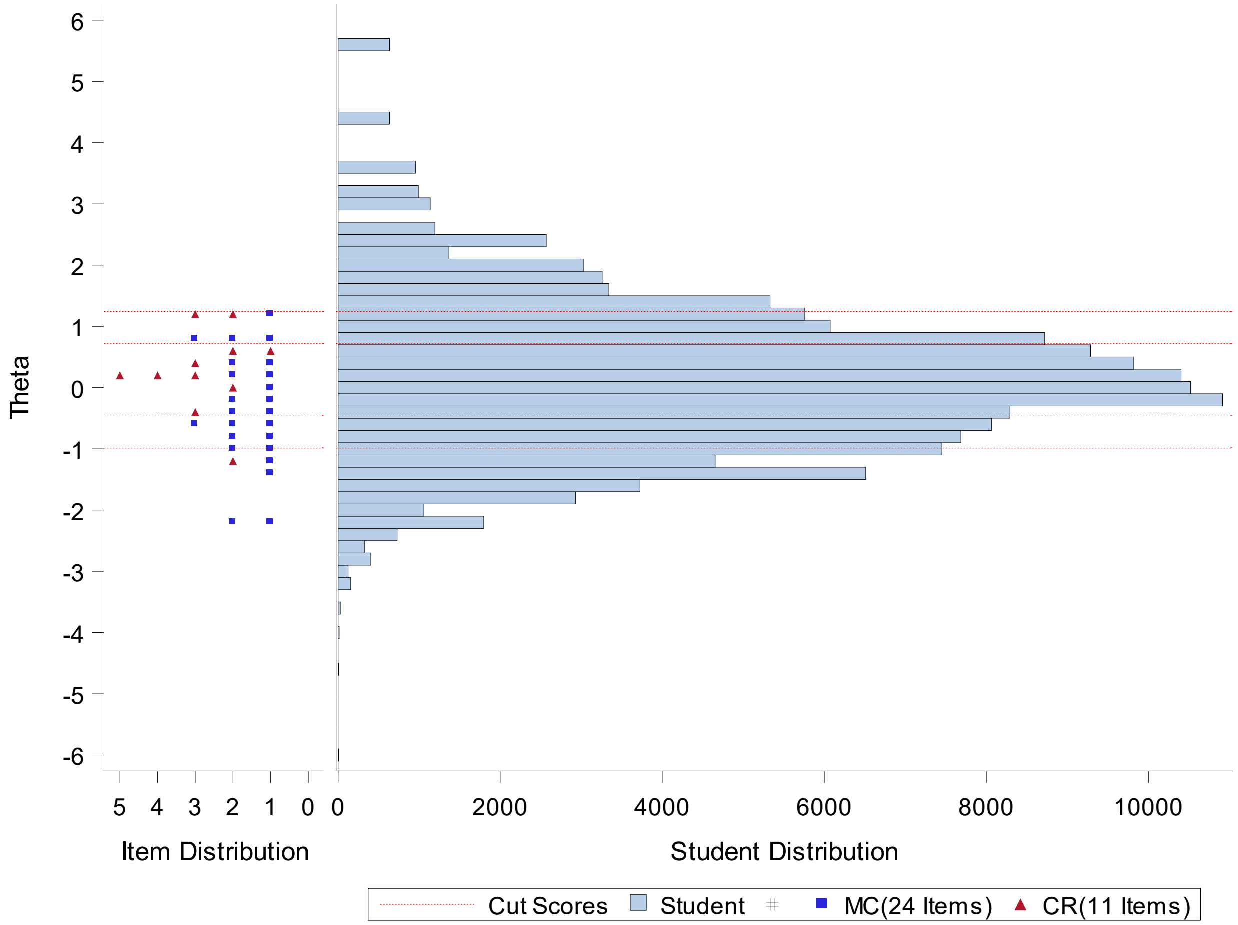


Figure 2 Student Performance Map: Regents Examination in Geometry

3.5 CHECKING RASCH ASSUMPTIONS

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

Unidimensionality

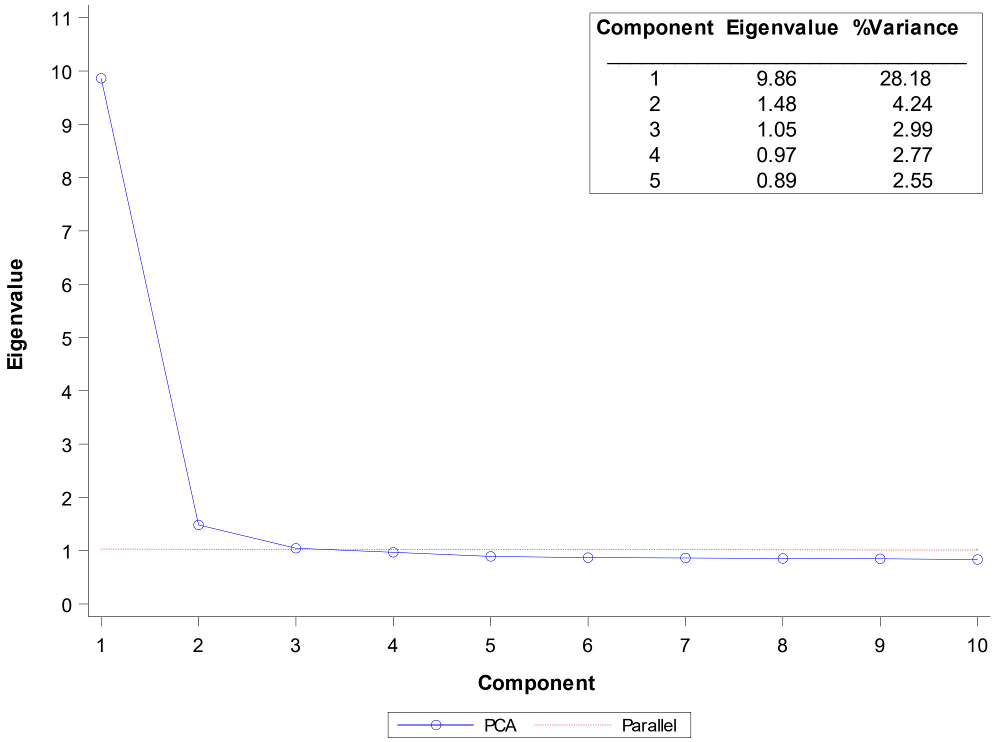
Rasch models assume that one dominant dimension determines the differences in students' performances. Principal Components Analysis (PCA) can be used to assess the unidimensionality assumption. The purpose of the analysis is to verify if other dominant components exist among the items. If any other dimensions are found, the unidimensionality of test content 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 to decide how many factors exist in principal components. For the parallel analysis, 100 random data sets of sizes equal to the original data were created. For each random data set, a PCA was performed and the resulting eigenvalues stored. Then for each component, the upper 95th percentile value of the distribution of the 100 eigenvalues from the random data sets was 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 shows the PCA and parallel analysis results for the Regents Examination in Geometry. 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 from a parallel analysis. Although the total number of components in PCA is same as the total number of items in a test, Figure 3 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 through 10 are much lower than the first component demonstrates that there is only one dominant component, showing evidence of unidimensionality.

As rule of thumb, Reckase (1979) proposed that the variance explained by the primary dimension should be greater than 20 percent 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, the primary dimension explained 28.18 percent of the total variance for the Regents Examination in Geometry. The eigenvalue of the second dimension is less than one-third of the first, at 1.48, and the second value is not significantly different from the parallel analysis. Overall, the PCA suggests that the test is reasonably unidimensional.



Component	Eigenvalue	%Variance
1	9.86	28.18
2	1.48	4.24
3	1.05	2.99
4	0.97	2.77
5	0.89	2.55

Figure 3 Scree Plot: Regents Examination in Geometry

Local Independence

Local independence (LI) is a fundamental assumption of IRT. This means that, for statistical purposes, an examinee's response to any one item should not depend on the examinee's response to any other item on the test. In formal statistical terms, a test *X* that is comprised of items X_1, X_2,…X_n is locally independent with respect to the latent variable *θ* if, for all $x = (x_1, x_2, \ldots x_n)$ 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 examinee'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 actually framed by WLI. For WLI, the conditional covariances of all pairs of item responses, conditioned on the abilities, are assumed to be equal to zero. When this assumption is met, the joint probability of responses to an item pair, conditioned on the abilities, is the product of the probabilities of responses to these two items, as shown below. Based on the WLI, the following expression can be derived:

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

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 Geometry 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 examinees’ 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. It is noted that the raw score residual correlation essentially corresponds to Yen’s Q_3 index, a popular statistic used to assess local independence. The expected value for the Q_3 statistic is approximately −1/(*k* − 1) when no local dependence exists, where *k* is test length (Yen, 1993). Thus, the expected Q_3 values should be approximately −0.03 for the items on the exam. Index values that are greater than 0.20 indicate a degree of local dependence that probably should be examined by test developers (Chen & Thissen, 1997).

Since the three residual correlations are very similar, the default “standardized residual correlation” in WINSTEPS was used for these analyses. Table 5 shows the summary statistics — mean, standard deviation, minimum, maximum, and several percentiles (P_{10}, P_{25}, P_{50}, P_{75}, P_{90}) — for all of the residual correlations for each test. The total number of item pairs (N) and the number of pairs with the residual correlations greater than 0.20 are also reported in this table. There were no item pairs with residual correlations greater than 0.20. The mean residual correlation was slightly negative, and the values were close to −0.02. The vast majority of the correlations were very small, suggesting that local item independence generally holds for the Regents Examination in Geometry.

Table 5 Summary of Item Residual Correlations: Geometry

Statistic Type	Value
N	595
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	0.06
>|0.20|	0

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.

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 either 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 6 presents the summary statistics of INFIT mean square statistics for the Regents Examination in Geometry, including the number of items, mean, standard deviation, and minimum and maximum values.

The number of items within a targeted range of [0.7, 1.3] is also reported in Table 6. 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 guide for ideal fit, fit values outside of the range are considered individually. A finding of all items falling in the ideal fit range indicates that the Rasch model fits the Regents Examination in Geometry item data well.

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

	INFIT Mean Square					
	N	Mean	SD	Min	Max	[0.7, 1.3]
Geometry	35	1.00	0.08	0.84	1.14	[35/35]

Items for the Regents Examination in Geometry were field tested in 2012–2017, and separate technical reports for each year were produced to document the full test development, scoring, scaling, and data analysis conducted.

3.6 SCALING OF OPERATIONAL TEST FORMS

Operational test items were selected based on content coverage, content accuracy, and statistical quality. The sets of items on each operational test conformed to the coverage determined by content experts working from the learning standards established by the New York State Education Department and explicated in the test blueprint. Each item's classical and Rasch statistics were used to assess item quality. Items were selected to vary in difficulty to accurately measure students' abilities across the ability continuum. Appendix A contains the operational test maps for the August 2017, January 2018, and June 2018 administrations. Note that 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 across the state. The small n-counts associated with such administrations are sufficient for reasonably accurate estimation of most items' parameters; however, for the six-point essay item, its parameters can be unstable when estimated across as small a sample as is typically used. Therefore, a set of constants is used for these items' parameters on operational examinations. These constants were set by NYSED and are based on the values in the bank for all essay items. For the August 2017 Geometry examination, there are two six-point items with fixed constants as follows: First six-point item ($D = 1.13$, $F_0 = 0.00$, $F_1 = 0.22$, $F_2 = -0.70$, $F_3 = -0.23$, $F_4 = -0.34$, $F_5 = 0.33$, and $F_6 = 0.73$); second six-point item ($D = 0.79$, $F_0 = 0.00$, $F_1 = -0.42$, $F_2 = 0.11$, $F_3 = -0.11$, $F_4 = 0.18$, $F_5 = 0.41$, and $F_6 = -0.17$). Starting in January 2018, only one six-point item appears on the Geometry examination with fixed constants as follows: $D = 1.13$, $F_0 = 0.00$, $F_1 = 0.22$, $F_2 = -0.70$, $F_3 = -0.23$, $F_4 = -0.34$, $F_5 = 0.33$, and $F_6 = 0.73$. For June 2018, the fixed constants for the six-point item were updated as follows: $D = 1.16$, $F_0 = 0.00$, $F_1 = -0.13$, $F_2 = -0.10$, $F_3 = -0.01$, $F_4 = -0.14$, $F_5 = 0.18$, and $F_6 = 0.21$.

