New York State Regents Examination in Geometry

2016 Technical Report

Prepared for the New York State Education Department
by Pearson

March 2017

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 on 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 2015 and January 2016 administrations. Chapters 1–5 usually detail results for the June administration; however, the Regents Examination in Geometry was not administered in June 2016. The last administration for the Geometry exam was given in January 2016. Results for the August 2015 and January 2016 administrations are provided in Appendices B and C, 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 in August and January to students enrolled in New York State schools. The examination is based on the Geometry Core Curriculum which is based on New York State Learning Standards for Mathematics.

1.2 PURPOSES OF THE EXAM

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 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

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

Geometry may also be used to satisfy various locally established requirements throughout the state.

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 in Geometry.

Table 1 provides a demographic breakdown of all students who took the August 2015 and January 2016 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. The results include those exams administered August 2015 and January 2016 (see http://data.nysed.gov/). If a student takes the same exam multiple times in the year, the highest score only 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. Tables and figures for the August 2015 and January 2016 administrations are included in Appendices D and E, respectively.

Table 1 Total Examinee Population: Regents Examination in Geometry

	August Admin*		January Admin**	
Demographics	**Number**	**Percent**	**Number**	**Percent**
All Students	11,120	100	10,221	100
Race/Ethnicity				
American Indian or Alaska Native	63	0.57	84	0.82
Asian/Native Hawaiian/Other Pacific Islander	905	8.14	1,141	11.17
Black/African American	3,091	27.82	3,136	30.69
Hispanic/Latino	2,790	25.11	3,593	35.16
Multiracial	132	1.19	110	1.08
White	4,131	37.18	2,155	21.09
English Language Learner				
No	10,778	96.92	9,462	92.57
Yes	342	3.08	759	7.43
Economically Disadvantaged				
No	5,380	48.38	3,408	33.34
Yes	5,740	51.62	6,813	66.66
Gender				
Female	5,864	52.77	5,514	53.96
Male	5,248	47.23	4,705	46.04
Student with Disabilities				

No	9,967	89.63	9,066	88.70
Yes	1,153	10.37	1,155	11.30

*Note: Eight students were not reported in the Ethnicity and Gender group, but they are 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. Additionally, this value might suggest that the item was relatively easy and/or that the students who attempted the item were relatively high achievers. For constructed-response (CR) items, mean scores can range from the minimum possible score (usually zero) to the maximum possible score. To facilitate average score comparability across MC and CR items, mean item performance for CR items is divided by the maximum score possible so that the *p*-values for all items are reported as a ratio from 0.0 to 1.0.

Although the *p*-value statistic does not consider individual student ability in its computation, it provides a useful view of overall item difficulty, and can provide an early and simple indication of items that are too difficult for the population of students taking the examination. Items with very high or very low *p*-values receive added scrutiny during all follow-up analyses, including item response theory analyses that factor student ability into estimates of item difficulty. Such items may be removed from the item pool during the test development process, as field testing typically reveals that they add very little measurement information. Refer to Appendices D and E for *p*-values from the August 2015 and January 2016 administration, respectively.

2.2 ITEM DISCRIMINATION

At the most general level, estimates of item discrimination indicate an item's ability to differentiate between high and low performance on an item. 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 overall) would be more likely to answer the same item incorrectly. Pearson's product-

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

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 that 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 the analysis. The point-biserial values on the distractors are typically negative. Positive values can indicate that higher-performing students are selecting an incorrect answer or that the item key for the correct answer should be checked.

Refer to Appendices D and E for point-biserial values on the correct response and three distractors for the August 2015 and January 2016 administration.

2.3 DISCRIMINATION ON DIFFICULTY SCATTER PLOTS

Scatter plots of item discrimination values (*y*-axis) and item difficulty values (*x*-axis) for the August 2015 and January 2016 administration are presented in Appendices D and E, respectively. The descriptive statistics of *p*-value and point-biserials, including mean, minimum, Q1, median, Q3, and maximum, are also presented in the appendices.

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 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 on 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

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

where θ_n represents 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 RPCM reduces to the standard Rasch model and the single step difficulty is referred to as the item's difficulty. The Rasch model predicts the probability of person n getting item i correct as follows:

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

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

3.2 SOFTWARE AND ESTIMATION ALGORITHM

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

3.3 CHARACTERISTICS OF THE TESTING POPULATION

The data analyses reported here are usually based on all students who took the Regents Examination in Geometry in the 2016 administration. 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 Appendices D and E for the August 2015 and January 2016 administrations, respectively. 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.