The New York State Regents Examination in Geometry has four cut scores, which are set at the scale scores of 55, 65, 80 (floating), and 85. One of the primary considerations during test construction was to select items so as to minimize changes in the raw scores corresponding to these 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 0.125. It should be noted that the raw scores corresponding to the scale score cut scores may still fluctuate, even if the mean Rasch difficulty level is maintained at the target value, due to differences in the distributions of the Rasch difficulty values among the items from administration to administration.

The relationship between raw and scale scores is explicated in the scoring tables for each administration. These tables for the August 2017, January 2018, and June 2018 administrations can be found in Appendix B. These tables are the end product of the following scaling procedure.

All Regents Examinations are equated back to a base scale, which is held constant from year to year. Specifically, they are equated to the base scale through the use of a calibrated item pool. The Rasch difficulties from the items' initial administration in a previous year's field test are used to equate the scale for the current administration to the base administration. For this examination, the base administration was the June 2015 administration. Scale scores from the August 2017, January 2018, and June 2018 administrations are on the same scale and can be directly compared to scale scores on all previous administrations back to the June 2015 administration.

When the base administration was concluded, the initial raw score to scale score relationship was established. Three raw scores were fixed at specific scale scores. Scale scores of 0 and 100 were fixed to correspond to the minimum and maximum possible raw scores. In addition, a standard setting had been held to determine the passing and passing with distinction cut scores in the raw score metric. The scale score points of 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:

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

where SS is the scaled score, RS is the raw score, and m0 through m4 are the transformation constants that convert the raw score into the scale score (please note that m0 will always be equal to zero in this application, since a raw score of zero corresponds to a scale score of zero). A subscript for a person on both dependent and independent variables is not present for simplicity. The above relationship and the values of m1 to m4 specific to this subject were then used to determine the scale scores corresponding to the remainder of the raw scores on the examination. This initial relationship between the raw and scale scores became the base scale.

The Rasch difficulty parameters for the items on the base form were then used to derive a raw score to Rasch student ability (theta score) relationship. This 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 (zero) and maximum possible raw scores are assigned scale scores of 0 and 100, respectively. In the event that there are raw scores less than the maximum with scale scores that round to 100, their scale scores are set equal to 99. A similar process is followed with the minimum score; if any raw scores other than zero have scale scores that round to zero, their scale scores are instead set equal to one.

With regard to the cuts, if two or more scale scores round to 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 cut at a scale score of 80. If no scale score rounds to these four critical cuts, then the raw score with the largest scale score that is less than the cut is set equal to the cut. The overarching principle, when two raw scores both round to either scale score cut, is that the lower of the raw scores is always assigned to be equal to the cut so that students are never penalized for this ambiguity.

Chapter 4: Reliability (Standard 2)

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

According to the *Standards for Educational and Psychological Testing*, “A number of factors can have significant effects on reliability/precision, and in some cases, these factors can lead to misinterpretations of test scores, if not taken into account” (AERA et al., 2014, p. 38). First, test length and the variability of observed scores can both influence reliability estimates. Tests with fewer items or with a lack of heterogeneity in scores tend to produce lower reliability estimates. Second, reliability is specifically 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. Of course, systematic error sources also exist.

The remainder of this chapter discusses reliability results for the Regents Examination in Geometry and three additional statistical measures to address the multiple factors affecting an interpretation of the exam’s reliability:

- standard errors of measurement
- decision consistency
- 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). Put differently, 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).

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

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

Coefficient Alpha

Reliability is most often estimated using the formula for Coefficient Alpha, which provides a practical internal consistency index. 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. Note that relative error is reflected in this index. Excessive variation in student performance from one sample of items to the next should be of particular concern for any achievement test user.

A general computational formula for Coefficient Alpha is as follows:

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

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), discussed further below.

Traditional Standard Error of Measurement

The standard error of measurement is defined as the standard deviation of the distribution of observed scores for students with identical true scores. Because the SEM is an index of the random variability in test scores in test score units, it represents important information for test score users.

The SEM formula is provided below.

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

This formula indicates that the value of the SEM depends on both the reliability coefficient (the Coefficient Alpha, as detailed previously) 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. In other words, a perfectly reliable test has no measurement error (Harvill, 1991). Additionally, the value of the SEM takes the group variation (i.e., score standard deviation) into account. Consider that a SEM of 3 on a 10-point test would be very different from a SEM of 3 on a 100-point test.

Traditional Standard Error of Measurement Confidence Intervals

The SEM is an index of the random variability in test scores reported in actual score units, which is why it has such great utility for test score users. SEMs allow statements regarding the precision of individual test scores. SEMs 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. As an example, students with a given true score will have observed scores that fall between ±1 SEM about two-thirds of the time.[4] For ±2 SEM confidence intervals, this increases to about 95 percent.

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

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

Subject	Coefficient Alpha	SEM
Geometry	0.91	5.51

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

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

Conditional Standard Error of Measurement

Every time an assessment is administered, the score the student receives 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 their true score (*θ*), the score obtained with no error), and the standard deviation of the distribution of their raw scores would be the conditional standard error. Since there is a one-to-one correspondence between the raw score and *θ* in the Rasch model, we can apply this concept more generally to all students who obtained a particular raw score and calculate the probability of obtaining each possible raw score, given the students' estimated *θ*. The standard deviation of this conditional distribution is defined as the conditional standard error of measurement (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 third-degree polynomial relationship between the raw and

[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 percent of the time.

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 third-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 that are 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, Zeng, and Hanson (1996) for dichotomously scored items and extended by Wang, Kolen, and Harris (2000) to polytomously scored items. For tests such as the Regents Examination in Geometry 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 an examinee with a certain performance level. If it were possible to measure this examinee's performance perfectly, without any error, this measure could be called the examinee's "true score," as discussed earlier. This score is equal to the expected raw score. However, whenever an examinee takes a test, the observed test score always includes some level of measurement error. Sometimes this error is positive, and the examinee achieves a higher score than would be expected given his or her level of *θ*; other times it is negative, and the examinee achieves a lower-than-expected score. If we could give an examinee the same test multiple times and record the observed test scores, the resulting distribution would be the conditional distribution of raw scores for that examinee's level of *θ* with a mean value equal to the examinee'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, Pommerich, Billeaud, and Williams (1995) and is also implemented in the computer program POLYCSEM (Kolen, 2004). The compound multinomial distribution yields the probabilities that an examinee with a given level of *θ* has of achieving 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*.

Conditional Standard Error of Measurement Confidence Intervals

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

Conditional Standard Error of Measurement Characteristics

The relationship between the scale score CSEM and *θ* depends both on the nature of the raw-to-scale score transformation (Kolen and Brennan, 2005; Kolen and Lee, 2011) and on 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 (2005, p. 357) state, "When, relative to raw scores, the transformation compresses the scale in the middle and stretches it at the ends, the pattern of the conditional standard errors of measurement will be concave up (U-shaped), even though the pattern for the raw scores was concave down (inverted-U shape)."

Results and Observations

The relationship between raw and scale scores for the Regents 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 about the lower 80 percent of the scale score range and then are compressed relative to the raw scores for about the remaining 20 percent 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 shows this type of CSEM variation for the Regents Examination in Geometry where the compression of raw score to scale scores between the cut scores of 65 and 85 changes the shape of the curve noticeably. This type of expansion and compression can be seen in Figure 4 by looking at the changing density of raw score points along the scale score range on the horizontal axis. Specifically, the raw scores are expanded up to a scale score of about 65 followed by very noticeable compression through a scale score of about 95.

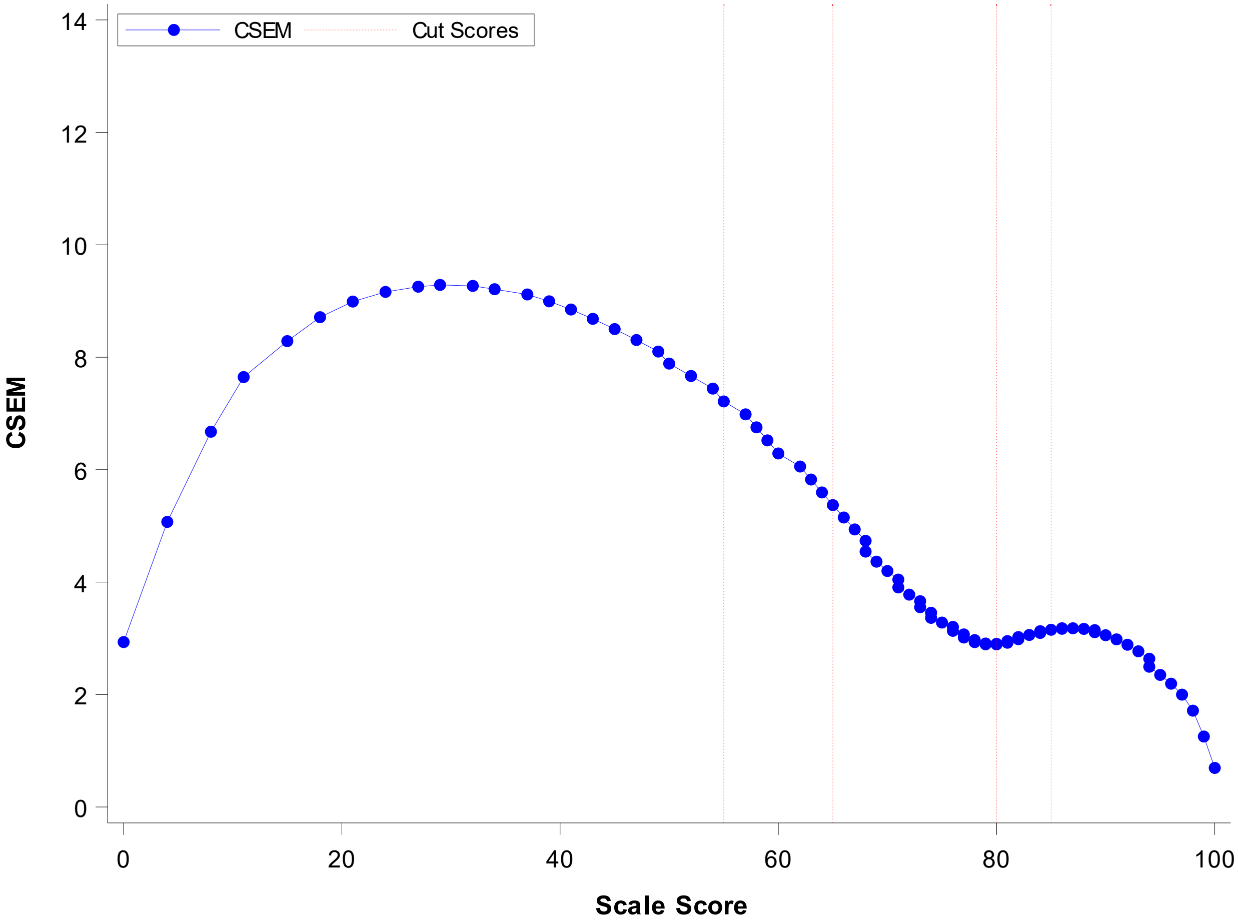


Figure 4 Conditional Standard Error Plot: Regents Examination in Geometry

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, which 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 achievement level for each student can be replicated upon retesting using an equivalent form (Huynh, 1976). Decision consistency answers the following question: What is the agreement in classifications between the two non-overlapping, 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 the following tables.

		TEST ONE		
		LEVEL I	LEVEL II	MARGINAL
TEST TWO	LEVEL I	φ11	φ12	φ1●
	LEVEL II	φ21	φ22	φ2●
	MARGINAL	φ●1	φ●2	1

Figure 5 Pseudo-Decision Table for Two Hypothetical Categories

		TEST ONE				
		LEVEL I	LEVEL II	LEVEL III	LEVEL IV	MARGINAL
TEST TWO	LEVEL I	φ11	φ12	φ13	φ14	φ1●
	LEVEL II	φ21	φ22	φ23	φ24	φ2●
	LEVEL III	φ31	φ32	φ33	φ34	φ3●
	LEVEL IV	φ41	φ42	φ43	φ44	φ4●
	MARGINAL	φ●1	φ●2	φ●3	φ●4	1

Figure 6 Pseudo-Decision Table for Four Hypothetical Categories

If a student is classified as being in one category based on Test One's score, how probable would it be that the student would be reclassified as being in the same category if he or she took Test Two (a non-overlapping, 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 two formulas, respectively:

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

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

The sum of the diagonal entries — that is, the proportion of students classified by the two forms into exactly the same achievement level — signifies the overall consistency.