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 between students' performances. Principal Components Analysis (PCA) can be used to assess the unidimensionality assumption. The purpose of the analysis is to verify if any other dominant components exist among the items. If any other dimensions are found, the unidimensionality assumption would be violated.

A parallel analysis (Horn, 1965) was conducted to help distinguish components that are real from components that are random. Parallel analysis is a technique 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.

Appendices D and E show the PCA and parallel analysis results for the Regents Examination in Geometry for the August 2015 and January 2016 administrations, respectively. The results include the eigenvalues and the percentage of variance explained for the first five components, as well as the scree plots. The scree plots show the eigenvalues plotted by component number and the results of a parallel analysis. Although the total number of components in PCA is same as the total number of items in a test, the figures show only 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 an evidence of unidimensionality.

As rule of thumb, Reckase (1979) proposed that the variance explained by the primary dimension should be greater than 20 percent, in order 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.

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:

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

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 is 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. Absolute 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. Appendices D and E show the summary statistics — mean, standard deviation, minimum, maximum, and several percentiles (P_{10}, P_{25}, P_{50}, P_{75}, P_{90}) — for all the residual correlations for each test from the August 2015 and January 2016 administrations, respectively. The total number of item pairs (N) and the number of pairs with the absolute residual correlations greater than 0.20 are also reported in this table.

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 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.

Appendices D and E present the summary statistics of INFIT mean square statistics for the Regents Examination in Geometry, including the mean, standard deviation, and minimum and maximum values for the August 2015 and January 2016 administrations, respectively. The number of items within a targeted range of [0.7, 1.3] is also reported in the appendices.

Items for the 2015 Regents Examination in Geometry were field tested in 2008–2015.

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 2015 and January 2016 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 the NYSED and are based on the values in the bank for all essay items. For the Regents Examination in Geometry, there is only one six-point item with fixed constants.

The New York State Regents Examination in Geometry has three cut scores, which are set at the scale scores of 55, 65, 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 0.279. 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 2015 and January 2016 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 2009 administration. Scale scores from the August 2015 and January 2016 administrations are on the same scale and can be directly compared to scale scores on all previous administrations back to the June 2009 administration.

When the base administration was concluded, the initial raw score to scale score relationship was established. Three raw scores were fixed at specific scale scores. Scale scores of 0 and 100 were fixed to correspond to the minimum and maximum possible raw scores. In addition, a standard setting had been held to determine the passing and passing with distinction cut scores in the raw score metric. The scale score points of 65, and 85 were set to correspond to those raw score cuts. A third-degree polynomial is required to fit a line exactly to five arbitrary points (e.g., the raw scores corresponding to the four critical scale scores of 0, 65, 85, and 100). The general form of this best-fitting line is:

$$SS = m3 * RS^3 + m2 * RS^2 + m1 * RS1 + m0,$$

where SS is the scaled score, RS is the raw score, and m0 through m3 are the transformation constants that convert the raw score into the scale score (please note that m0 will always be equal to zero in this application, since a raw score of zero corresponds to a scale score of zero). A subscript for a person on both dependent and independent variables is not present for simplicity. The above relationship and the values of m1 to m3 specific to this subject were then used to determine the scale scores corresponding to the remainder of the raw scores on the examination. This initial relationship between the raw and scale scores became the base scale.

The Rasch difficulty parameters for the items on the base form were then used to derive a raw score to Rasch student ability (theta score) relationship. This 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, 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 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 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. It can be conceptualized as the extent to which an exchangeable set of items from the same domain would result in a similar rank ordering of students. Note that relative error is reflected in this index. Excessive variation in student performance from one sample of items to the next should be of particular concern for any achievement test user.

A general computational formula for Coefficient Alpha is as follows:

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

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 Appendices D and E for the August 2015 and January 2016 administrations, respectively.

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 that 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 student's 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 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 a simple mathematical form. In the absence of a

[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.

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 Exams does, however, make it possible to calculate CSEMs 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 have a one-to-one relationship between raw (θ) and scale scores, the CSEM for each achievable scale score can be calculated 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, their 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 their 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 their 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 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 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 Exams tends to be roughly linear from scale scores of 0 to 20 and then concave down from about 20 to 100. In other words, the scale scores track linearly with the raw scores for the first quarter of the scale score range and then are compressed relative to the raw scores for the remaining three quarters of the range, though there are slight variations. The CSEMs for the Regents Exams can be expected to have inverted-U shaped patterns, with some variations.