Classification accuracy refers to the agreement of the observed classifications of students with the classifications made on the basis of their true scores. As discussed above, 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 his or her true score plus measurement error, or $Observed = True + Error$. Decision accuracy is an index to determine the extent to which measurement error causes a classification different from the one expected from the true score.

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

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

Results and Observations

The results for the dichotomies created by the four cut scores are presented in Table 8. The tabled values are derived with the program *BB-Class* (Brennan, 2004) using the Livingston and Lewis method. The overall decision consistency ranged from 0.88 to 0.92, and the decision accuracy ranged from 0.92 to 0.94. Both decision consistency and accuracy values indicate good consistency and accuracy of examinee classifications.

Table 8 Decision Consistency and Accuracy Results: Regents Examination in Geometry

Statistic	1/2	2/3	3/4	4/5
Consistency	0.90	0.88	0.90	0.92
Accuracy	0.93	0.92	0.93	0.94

4.4 GROUP MEANS (STANDARD 2.17)

Mean and SD scale scores were computed based on reported gender, race/ethnicity, English language learner/multilingual learner status, economically disadvantaged status, and student with a disability status. The results are reported in Table 9.

Table 9 Group Means: Regents Examination in Geometry

Demographics	Number	Mean Scale Score	SD Scale Score
All Students*	149,890	70.97	15.71
Ethnicity			
American Indian/Alaska Native	891	66.34	15.73
Asian/Native Hawaiian/Other Pacific Islander	17,472	77.87	14.94
Black/African American	20,751	60.05	15.62
Hispanic/Latino	30,506	63.50	15.78
Multiracial	2,503	72.16	14.99
White	77,765	75.29	13.02
English Language Learner/Multilingual Learner			
No	144,519	71.50	15.34
Yes	5,371	56.77	18.63
Economically Disadvantaged			
No	86,704	75.05	13.98
Yes	63,186	65.39	16.24
Gender			
Female	78,767	71.55	15.56
Male	71,121	70.34	15.86
Student with a Disability			
No	139,526	72.05	15.08
Yes	10,364	56.54	16.89

*Note: Two students were not reported in the Ethnicity and Gender group, but they are reflected in “All Students.”

4.5 STATE PERCENTILE RANKINGS

State percentile rankings based on raw score distributions are noted in Table 10. The percentiles are based on the distribution of all students taking the Regents Examination in Geometry for the June 2018 administration. Note that the scale scores for the Regents Examination range from 0 to 100, but 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 10 State Percentile Ranking for Raw Score: Regents Examination in Geometry

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

Chapter 5: Validity (Standard 1)

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

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

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

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

5.1 EVIDENCE BASED ON TEST CONTENT

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

The content standards associated with Geometry are based on the New York State Learning Standards for Mathematics and the PARCC Model Content Framework for Geometry located at https://www.engageny.org/resource/new-york-state-p-12-common-core-learning-standards-for-mathematics.

The content standards define what students should understand and be able to do at the high school level; the Model Content Framework describes which content is included and emphasized within the Geometry course, specifically. More information about the relationship between the NYS CCLS and the PARCC Model Content Frameworks can be found in https://www.engageny.org/file/741/download/nysp12cclsmath.pdf and https://www.engageny.org/resource/parcc-model-content-frameworks-for-educators.

Content Validity

Content validity is necessarily concerned with the proper definition of the construct and evidence that the test provides an accurate measure of examinee performance within the defined construct. The test blueprint for the Regents Examination in Geometry is essentially the design document for constructing the exam. It provides an explicit definition of the content domain that is to be represented on the exam. The test development process (discussed in the next section) is in place to ensure, to the extent possible, that the blueprint is met in all operational forms of the exam.

Table 11 displays the Conceptual category, domains, and target percent of each for the Regents Examination in Geometry. Geometry is associated with the high school content standards within the *conceptual category* of Geometry. This conceptual category contains *domains* of related *clusters* of standards. This chart shows the high school mathematics domains included in Geometry, as well as the corresponding percent of credits on the Regents Examination in Geometry.

Table 11 Test Blueprint, Regents Examination in Geometry

Conceptual Category	Percent of Test By Credit	Domains in Geometry
Geometry	27–34%	Congruence (G-CO)
	29–37%	Similarity, Right Triangles, and Trigonometry (G-SRT)
	2–8%	Circles (G-C)
	12–18%	Expressing Geometric Properties with Equations (G-GPE)
	2–8%	Geometric Measurement and Dimensions (G-GMD)
	8–15%	Modeling with Geometry (G-MG)

Item Development Process

Test development for the Regents Examination in Geometry 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 New York State 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 New York State-certified educators may participate in this process. The New York State Education Department 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 around the state are retained to write all items for the Regents Examination in Geometry, under strict guidelines that leverage best practices (see Appendix C). State educators also conduct all item quality and bias reviews, in order 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, in order to select the highest quality items for use in the operational test.

Figure 7 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.

Initial item development is conducted under the criteria and guidance provided by multiple documents, including the blueprint noted in Table 10 and Item Writing Guidelines noted in Appendix A. To facilitate the alignment of items during development with standards, Standards Interpretations are also provided to developers. These interpretations are noted in Appendix B. Both multiple-choice and constructed-response items are included in the Regents Examination in Geometry, in order to ensure appropriate coverage of the construct domain.

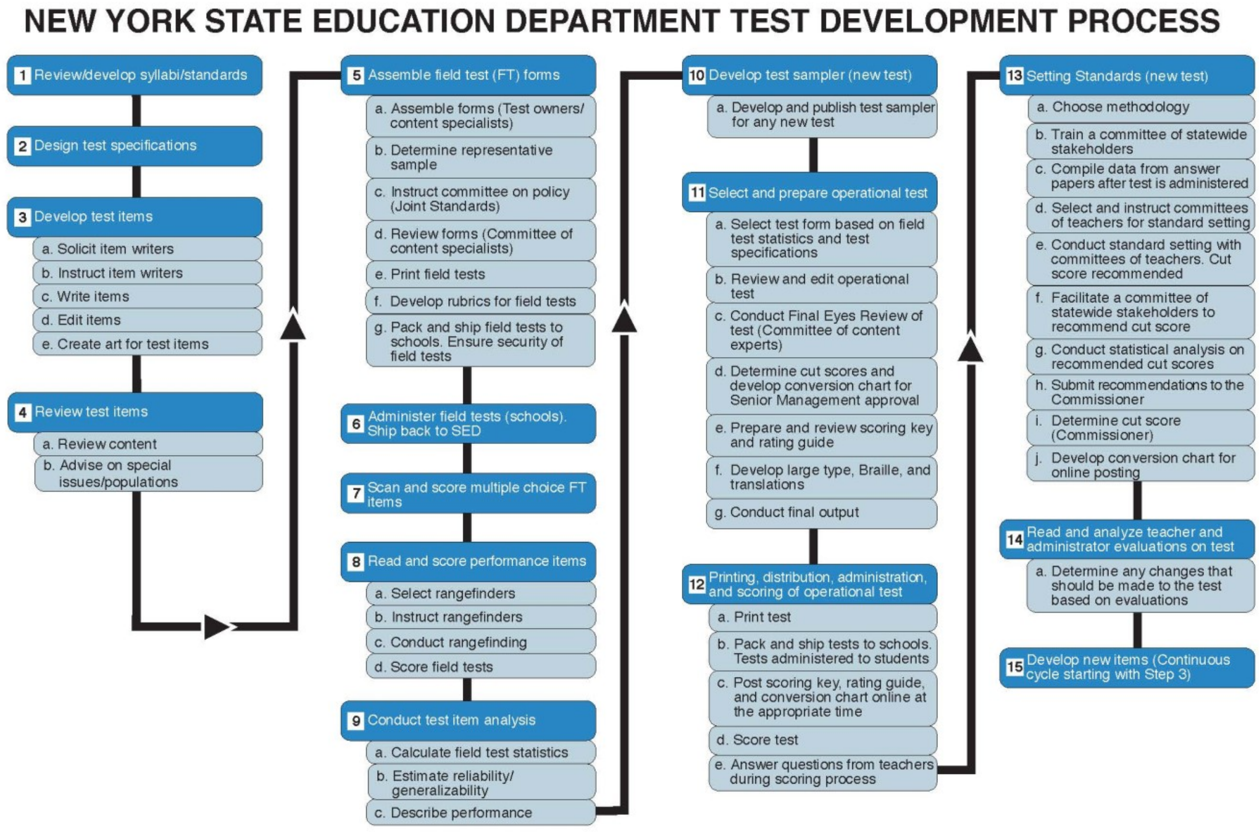

NEW YORK STATE EDUCATION DEPARTMENT TEST DEVELOPMENT PROCESS
1 Review/develop syllabi/standards
2 Design test specifications
3 Develop test items
a. Solicit item writers
b. Instruct item writers
c. Write items
d. Edit items
e. Create art for test items
4 Review test items
a. Review content
b. Advise on special issues/populations
5 Assemble field test (FT) forms
a. Assemble forms (Test owners/ content specialists)
b. Determine representative sample
c. Instruct committee on policy (Joint Standards)
d. Review forms (Committee of content specialists)
e. Print field tests
f. Develop rubrics for field tests
g. Pack and ship field tests to schools. Ensure security of field tests
6 Administer field tests (schools). Ship back to SED
7 Scan and score multiple choice FT items
8 Read and score performance items
a. Select rangefinders
b. Instruct rangefinders
c. Conduct rangefinding
d. Score field tests
9 Conduct test item analysis
a. Calculate field test statistics
b. Estimate reliability/ generalizability
c. Describe performance
10 Develop test sampler (new test)
a. Develop and publish test sampler for any new test
11 Select and prepare operational test
a. Select test form based on field test statistics and test specifications
b. Review and edit operational test
c. Conduct Final Eyes Review of test (Committee of content experts)
d. Determine cut scores and develop conversion chart for Senior Management approval
e. Prepare and review scoring key and rating guide
f. Develop large type, Braille, and translations
g. Conduct final output
12 Printing, distribution, administration, and scoring of operational test
a. Print test
b. Pack and ship tests to schools. Tests administered to students
c. Post scoring key, rating guide, and conversion chart online at the appropriate time
d. Score test
e. Answer questions from teachers during scoring process
13 Setting Standards (new test)
a. Choose methodology
b. Train a committee of statewide stakeholders
c. Compile data from answer papers after test is administered
d. Select and instruct committees of teachers for standard setting
e. Conduct standard setting with committees of teachers. Cut score recommended
f. Facilitate a committee of statewide stakeholders to recommend cut score
g. Conduct statistical analysis on recommended cut scores
h. Submit recommendations to the Commissioner
i. Determine cut score (Commissioner)
j. Develop conversion chart for online posting
14 Read and analyze teacher and administrator evaluations on test
a. Determine any changes that should be made to the test based on evaluations
15 Develop new items (Continuous cycle starting with Step 3)

Figure 7 New York State Education Department Test Development Process

Item Review Process

The item review process helps to ensure 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. This process allows high-quality items to be continually developed in a manner that is consistent with the test blueprint. Item review guidelines for multiple-choice items are included in Appendix C.

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 examinees 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 examinee is being asked to demonstrate by responding to each item. Each of these elements of the review process is in place, ultimately, to target fairness for all students by targeting consistency in examinee scores and providing evidence of the validity of their interpretations.

Specifically, the item review process articulates the four major item characteristics that the New York State Education Department looks for when developing quality items:

1. language and graphical appropriateness
2. sensitivity/bias
3. fidelity of measurement to standards
4. conformity to the expectations for the specific item types and formats (e.g., multiple-choice questions, 2-point constructed-response questions, 4-point constructed-response questions, and 6-point constructed-response questions)

Each 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 question and that the language in the question does not adversely affect a student's ability to perform the required task. Similarly, sensitivity/bias criteria are used to evaluate whether questions are unbiased, non-offensive, and not disadvantageous to any given subgroup(s).