Appendices D and E show this type of CSEM variation for the Regents Examination in Geometry for the August 2015 and January 2016 administrations, respectively, in which the compression of raw score to scale scores around the cut score of 85 changes the shape of the curve slightly. This type of expansion and compression can be seen in the appendices by looking at the changing density of raw score points along the scale score range on the horizontal axis. Specifically, the greatest compression is noted around the cut scores indicated in red.

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 tables below.

		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 1 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 2 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 using four performance levels would allow more opportunity to change achievement levels. Hence, there would be more classification errors and less accuracy with four performance levels, resulting in lower consistency indices.

Results and Observations The results for the dichotomies created by the three cut scores for the August 2015 and January 2016 administrations are presented in Appendices D and E, respectively. The tabled values are derived with the program *BB-Class* (Brennan, 2004) using the Livingston and Lewis method.

4.4 GROUP MEANS (STANDARD 2.17)

Mean scale scores were computed based on reported gender, race/ethnicity, English Language Learner status, economically disadvantaged status, and student with disability status. The results for the August 2015 and January 2016 administrations are reported in Appendices D and E, respectively.

4.5 STATE PERCENTILE RANKINGS

State percentile rankings based on scale score distribution are noted in this section. The percentiles presented here are based on the distribution of all students taking the Regents Examination in Geometry for the June 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. However, the percentile ranking table was not generated, as the Regents Examination in Geometry was not administered in June 2016. The percentile ranks are usually 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.

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 the 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 Regents Examination in Geometry measures student achievement on the New York State Learning Standards for Mathematics. The Geometry standards can be found at: http://www.nysed.gov/curriculum-instruction/new-york-state-next-generation-mathematics-learning-standards.

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 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 2 displays the targeted proportions of content bands on the exam.

Table 2 Test Blueprint, Regents Examination in Geometry

Content Band	Percentage of Total
Geometric Relationships	8–12
Constructions	3–7
Locus	4–8
Informal and Formal Proofs	41–47
Transformational Geometry	8–13
Coordinate Geometry	23–28

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, to ensure that item content is appropriate to the construct being measured and fair for all students. Finally, educators use the defined standards, test blueprint targets, and statistical information generated during field testing to select the highest quality items for use in the operational test.

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 the Department. Both multiple-choice and constructed-response items are included in the Regents Examination in Geometry, to ensure appropriate coverage of the construct domain.