The third category of item review, alignment, addresses how each item measures a given standard. This category asks the reviewer to comment on key aspects of how the item addresses and calls for the skills demanded by the standards.

The fourth category addresses the specific demands for different item types and formats. Reviewers evaluate each item to ensure that it conforms to the given requirements. For example, multiple-choice items must have, among other characteristics, one unambiguously correct answer and several plausible, but incorrect, answer choices. Following these reviews, only items that are approved by an assigned educator panel move forward for field testing.

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, the New York State Education Department is committed to ongoing standards review over time and periodically solicits thoughtful, specific responses from stakeholders about individual standards within the NYS P–12 Standards.

5.2 EVIDENCE BASED ON RESPONSE PROCESSES

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

The controls and monitoring in place for the Regents Examination in Geometry include the item development process, with attention paid to mitigating the introduction of construct-irrelevant variance. The development process described in the previous sections details the process and attention given to reducing the potential for construct irrelevance in response processes by attending to the quality and alignment of test content to the test blueprint and to the item development guidelines. Further evidence is documented in the test administration and scoring procedures, as well as the results of statistical analyses, which are covered in the following two sections.

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, which are contained in the *School Administrator's Manual, Secondary Level Examinations* (http://www.p12.nysed.gov/assessment/manuals/), have been developed and implemented for the New York State Regents testing program. All secondary-level Regents Examinations are administered under these standard conditions, in order to support valid inferences for all students. These standard 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 http://www.p12.nysed.gov/assessment/hsgen/.

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

1) Quality training (Hoyt & Kerns, 1999; Lumley & McNamara, 1995; Wang, Wong, and Kwong, 2010; Gorman & Rentsch, 2009; Schleicher, Day, Bronston, Mayes, and Riggo, 2002; Woehr & Huffcutt, 1994; Johnson, Penny, and Gordon, 2008; Weigle, 1998)
2) Detection and correction of rating bias (McQueen & Congdon, 1997; Congdon & McQueen, 2000; Myford & Wolfe, 2009; Barkaoui, 2011; Patz, Junker, Johnson, and Mariano, 2002)
3) Consistency or reliability of ratings (Congdon & McQueen, 2000; Harik, Clauser, Grabovsky, Nungester, Swanson, & Nandakumar, 2009; McQueen & Congdon, 1997; Myford & Wolfe, 2009; Mero & Motowidlo, 1995; Weinrott & Jones, 1984)
4) Rubric designs that facilitate consistency of ratings (Pecheone & Chung, 2006; Wolfe & Gitomer, 2000; Cronbach, Linn, Brennan, & Haertel, 1995; Cook & Beckman, 2009; Penny, Johnson, & Gordon, 2000; Smith, 1993; Leacock, Gonzalez, and Conarroe, 2014)

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

To review and select these anchor papers, NYS educators serve as table leaders during the range-finding session. In the range-finding process, committees of educators receive a set of student papers for each field-tested question. Committee members familiarize themselves with each item type and score a number of responses that are representative of each of the different score points. After the independent scoring is completed, the committee reviews and discusses their results and determines consensus scores for the student responses. During this process, atypical responses are important to identify and annotate for use in training and live scoring. The range-finding results are then used to build training materials for the vendor's scorers, who then score the rest of the field test responses to constructed-response items. The final model response sets for the August 2016, January 2017, and June 2017 administration of the Regents Examination in Geometry are located at http://www.nysedregents.org/geometryre/.

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

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

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 constructed-response 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 their assigned staff against model responses and to be available for consultation throughout the scoring process.

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

- The rubric should clearly specify the criteria for awarding each credit.

- The rubric should be aligned to what is asked for in the item and correspond to the knowledge or skill being assessed.
- Whenever possible, the rubric should be written to allow for alternate approaches and other legitimate methods.

In support of the goal of valid score interpretations for each examinee, then, such scoring training procedures are implemented for the Regents Examination in Geometry. Operational raters are selected based on expertise in the exam subject and are assigned a specific set of items to score. No more than approximately one-half of the items on the test are assigned to any one rater. This has the effect of increasing the consistency of scoring across examinee responses by allowing each rater to focus on a subset of items. It also assures 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 examinees. Additionally, no rater is allowed to score the responses of his or her own students.

Statistical Analysis

One statistic that is useful for evaluating the response processes for multiple-choice 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 examinees are not responding the way that the item developer intended them to. As documented in Table 2, the point-biserial statistics for distractors in the multiple-choice items all appear to be negative or very close to 0, indicating that examinees 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. This requires that test developers evaluate the test structure, in order to ensure that the test 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 subgroups of examinees 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 Geometry:

- item difficulty
- item discrimination
- differential item functioning
- IRT model fit
- test reliability
- classification consistency
- test dimensionality

Item Difficulty

Multiple analyses allow 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 2017 Regents Examination in Geometry show a range of *p*-values consistent with the targeted exam difficulty. Item *p*-value ranged from 0.26 to 0.85, with a mean of 0.54. The difficulty distribution illustrated in Figure 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.

Item Discrimination

How well the items on a test discriminate between high- and low-performing examinees is an important measure of the structure of a test. Items that do not discriminate well generally provide less reliable information about student performance. Table 2 and Table 3 provide point-biserial values on the correct responses, and Table 2 also provides point-biserial values on the three distractors for multiple-choice items. The values for all items indicate that they are discriminating well between high- and low-performing examinees. Point-biserials for all distractors are negative or very close to zero, indicating that examinees are responding to the items as expected during item and rubric development. Refer to Chapter 2 of this report for additional details.

Differential Item Functioning

Differential item functioning (DIF) for gender was conducted following field testing of the items in 2012–2017. Sample sizes for subgroups based on ethnicity and English language learner/multilingual learner status were, unfortunately, too small to reliably compute DIF statistics, so only gender DIF analyses were conducted. The Mantel-Haenszel χ^2 and standardized mean difference were used to detect items that may function differently for any of these subgroups. The Mantel-Haenszel χ^2 is a conditional mean comparison of the ordered response categories for reference and focal groups combined over values of the matching variable score. “Ordered” means that a response earning a score of “1” on an item is better than a response earning a score of “0,” 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 our analysis.

One operational item for the June 2018 administration had DIF flags from the field test. One item (# 24) had moderate DIF favoring male students. The item was subsequently reviewed by content specialists. They were unable to identify content-based reasons why the item might be functioning differently between male students and female students, and they did not see any issue with using it for the operational exam.

Full differential item functioning results are reported in Appendix C of the 2012 technical report, in Appendix E of the 2012, 2013, 2014, and 2015 technical reports, and in Appendix F of the 2016 and 2017 technical reports.

IRT Model Fit

Model fit for the Rasch method used to estimate location (difficulty) parameters for the items on the Regents Examination in Geometry provide important evidence that the internal structure of the test is of high technical quality. The INFIT values for all items fall within a targeted range of [0.7, 1.3]. The mean INFIT value is 1.00. A finding of 35 out of 35 items falling in the ideal fit range indicates that the Rasch model fits the Regents Examination in Geometry item data well.

Test Reliability

As discussed, 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 examinee score estimates maximize consistency and therefore minimize error or, theoretically speaking, that examinees who take a test multiple times would get the same score each time. Assessments that include items with higher maximum possible score points may show slightly lower reliabilities than assessments with dichotomous and low maximum possible scores points. The Regents Examination in Geometry contains two constructed-response items with maximum possible points of 4 and 6. The reliability estimate for the Regents Examination in Geometry is 0.91. Refer to Chapter 4 of this report for additional details related to evaluating the standard errors of measurement, and the consistency and accuracy of examinee scores.

Classification Consistency and Accuracy

A decision consistency analysis measures the agreement between the classifications based on two non-overlapping, 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. Decision accuracy is an index to determine 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.

The results for the dichotomies created by the four cut scores, are presented in Table 7. The tabled values are derived with the program *BB-Class* (Brennan, 2004) using the Livingston and Lewis method. The overall decision consistency ranged from 0.88 to 0.92, and the decision accuracy ranged from 0.92 to 0.94. Both decision consistency and accuracy values indicate good consistency and accuracy of examinee classifications.

Dimensionality

In addition to model fit, a strong assumption of the Rasch model is that the construct measured by a test is unidimensional. Violation of this assumption might suggest that the test is measuring something other than the intended content and indicate that the quality of the test structure is compromised. A principal components analysis was conducted to test the assumption of unidimensionality, and the results provide strong evidence that a single dimension in the Regents Examination in Geometry 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 Geometry, strong evidence exists that the exam is functioning as intended and is providing reasonably valid and reliable information about examinee 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 Geometry, these claims could be supported by providing evidence that performance on the Geometry 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 examinee score data for other tests measuring the same content as the Regents Examination in Geometry is ideal, but 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 New York State educators, deeply familiar with both the curriculum standards and their enactment in the classroom, develop all content for the Regents Examination in Geometry.

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 to 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 examinee scores on other tests is not directly supported by either the stated purposes or by the development process and research conducted on examinee data. A brief survey of websites of New York State universities and colleges finds that, beyond the explicitly defined use as a testing requirement toward graduation for students who have completed a course in Geometry, 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 Geometry 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 Geometry and to assess their appropriateness for the inferences being made about course placement.

As stated, the nature of validity arguments is not absolute, but 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 Geometry scores for the purposes described.

References

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

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

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163–178.

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

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

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

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

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

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

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

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

Hanson, B. A. (1994). Extension of Lord-Wingersky algorithm to computing test scores for polytomous items. Retrieved February 17, 2016 from http://www.b-a-h.com/papers/note9401.pdf.

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009, Spring). 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*, 403–424.

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

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

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

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

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, Claudia, Gonzalez, Erin, Conarroe, Mike. (2014). *Developing effective scoring rubrics for AI short answer scoring*. CTB McGraw-Hill Education Innovative Research and Development Grant.

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*. Hillsdale, NJ: Lawrence Erlbaum.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*, 54–72.

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of 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 and standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5–8.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale use. *Journal of Educational Measurement*, *46*(4), 371–389.

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

Patz, R. J., Junker, B. W., Johnson, M. S., and 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: 341.

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 Wei, R. R. (2007). Performance assessment for California teachers: Summary of validity and reliability studies for the 2003−04 pilot year. Palo Alto, CA: 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.

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.

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*. New York, NY: 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, 263–287.

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

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

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

Appendix A: Operational Test Maps

Table A.1 Test Map for August 2017 Administration

Position	Item Type	Max Points	Weight	Cluster	Mean	Point-Biserial	Rasch Difficulty	INFIT
1	MC	1	2	G-GMD.B	0.59	0.33	-1.2407	1.13
2	MC	1	2	G-CO.B	0.73	0.41	-1.9664	0.95
3	MC	1	2	G-GPE.B	0.59	0.36	-1.2080	1.11
4	MC	1	2	G-C.A	0.69	0.46	-1.3785	0.87
5	MC	1	2	G-SRT.B	0.58	0.41	-1.1840	0.99
6	MC	1	2	G-CO.A	0.53	0.41	-1.0068	0.99
7	MC	1	2	G-SRT.B	0.52	0.46	-0.9871	0.94
8	MC	1	2	G-CO.C	0.55	0.49	-1.0028	0.95
9	MC	1	2	G-SRT.B	0.47	0.51	-0.7267	0.92
10	MC	1	2	G-SRT.A	0.40	0.27	-0.3756	1.11
11	MC	1	2	G-CO.C	0.48	0.46	-0.6915	0.97
12	MC	1	2	G-C.A	0.39	0.40	-0.2459	1.04
13	MC	1	2	G-GMD.B	0.37	0.27	-0.2098	1.14
14	MC	1	2	G-CO.C	0.35	0.41	-0.1404	0.99
15	MC	1	2	G-SRT.C	0.35	0.34	0.1528	1.11
16	MC	1	2	G-CO.C	0.45	0.43	-0.6317	0.99
17	MC	1	2	G-GPE.B	0.30	0.47	0.1568	0.92
18	MC	1	2	G-SRT.B	0.29	0.36	0.2702	1.00
19	MC	1	2	G-SRT.C	0.33	0.32	0.3181	1.07
20	MC	1	2	G-MG.A	0.22	0.35	0.7053	1.00
21	MC	1	2	G-SRT.C	0.22	0.39	0.7094	1.03
22	MC	1	2	G-CO.A	0.20	0.42	0.8115	0.99
23	MC	1	2	G-C.B	0.19	0.20	0.8671	1.09
24	MC	1	2	G-GPE.B	0.21	0.32	0.9692	1.03
25	CR	2	1	G-GMD.A	1.06	0.56	-0.9383	1.04
26	CR	2	1	G-CO.C	0.74	0.63	-0.4071	0.87
27	CR	2	1	G-CO.A	0.71	0.62	-0.1597	0.89
28	CR	2	1	G-CO.D	0.47	0.52	0.2147	1.06
29	CR	2	1	G-SRT.A	0.56	0.63	0.3391	0.85
30	CR	2	1	G-CO.B	0.41	0.56	0.5002	0.91
31	CR	2	1	G-GPE.A	0.19	0.49	1.2338	1.02
32	CR	4	1	G-GPE.B	0.56	0.72	0.4220	0.81
33	CR	4	1	G-SRT.B	0.27	0.62	1.0462	0.87
34	CR	4	1	G-SRT.C	0.22	0.58	1.5060	0.94
35	CR	6	1	G-CO.C	1.03	0.83	0.7347	0.74
36	CR	6	1	G-MG.A	0.31	0.65	1.0940	0.72

Table A.2 Test Map for January 2018 Administration

Position	Item Type	Max Points	Weight	Cluster	Mean	Point-Biserial	Rasch Difficulty	INFIT
1	MC	1	2	G-CO.B	0.83	0.40	-2.7368	0.93
2	MC	1	2	G-CO.C	0.81	0.34	-2.5329	1.02
3	MC	1	2	G-CO.B	0.61	0.37	-1.2751	1.04
4	MC	1	2	G-SRT.C	0.56	0.56	-1.1155	0.87
5	MC	1	2	G-GMD.B	0.50	0.33	-0.7296	1.07
6	MC	1	2	G-GPE.B	0.51	0.45	-0.7423	0.94
7	MC	1	2	G-MG.A	0.48	0.39	-0.6260	1.12
8	MC	1	2	G-CO.B	0.54	0.47	-0.6992	0.96
9	MC	1	2	G-CO.C	0.48	0.46	-0.6502	0.98
10	MC	1	2	G-GMD.B	0.50	0.39	-0.6096	1.09
11	MC	1	2	G-SRT.A	0.46	0.40	-0.5526	1.06
12	MC	1	2	G-GPE.A	0.43	0.41	-0.5500	1.02
13	MC	1	2	G-SRT.B	0.49	0.48	-0.4391	0.98
14	MC	1	2	G-SRT.A	0.39	0.43	-0.3169	1.06
15	MC	1	2	G-CO.A	0.40	0.42	-0.2649	1.04
16	MC	1	2	G-C.A	0.36	0.31	-0.0688	1.13
17	MC	1	2	G-SRT.B	0.32	0.38	-0.0075	1.00
18	MC	1	2	G-CO.C	0.33	0.27	0.0080	1.10
19	MC	1	2	G-CO.C	0.39	0.39	0.1148	1.04
20	MC	1	2	G-GPE.B	0.32	0.44	0.1329	0.99
21	MC	1	2	G-C.A	0.29	0.40	0.1767	1.02
22	MC	1	2	G-GMD.A	0.31	0.55	0.2347	0.88
23	MC	1	2	G-SRT.B	0.37	0.23	0.2742	1.16
24	MC	1	2	G-SRT.A	0.22	0.45	0.9276	0.98
25	CR	2	1	G-CO.C	0.75	0.67	-0.2200	0.79
26	CR	2	1	G-CO.D	0.42	0.60	0.4017	0.88
27	CR	2	1	G-SRT.C	0.46	0.59	0.4259	0.96
28	CR	2	1	G-C.B	0.24	0.60	0.8940	0.82
29	CR	2	1	G-MG.A	0.22	0.50	0.9922	0.93
30	CR	2	1	G-CO.B	0.27	0.42	1.1502	1.07
31	CR	2	1	G-SRT.C	0.21	0.54	1.9542	0.88
32	CR	4	1	G-SRT.A	1.27	0.68	0.2259	0.97
33	CR	4	1	G-MG.A	1.11	0.63	0.4326	1.05
34	CR	4	1	G-SRT.C	0.45	0.65	0.8637	0.98
35	CR	6	1	G-GPE.B	0.59	0.72	0.8687	0.87

Table A.3 Test Map for June 2018 Administration

Position	Item Type	Max Points	Weight	Cluster	Mean	Point-Biserial	Rasch Difficulty	INFIT
1	MC	1	2	G-CO.B	0.76	0.40	-2.1614	0.99
2	MC	1	2	G-CO.C	0.57	0.48	-1.1038	0.92
3	MC	1	2	G-CO.A	0.77	0.39	-2.1195	1.01
4	MC	1	2	G-SRT.B	0.65	0.42	-1.3993	0.98
5	MC	1	2	G-SRT.A	0.57	0.49	-0.9712	0.88
6	MC	1	2	G-SRT.C	0.54	0.46	-0.8453	1.05
7	MC	1	2	G-MG.A	0.47	0.34	-0.5960	1.13
8	MC	1	2	G-SRT.C	0.51	0.48	-0.7233	1.00
9	MC	1	2	G-SRT.B	0.47	0.40	-0.4995	0.99
10	MC	1	2	G-GMD.A	0.49	0.43	-0.5476	0.99
11	MC	1	2	G-SRT.B	0.45	0.41	-0.4597	1.02
12	MC	1	2	G-GPE.B	0.47	0.41	-0.3179	1.06
13	MC	1	2	G-CO.C	0.41	0.48	-0.2493	0.95
14	MC	1	2	G-GPE.B	0.37	0.40	-0.0392	1.07
15	MC	1	2	G-GPE.B	0.41	0.38	-0.2004	1.09
16	MC	1	2	G-GMD.B	0.36	0.38	0.0284	1.08
17	MC	1	2	G-C.A	0.36	0.39	0.2270	1.08
18	MC	1	2	G-CO.C	0.28	0.40	0.2525	1.00
19	MC	1	2	G-CO.A	0.26	0.36	0.3963	1.06
20	MC	1	2	G-GPE.A	0.28	0.40	0.4297	1.00
21	MC	1	2	G-SRT.B	0.26	0.30	0.7598	1.07
22	MC	1	2	G-C.B	0.22	0.42	0.8071	0.96
23	MC	1	2	G-SRT.B	0.20	0.25	0.8751	1.05
24	MC	1	2	G-SRT.A	0.14	0.25	1.2187	1.04
25	CR	2	1	G-CO.B	1.23	0.61	-1.1658	1.02
26	CR	2	1	G-SRT.A	0.67	0.56	0.0272	1.03
27	CR	2	1	G-CO.A	0.88	0.53	-0.4659	1.03
28	CR	2	1	G-C.A	0.58	0.56	0.1840	1.04
29	CR	2	1	G-CO.D	0.56	0.61	0.3811	1.09
30	CR	2	1	G-SRT.B	0.52	0.63	0.5473	0.88
31	CR	2	1	G-MG.A	0.32	0.50	0.9776	1.02
32	CR	4	1	G-GPE.B	1.04	0.78	0.1725	0.79
33	CR	4	1	G-SRT.C	0.58	0.74	0.5729	0.88
34	CR	4	1	G-MG.A	0.64	0.68	1.2504	0.89
35	CR	6	1	G-CO.C	0.64	0.69	0.9848	0.92

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

Table B.1 Score Table for August 2017 Administration

Raw Score	Ability	Scale Score
0	-5.7982	0.000
1	-4.5746	4.539
2	-3.8537	8.241
3	-3.4206	11.600
4	-3.1053	14.692
5	-2.8548	17.572
6	-2.6450	20.276
7	-2.4636	22.829
8	-2.3028	25.257
9	-2.1577	27.578
10	-2.0253	29.809
11	-1.9028	31.956
12	-1.7887	34.024
13	-1.6816	36.029
14	-1.5804	37.964
15	-1.4844	39.829
16	-1.3929	41.635
17	-1.3053	43.385
18	-1.2211	45.086
19	-1.1402	46.735
20	-1.0619	48.326
21	-0.9862	49.857
22	-0.9129	51.344
23	-0.8416	52.786
24	-0.7722	54.167
25	-0.7047	55.498
26	-0.6388	56.791
27	-0.5745	58.036
28	-0.5117	59.234
29	-0.4503	60.385
30	-0.3902	61.486
31	-0.3314	62.551
32	-0.2739	63.574
33	-0.2175	64.548
34	-0.1623	65.476
35	-0.1083	66.360
36	-0.0554	67.214
37	-0.0035	68.027
38	0.0473	68.809
39	0.0971	69.551
40	0.1460	70.261
41	0.1940	70.943
42	0.2411	71.590
43	0.2874	72.212
44	0.3330	72.804
45	0.3778	73.379
46	0.4220	73.940
47	0.4656	74.478
48	0.5087	74.996
49	0.5512	75.495
50	0.5935	75.991
51	0.6354	76.475
52	0.6770	76.948
53	0.7184	77.415
54	0.7598	77.867
55	0.8011	78.324
56	0.8425	78.775
57	0.8841	79.227
58	0.9259	79.683
59	0.9681	80.133
60	1.0109	80.587
61	1.0542	81.048
62	1.0984	81.516
63	1.1435	81.993
64	1.1897	82.478
65	1.2373	82.969
66	1.2864	83.471
67	1.3373	83.987
68	1.3903	84.521
69	1.4459	85.076
70	1.5042	85.642
71	1.5660	86.232
72	1.6316	86.845
73	1.7020	87.480
74	1.7780	88.146
75	1.8606	88.839
76	1.9515	89.577
77	2.0523	90.354
78	2.1658	91.173
79	2.2953	92.048
80	2.4460	92.974
81	2.6254	93.950
82	2.8465	94.993
83	3.1332	96.101
84	3.5385	97.280
85	4.2324	98.542
86	5.4373	100.000

Table B.2 Score Table for January 2018 Administration

Raw Score	Ability	Scale Score
0	-6.0612	0.000
1	-4.8204	3.578
2	-4.0763	7.188
3	-3.6219	10.589
4	-3.2873	13.926
5	-3.0192	17.173
6	-2.7938	20.316
7	-2.5983	23.347
8	-2.4250	26.260
9	-2.2689	29.055
10	-2.1265	31.732
11	-1.9953	34.290
12	-1.8734	36.737
13	-1.7594	39.073
14	-1.6522	41.301
15	-1.5508	43.428
16	-1.4545	45.456
17	-1.3626	47.389
18	-1.2748	49.233
19	-1.1904	50.990
20	-1.1092	52.666
21	-1.0308	54.263
22	-0.9550	55.785
23	-0.8815	57.235
24	-0.8101	58.616
25	-0.7406	59.932
26	-0.6729	61.183
27	-0.6068	62.377
28	-0.5422	63.512
29	-0.4790	64.592
30	-0.4170	65.621
31	-0.3561	66.600
32	-0.2965	67.531
33	-0.2378	68.417
34	-0.1801	69.260
35	-0.1233	70.064
36	-0.0674	70.828
37	-0.0124	71.558
38	0.0418	72.252
39	0.0953	72.916
40	0.1480	73.551
41	0.2000	74.160
42	0.2513	74.746
43	0.3019	75.310
44	0.3520	75.856
45	0.4015	76.385
46	0.4504	76.900
47	0.4989	77.403
48	0.5470	77.898
49	0.5948	78.385
50	0.6424	78.868
51	0.6900	79.349
52	0.7375	79.829
53	0.7852	80.311
54	0.8332	80.798
55	0.8816	81.292
56	0.9307	81.793
57	0.9806	82.307
58	1.0316	82.832
59	1.0839	83.371
60	1.1377	83.926
61	1.1934	84.499
62	1.2514	85.090
63	1.3118	85.701
64	1.3753	86.333
65	1.4424	86.986
66	1.5136	87.658
67	1.5896	88.350
68	1.6713	89.063
69	1.7599	89.797
70	1.8565	90.551
71	1.9631	91.327
72	2.0820	92.118
73	2.2165	92.932
74	2.3716	93.773
75	2.5548	94.647
76	2.7790	95.563
77	3.0681	96.537
78	3.4753	97.555
79	4.1713	98.641
80	5.3782	100.00