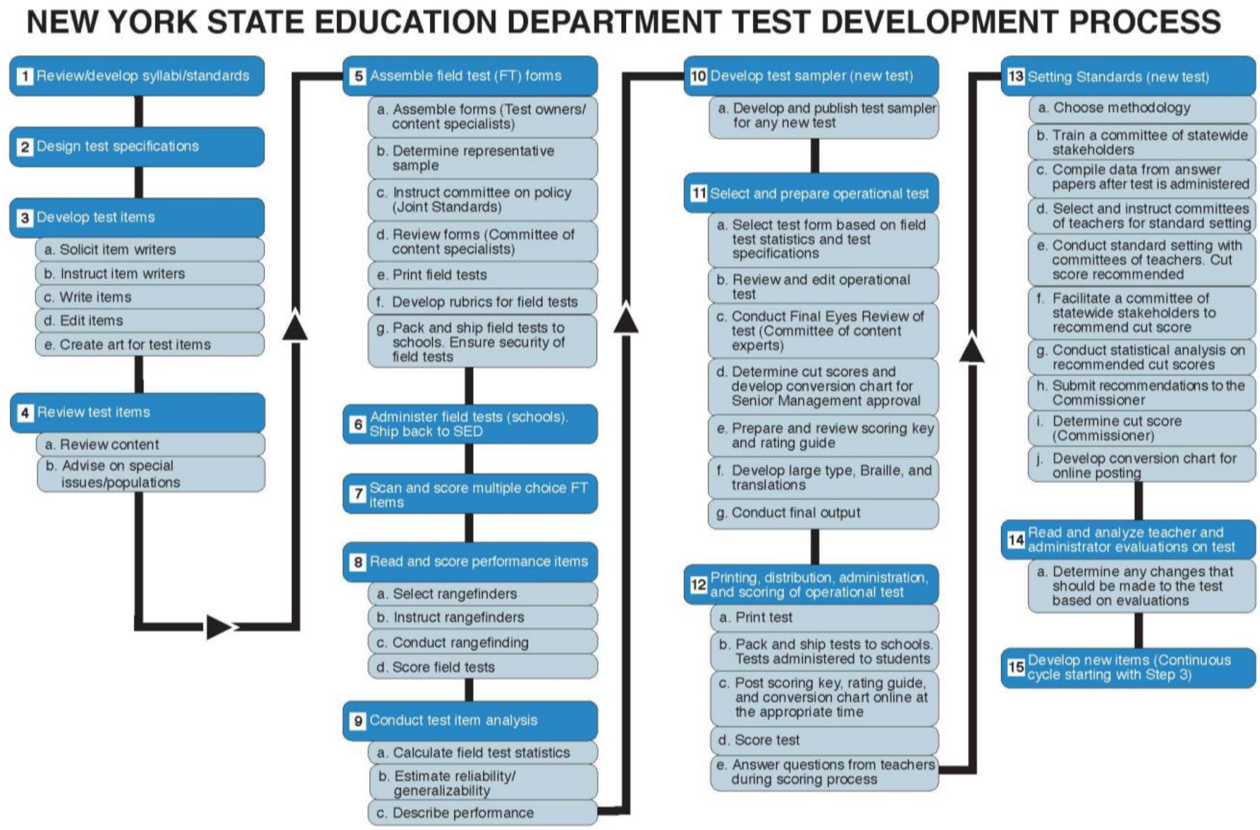

Figure 3 New York State Education Department Test Development Process

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. 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 in 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

Each section of the criteria includes pertinent questions that help reviewers determine whether or not 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 or not 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 (Appendix C). 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/sam/secondary/hssam-update.html), 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 2016 administrations of the Regents Examination in Geometry are located at http://www.nysedregents.org/geometry/.

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, the NYSED training for writing items includes specific attention to rubric development as follows:

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

In support of the goal of valid score interpretations for each examinee, then, such scoring 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 are all negative, indicating that, for the most part, examinees are not being drawn to an unintended construct. INFIT statistics are provided in Table 5. Values for all but two items indicate good model fit.

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 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 Regents Examination in Geometry show a range of *p*-values consistent with the targeted exam difficulty. Refer to section 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. Refer to section 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 2008 through 2015. Sample sizes for subgroups based on ethnicity and English language 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.

Full differential item functioning results are reported in Appendix E (or C) of the 2012–2015 field test 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.

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. Refer to section 4 of this report for additional details.

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 that 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 than expected from the true score. High decision consistency and accuracy provide strong evidence that the internal structure of a test is sound.

For the Regents Examination in Geometry, both decision consistency and accuracy values are high, indicating very 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 section 3 for details of this analysis.

Considering this collection of detailed analyses on the internal structure of the Regents Examination in Geometry, strong evidence exists that the exam is functioning as intended and is providing 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 this 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 are 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., 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. Retrieved February 17, 2016, from www.cse.ucla.edu/products/evaluation/cresst_ec1995_3.pdf.

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.

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 http://www.education.uiowa.edu/casma/computer_programs.htm.

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.

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

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

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

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.