Table B.3 Score Table for June 2018 Administration

Raw Score	Ability	Scale Score
0	-5.8839	0.000
1	-4.6561	4.209
2	-3.9295	8.089
3	-3.4912	11.759
4	-3.1713	15.236
5	-2.9164	18.536
6	-2.7028	21.673
7	-2.5179	24.658
8	-2.3539	27.501
9	-2.2060	30.213
10	-2.0709	32.797
11	-1.9460	35.265
12	-1.8297	37.624
13	-1.7205	39.877
14	-1.6174	42.029
15	-1.5196	44.085
16	-1.4263	46.050
17	-1.3371	47.927
18	-1.2515	49.719
19	-1.1691	51.432
20	-1.0897	53.065
21	-1.0130	54.624
22	-0.9387	56.109
23	-0.8667	57.524
24	-0.7968	58.871
25	-0.7289	60.150
26	-0.6629	61.368
27	-0.5986	62.524
28	-0.5360	63.620
29	-0.4748	64.662
30	-0.4151	65.651
31	-0.3568	66.590
32	-0.2996	67.482
33	-0.2436	68.330
34	-0.1886	69.137
35	-0.1345	69.907
36	-0.0812	70.640
37	-0.0287	71.343
38	0.0234	72.016
39	0.0748	72.662
40	0.1258	73.285
41	0.1765	73.887
42	0.2269	74.470
43	0.2771	75.035
44	0.3272	75.587
45	0.3772	76.128
46	0.4272	76.657
47	0.4773	77.180
48	0.5275	77.698
49	0.5779	78.213
50	0.6286	78.728
51	0.6797	79.245
52	0.7311	79.764
53	0.7830	80.290
54	0.8355	80.823
55	0.8887	81.364
56	0.9427	81.917
57	0.9976	82.482
58	1.0537	83.059
59	1.1110	83.651
60	1.1698	84.257
61	1.2306	84.879
62	1.2935	85.517
63	1.3590	86.171
64	1.4277	86.843
65	1.5001	87.533
66	1.5768	88.236
67	1.6587	88.956
68	1.7468	89.691
69	1.8421	90.441
70	1.9462	91.205
71	2.0606	91.982
72	2.1879	92.771
73	2.3312	93.569
74	2.4953	94.381
75	2.6876	95.209
76	2.9207	96.059
77	3.2180	96.937
78	3.6325	97.854
79	4.3339	98.830
80	5.5430	100.00

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 of 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, but 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 of 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 of 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. $\angle f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right)$

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 Common Core 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.).				

Math Art	Yes	No	n/a	Explain or Describe
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.				

Item Alignment	Yes	No	n/a	Explain or Describe
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 Representations(s). *If "No," indicate the correct MR code(s).*				
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.*				

Application/Modeling Items	Yes	No	n/a	Explain or Describe
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).				
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 student 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're trying to elicit from the student.				
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 (won't impact stats)?* *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 = ½ mv2).* 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 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 Common Core 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 Common Core 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.

Appendix D: Tables and Figures for August 2017 Administration

Table D.1 Multiple-Choice Item Analysis Summary: Regents Examination in Geometry

Item	Number	*p*-Value	SD	Point-Biserial	Point-Biserial Distractor 1	Point-Biserial Distractor 2	Point-Biserial Distractor 3
1	19,481	0.59	0.49	0.27	-0.14	-0.11	-0.15
2	19,481	0.72	0.45	0.33	-0.11	-0.14	-0.25
3	19,481	0.64	0.48	0.29	-0.17	-0.15	-0.11
4	19,481	0.56	0.50	0.46	-0.26	-0.25	-0.14
5	19,481	0.58	0.49	0.30	-0.15	-0.12	-0.16
6	19,481	0.54	0.50	0.38	-0.15	-0.17	-0.20
7	19,481	0.55	0.50	0.40	-0.13	-0.21	-0.22
8	19,481	0.54	0.50	0.44	-0.24	-0.18	-0.19
9	19,481	0.35	0.48	0.46	-0.15	-0.28	-0.12
10	19,481	0.43	0.50	0.33	-0.14	-0.15	-0.13
11	19,481	0.47	0.50	0.49	-0.12	-0.17	-0.33
12	19,481	0.30	0.46	0.31	-0.08	-0.10	-0.18
13	19,481	0.42	0.49	0.19	-0.05	0.02	-0.19
14	19,481	0.39	0.49	0.42	-0.15	-0.19	-0.19
15	19,481	0.57	0.49	0.40	-0.26	-0.18	-0.11
16	19,481	0.42	0.49	0.25	-0.08	-0.02	-0.23
17	19,481	0.45	0.50	0.45	-0.27	-0.16	-0.14
18	19,481	0.29	0.45	0.30	-0.12	-0.08	-0.12
19	19,481	0.45	0.50	0.46	-0.16	-0.25	-0.19
20	19,481	0.30	0.46	0.39	-0.17	-0.22	-0.02
21	19,481	0.45	0.50	0.34	-0.10	-0.20	-0.19
22	19,481	0.30	0.46	0.41	-0.10	-0.20	-0.15
23	19,481	0.18	0.38	0.25	-0.04	-0.12	-0.08
24	19,481	0.25	0.43	0.31	-0.12	-0.11	-0.08

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

Item	Min. score	Max. score	Number of Students	Mean	SD	*p*-Value	Point-Biserial
25	0	2	19,481	1.19	0.86	0.59	0.36
26	0	2	19,481	0.49	0.82	0.25	0.63
27	0	2	19,481	0.88	0.78	0.44	0.44
28	0	2	19,481	1.10	0.95	0.55	0.41
29	0	2	19,481	0.60	0.77	0.30	0.56
30	0	2	19,481	0.69	0.74	0.34	0.51
31	0	2	19,481	0.50	0.77	0.25	0.55
32	0	4	19,481	0.64	1.20	0.16	0.68
33	0	4	19,481	0.30	0.78	0.07	0.62
34	0	4	19,481	0.31	0.91	0.08	0.65
35	0	6	19,481	0.66	1.36	0.11	0.71
36	0	6	19,481	0.59	1.27	0.10	0.71

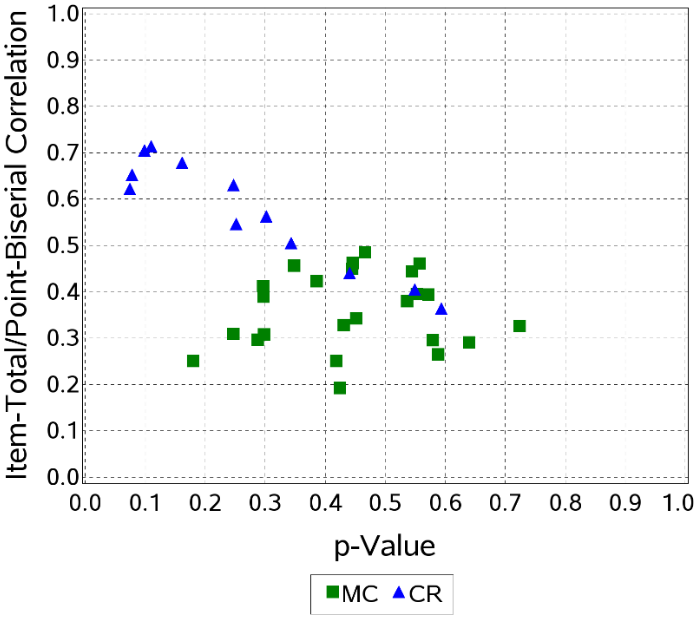


Figure D.1 Scatter Plot: Regents Examination in Geometry

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

Statistics	N	Mean	Min	Q1	Median	Q3	Max
p-value	36	0.39	0.07	0.27	0.42	0.55	0.72
Point-Biserial	36	0.43	0.19	0.32	0.41	0.50	0.71

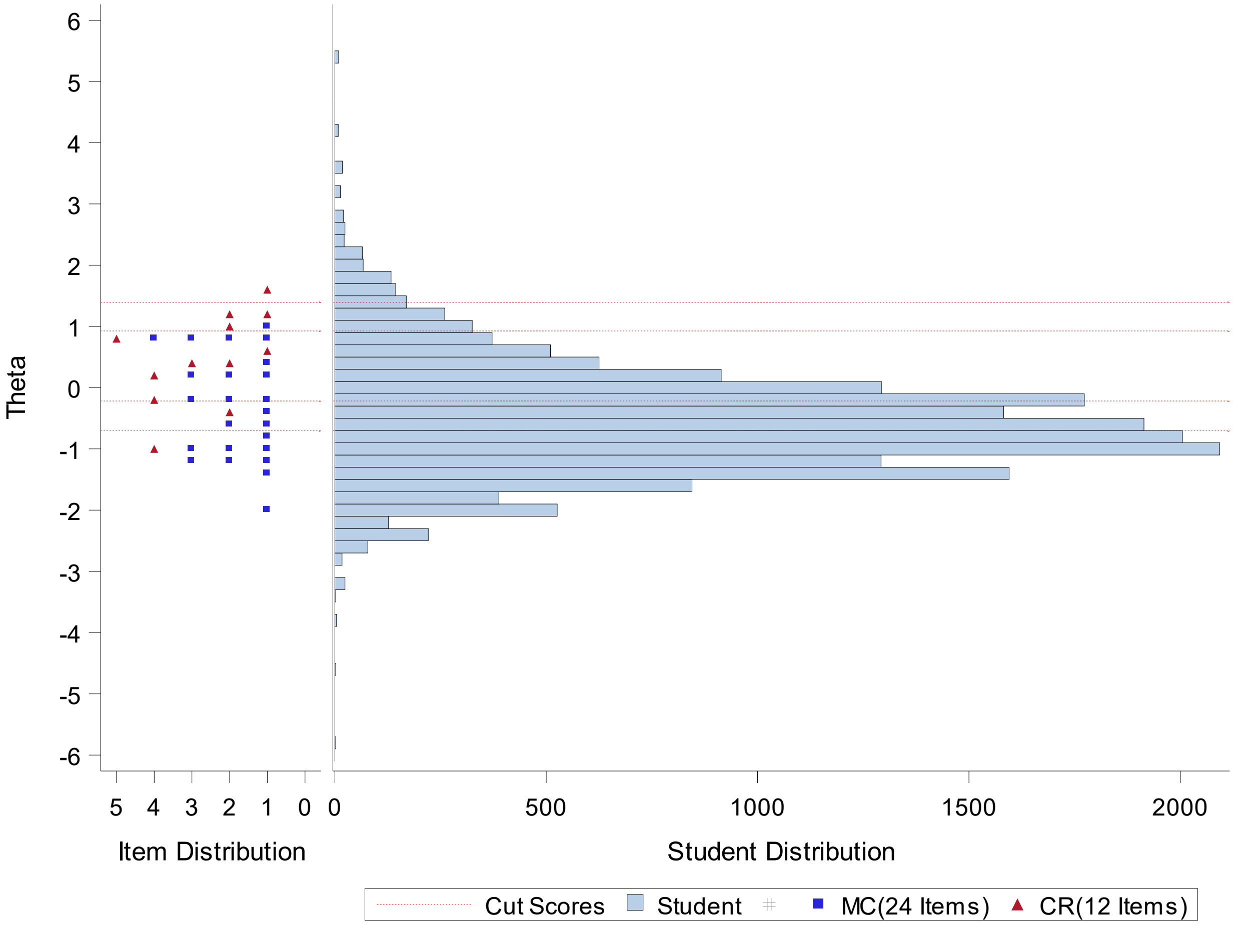


Figure D.2 Student Performance Map: Regents Examination in Geometry

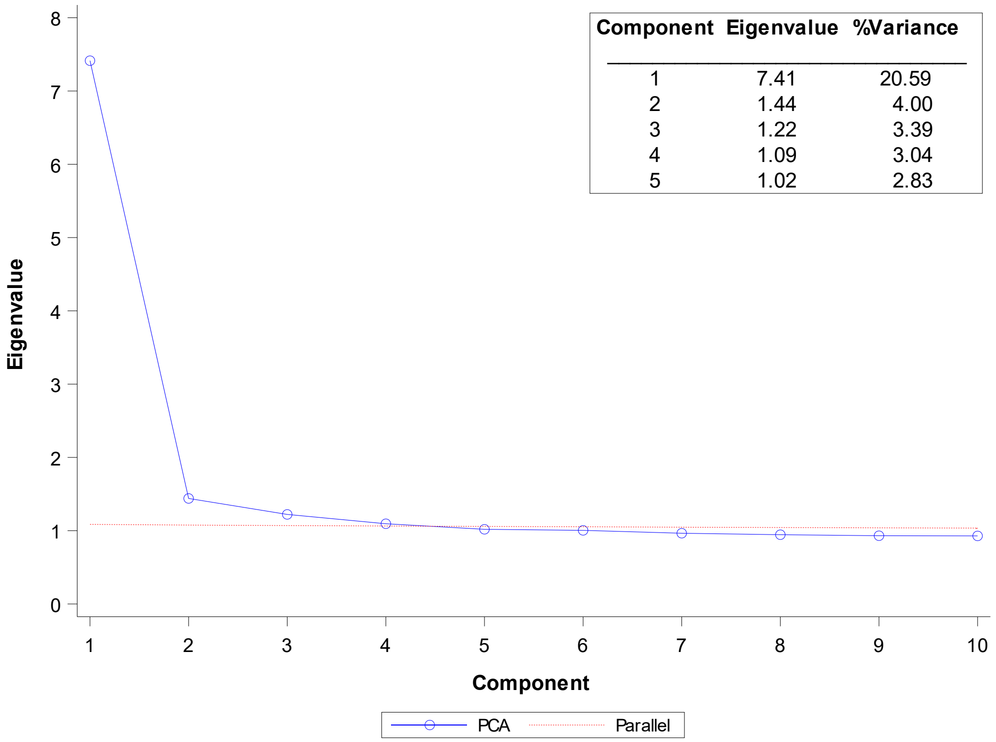


Figure D.3 Scree Plot: Regents Examination in Geometry

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

Statistic Type	Value
N	630
Mean	-0.03
SD	0.03
Minimum	-0.12
P_{10}	-0.06
P_{25}	-0.04
P_{50}	-0.03
P_{75}	-0.01
P_{90}	0.01
Maximum	0.16
>|0.20|	0

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

	INFIT Mean Square					
	N	Mean	SD	Min	Max	[0.7, 1.3]
Geometry	36	1.00	0.08	0.84	1.19	[36/36]

Table D.6 Reliabilities and Standard Errors of Measurement: Regents Examination in Geometry

Subject	Coefficient Alpha	SEM
Geometry	0.88	5.08

Table D.7 Decision Consistency and Accuracy Results: Regents Examination in Geometry

Statistic	1/2	2/3	3/4	4/5
Consistency	0.80	0.89	0.97	0.98
Accuracy	0.86	0.92	0.98	0.99

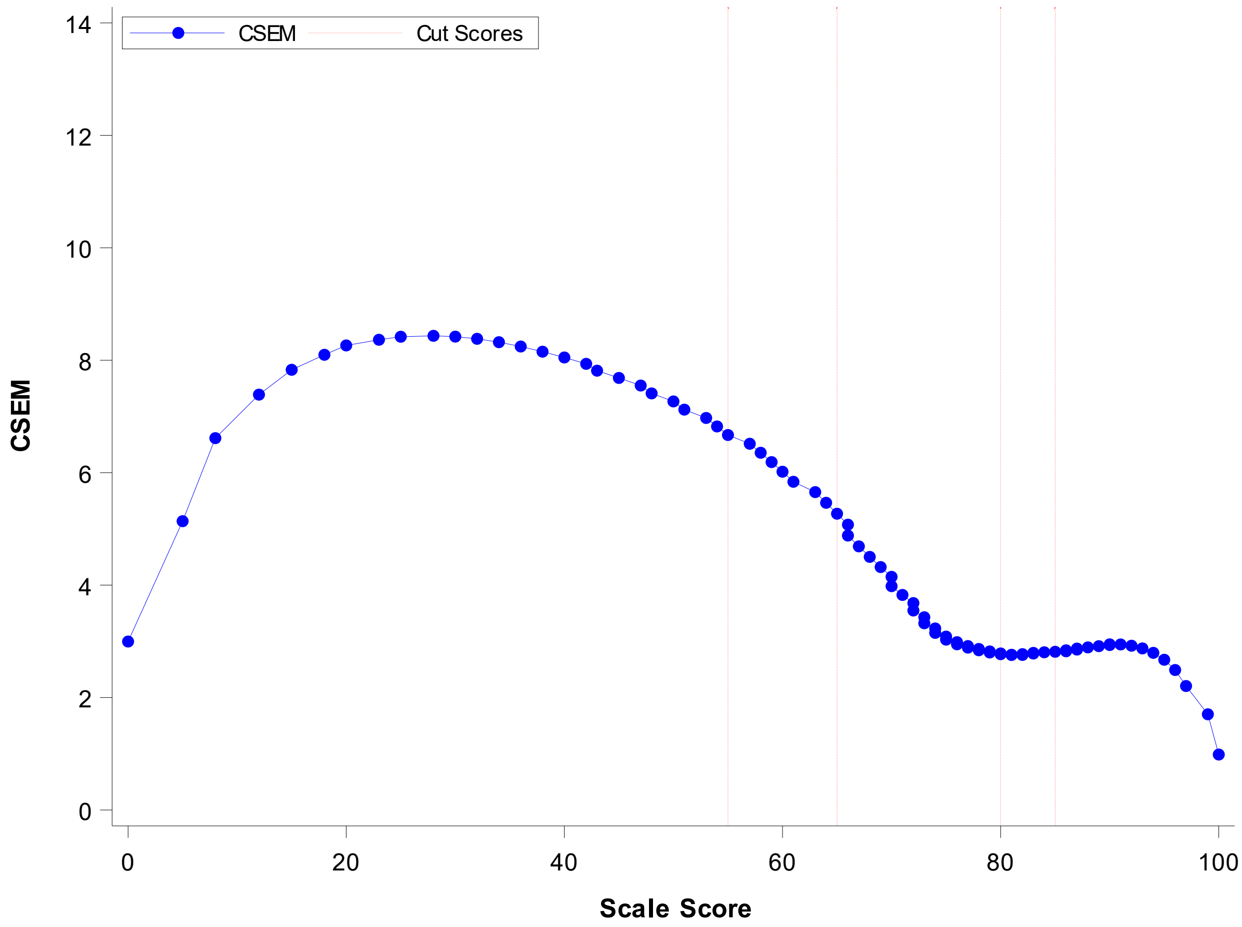


Figure D.4 Conditional Standard Error Plot: Regents Examination in Geometry

Table D.8 Group Means: Regents Examination in Geometry

Demographics	Number	Mean Scale Score	SD Scale Score
All Students*	19,481	56.99	14.65
Ethnicity			
American Indian/Alaska Native	129	57.37	15.18
Asian/Native Hawaiian/Other Pacific Islander	1,777	63.06	16.36
Black/African American	4,190	51.70	13.36
Hispanic/Latino	4,942	52.67	13.30
Multiracial	312	56.99	14.73
White	8,127	61.01	13.94
Gender			
Female	10,367	56.91	14.85
Male	9,110	57.08	14.42
Economically Disadvantaged			
No	9,776	60.16	14.71
Yes	9,705	53.79	13.88
English Language Learner/Multilingual Learner			
No	18,998	57.23	14.54
Yes	483	47.28	15.63
Student with a Disability			
No	17,822	57.76	14.50
Yes	1,659	48.65	13.65

*Note: Four students were not reported in the Ethnicity and Gender group, but they are reflected in “All Students.”

Appendix E: Tables and Figures for January 2018 Administration

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

Item	Number	*p*-Value	SD	Point Biserial	Point-Biserial Distractor 1	Point-Biserial Distractor 2	Point-Biserial Distractor 3
1	14,685	0.87	0.34	0.25	-0.17	-0.14	-0.09
2	14,685	0.81	0.40	0.22	-0.16	-0.14	-0.06
3	14,685	0.55	0.50	0.34	-0.19	-0.15	-0.15
4	14,685	0.51	0.50	0.48	-0.23	-0.27	-0.14
5	14,685	0.51	0.50	0.25	-0.04	-0.06	-0.20
6	14,685	0.53	0.50	0.41	-0.22	-0.15	-0.20
7	14,685	0.55	0.50	0.36	-0.18	-0.12	-0.21
8	14,685	0.52	0.50	0.40	-0.22	-0.14	-0.20
9	14,685	0.46	0.50	0.31	-0.12	-0.16	-0.15
10	14,685	0.30	0.46	0.40	-0.09	-0.22	-0.11
11	14,685	0.40	0.49	0.30	-0.13	-0.15	-0.12
12	14,685	0.51	0.50	0.35	-0.17	-0.18	-0.12
13	14,685	0.44	0.50	0.39	-0.11	-0.20	-0.20
14	14,685	0.36	0.48	0.31	-0.16	-0.09	-0.12
15	14,685	0.33	0.47	0.39	-0.14	-0.11	-0.18
16	14,685	0.38	0.48	0.20	-0.05	-0.04	-0.14
17	14,685	0.21	0.41	0.39	-0.08	-0.18	-0.15
18	14,685	0.33	0.47	0.33	-0.24	0.04	-0.19
19	14,685	0.43	0.49	0.31	-0.19	-0.10	-0.11
20	14,685	0.30	0.46	0.42	-0.13	-0.13	-0.20
21	14,685	0.26	0.44	0.31	-0.12	-0.23	-0.01
22	14,685	0.17	0.38	0.49	-0.15	-0.28	-0.02
23	14,685	0.39	0.49	0.27	-0.10	-0.10	-0.12
24	14,685	0.28	0.45	0.34	-0.20	-0.13	-0.04

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

Item	Min. score	Max. score	Number of Students	Mean	SD	*p*-Value	Point-Biserial
25	0	2	14,685	0.43	0.68	0.22	0.60
26	0	2	14,685	0.52	0.83	0.26	0.49
27	0	2	14,685	0.57	0.78	0.28	0.40
28	0	2	14,685	0.25	0.59	0.13	0.63
29	0	2	14,685	0.26	0.56	0.13	0.59
30	0	2	14,685	0.36	0.62	0.18	0.39
31	0	2	14,685	0.53	0.77	0.26	0.65
32	0	4	14,685	1.22	1.06	0.31	0.57
33	0	4	14,685	1.11	1.25	0.28	0.64
34	0	4	14,685	1.17	1.59	0.29	0.70
35	0	6	14,685	0.94	1.66	0.16	0.74

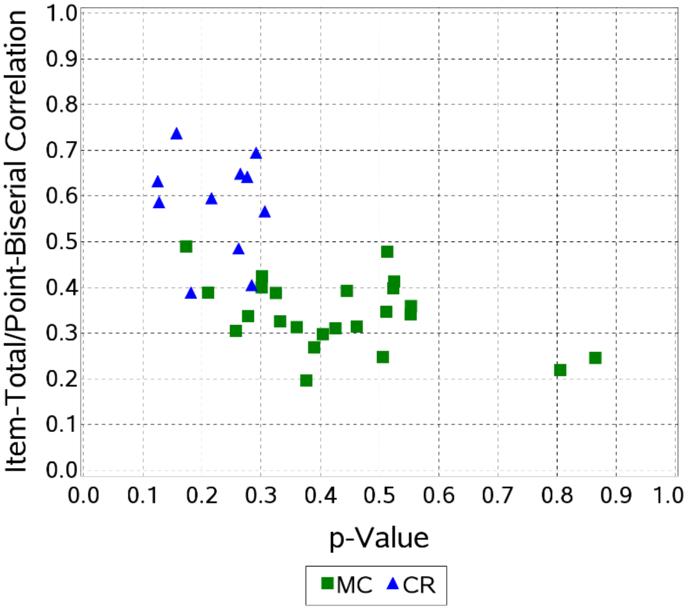


Figure E.1 Scatter Plot: Regents Examination in Geometry

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

Statistics	N	Mean	Min	Q1	Median	Q3	Max
p-value	35	0.37	0.13	0.26	0.33	0.51	0.87
Point-Biserial	35	0.42	0.20	0.31	0.39	0.49	0.74

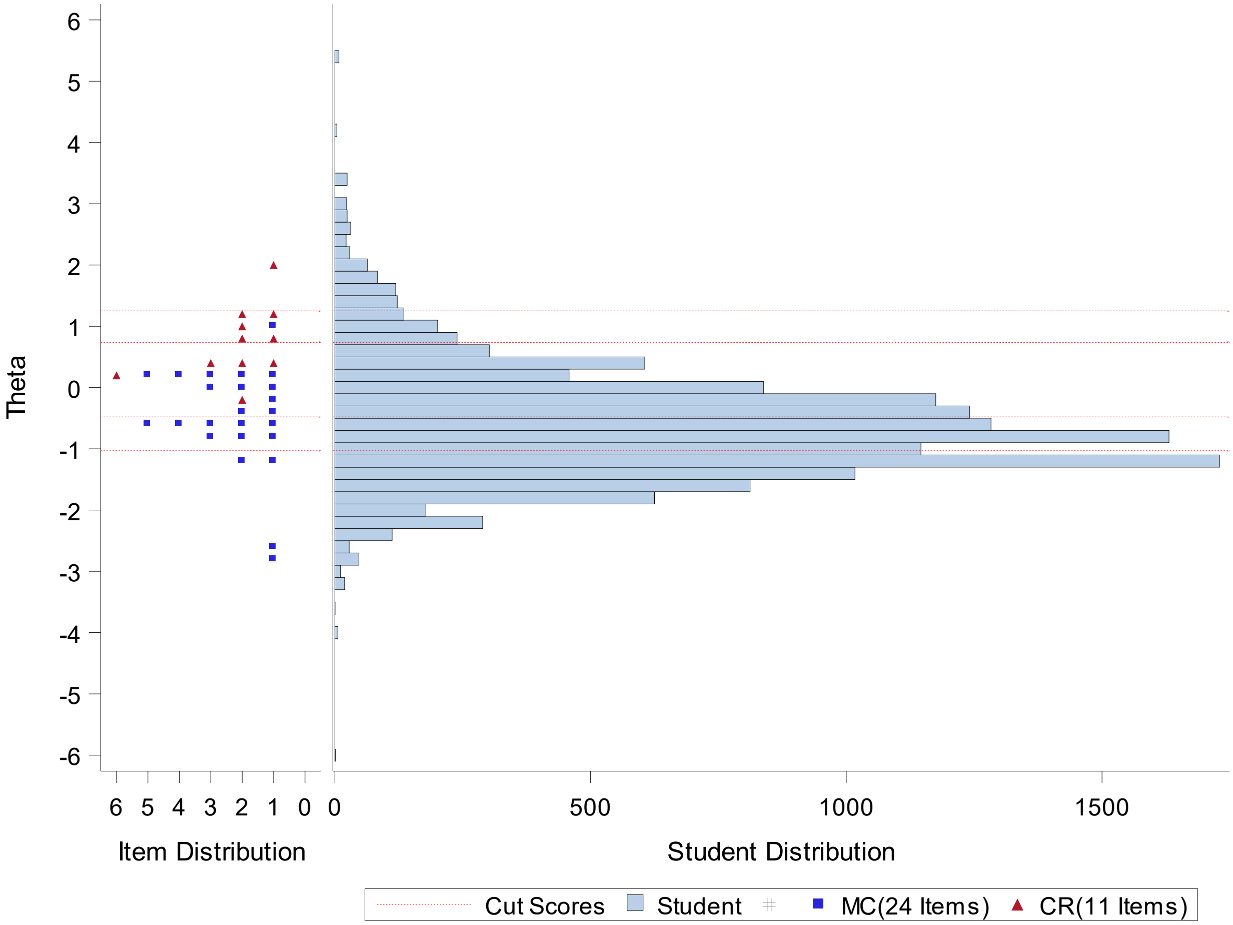


Figure E.2 Student Performance Map: Regents Examination in Geometry

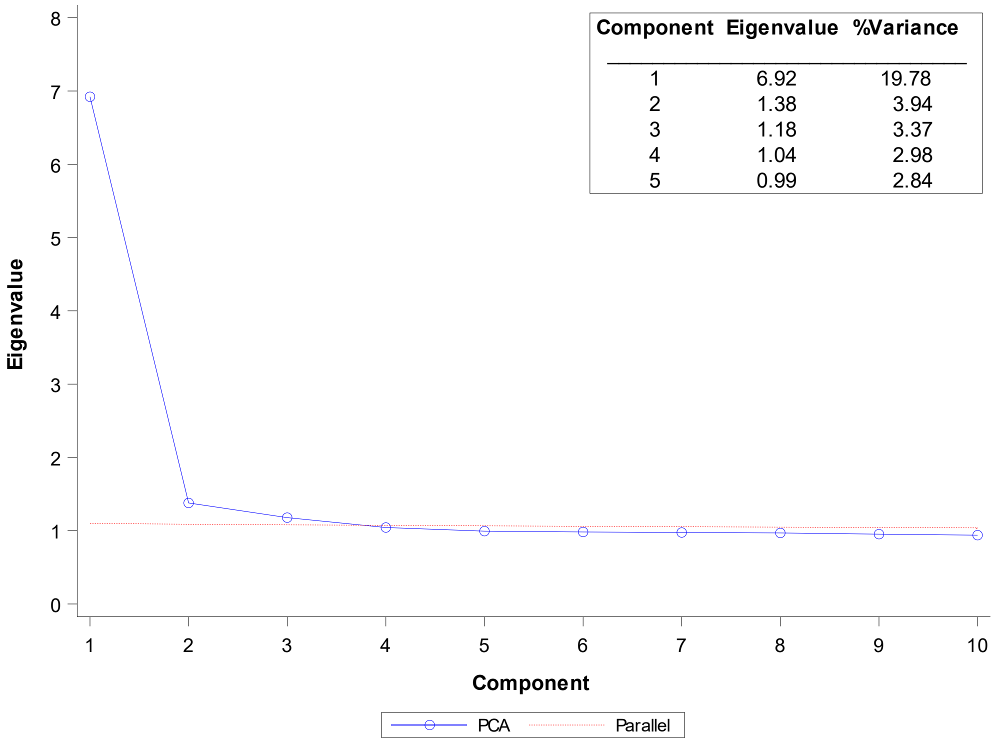


Figure E.3 Scree Plot: Regents Examination in Geometry

Table E.4 Summary of Item Residual Correlations: Regents Examination in Geometry

Statistic Type	Value
N	595
Mean	-0.03
SD	0.04
Minimum	-0.13
P_{10}	-0.07
P_{25}	-0.05
P_{50}	-0.02
P_{75}	0.00
P_{90}	0.01
Maximum	0.31
>|0.20|	1

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

	INFIT Mean Square					
	N	Mean	SD	Min	Max	[0.7, 1.3]
Geometry	35	1.00	0.10	0.77	1.17	[35/35]

Table E.6 Reliabilities and Standard Errors of Measurement: Regents Examination in Geometry

Subject	Coefficient Alpha	SEM
Geometry	0.87	5.11

Table E.7 Decision Consistency and Accuracy Results: Regents Examination in Geometry

Statistic	1/2	2/3	3/4	4/5
Consistency	0.74	0.86	0.96	0.98
Accuracy	0.80	0.90	0.98	0.99

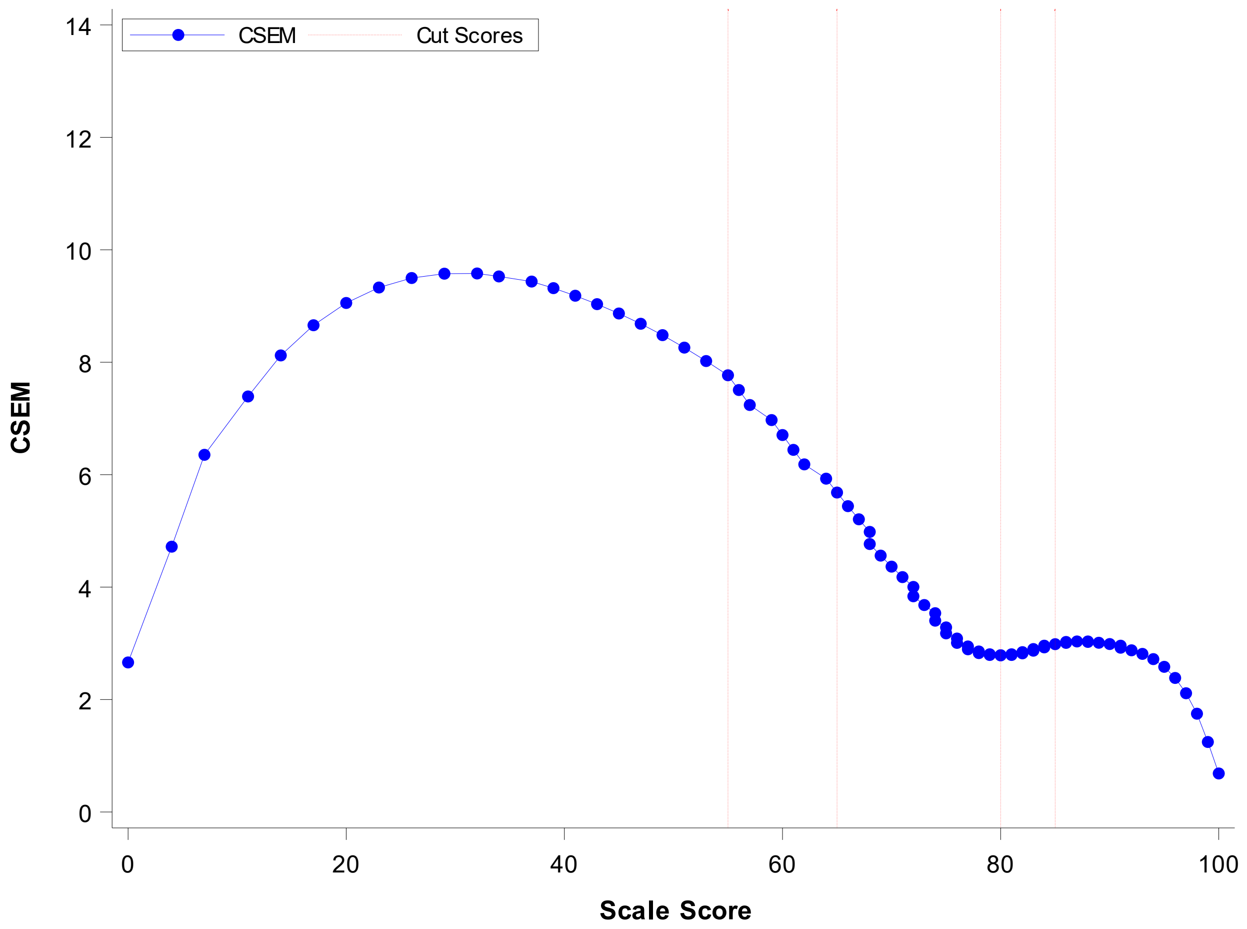


Figure E.4 Conditional Standard Error Plot: Regents Examination in Geometry

Table E.8 Group Means: Regents Examination in Geometry

Demographics	Number	Mean Scale Score	SD Scale Score
All Students	14,685	59.66	14.44
Ethnicity			
American Indian/Alaska Native	121	58.63	13.81
Asian/Native Hawaiian/Other Pacific Islander	1,631	65.60	15.74
Black/African American	3,844	56.14	13.54
Hispanic/Latino	4,417	56.15	13.12
Multiracial	240	61.56	14.74
White	4,431	63.96	14.05
Gender			
Female	7,808	59.72	14.17
Male	6,876	59.59	14.74
Economically Disadvantaged			
No	5,544	62.49	14.30
Yes	9,141	57.95	14.25
English Language Learner/Multilingual Learner			
No	13,816	59.81	14.15
Yes	869	57.25	18.22
Student with Disabilities			
No	13,345	60.47	14.21
Yes	1,340	51.63	14.21

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