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 2015 Administration

Position	Item Type	Max Points	Weight	Strand	PI	Mean	Point-Biserial	Rasch Difficulty	INFIT
1	MC	1	2	Informal & Formal Proofs	G.G.29	0.73	0.39	-0.9688	1.06
2	MC	1	2	Coordinate Geometry	G.G.73	0.71	0.44	-0.7691	0.95
3	MC	1	2	Informal & Formal Proofs	G.G.52	0.68	0.50	-0.5335	0.87
4	MC	1	2	Transformational Geometry	G.G.54	0.71	0.42	-0.6046	1.00
5	MC	1	2	Informal & Formal Proofs	G.G.25	0.67	0.41	-0.4600	0.99
6	MC	1	2	Transformational Geometry	G.G.57	0.71	0.50	-0.7559	0.89
7	MC	1	2	Informal & Formal Proofs	G.G.46	0.63	0.41	-0.2358	1.00
8	MC	1	2	Geometric Relationships	G.G.4	0.62	0.33	-0.3103	1.06
9	MC	1	2	Coordinate Geometry	G.G.65	0.56	0.51	0.0481	0.90
10	MC	1	2	Informal & Formal Proofs	G.G.31	0.65	0.49	-0.1966	0.92
11	MC	1	2	Informal & Formal Proofs	G.G.45	0.60	0.47	-0.1854	1.01
12	MC	1	2	Constructions	G.G.19	0.64	0.28	-0.1816	1.19
13	MC	1	2	Informal & Formal Proofs	G.G.24	0.59	0.40	-0.0828	1.04
14	MC	1	2	Geometric Relationships	G.G.2	0.58	0.37	-0.0688	1.13
15	MC	1	2	Transformational Geometry	G.G.59	0.60	0.45	-0.0648	1.04
16	MC	1	2	Coordinate Geometry	G.G.70	0.61	0.34	-0.2239	1.13
17	MC	1	2	Informal & Formal Proofs	G.G.41	0.56	0.48	0.0009	1.01
18	MC	1	2	Informal & Formal Proofs	G.G.51	0.57	0.41	0.0177	1.00
19	MC	1	2	Informal & Formal Proofs	G.G.35	0.54	0.43	0.1290	0.96
20	MC	1	2	Coordinate Geometry	G.G.72	0.55	0.55	0.0959	0.91
21	MC	1	2	Informal & Formal Proofs	G.G.37	0.52	0.44	0.2428	1.02
22	MC	1	2	Locus	G.G.22	0.48	0.56	0.3401	0.92
23	MC	1	2	Geometric Relationships	G.G.11	0.50	0.25	0.3092	1.14
24	MC	1	2	Informal & Formal Proofs	G.G.34	0.49	0.47	0.5730	1.02
25	MC	1	2	Informal & Formal Proofs	G.G.50	0.47	0.37	0.5733	1.07
26	MC	1	2	Informal & Formal Proofs	G.G.43	0.47	0.35	0.4449	1.03
27	MC	1	2	Informal & Formal Proofs	G.G.33	0.49	0.46	0.3978	1.00
28	MC	1	2	Informal & Formal Proofs	G.G.47	0.43	0.40	0.7892	1.08
29	CR	2	1	Transformational Geometry	G.G.60	0.99	0.56	0.3779	1.00
30	CR	2	1	Geometric Relationships	G.G.15	0.91	0.66	0.5357	0.96
31	CR	2	1	Informal & Formal Proofs	G.G.44	0.86	0.68	0.5753	0.93
32	CR	2	1	Constructions	G.G.20	0.86	0.51	0.6850	1.00
33	CR	2	1	Coordinate Geometry	G.G.64	0.79	0.73	0.7388	0.81
34	CR	2	1	Coordinate Geometry	G.G.69	0.55	0.66	1.1827	0.95
35	CR	4	1	Locus	G.G.23	1.83	0.77	0.6999	0.93
36	CR	4	1	Coordinate Geometry	G.G.68	0.95	0.73	1.3993	0.83
37	CR	4	1	Coordinate Geometry	G.G.74	1.30	0.78	0.9713	0.76
38	CR	6	1	Informal & Formal Proofs	G.G.38	1.05	0.80	1.6543	0.70

Table A.2 Test Map for January 2016 Administration

Position	Item Type	Max Points	Weight	Strand	PI	Mean	Point-Biserial	Rasch Difficulty	INFIT
1	MC	1	2	Coordinate Geometry	G.G.71	0.77	0.40	-1.0183	1.02
2	MC	1	2	Informal & Formal Proofs	G.G.51	0.72	0.45	-0.8566	0.95
3	MC	1	2	Informal & Formal Proofs	G.G.39	0.56	0.39	0.0562	1.04
4	MC	1	2	Locus	G.G.23	0.61	0.48	-0.2462	0.95
5	MC	1	2	Informal & Formal Proofs	G.G.26	0.64	0.35	-0.3241	1.08
6	MC	1	2	Geometric Relationships	G.G.16	0.64	0.51	-0.3992	0.93
7	MC	1	2	Informal & Formal Proofs	G.G.34	0.59	0.50	-0.1091	0.91
8	MC	1	2	Informal & Formal Proofs	G.G.40	0.53	0.42	0.1629	1.04
9	MC	1	2	Locus	G.G.22	0.58	0.41	-0.0371	1.00
10	MC	1	2	Coordinate Geometry	G.G.62	0.58	0.44	-0.0366	0.95
11	MC	1	2	Informal & Formal Proofs	G.G.42	0.56	0.46	0.1102	0.95
12	MC	1	2	Informal & Formal Proofs	G.G.35	0.64	0.40	-0.1402	1.02
13	MC	1	2	Coordinate Geometry	G.G.63	0.54	0.51	0.1575	0.92
14	MC	1	2	Coordinate Geometry	G.G.74	0.52	0.47	0.2665	0.97
15	MC	1	2	Informal & Formal Proofs	G.G.32	0.51	0.45	0.3374	1.01
16	MC	1	2	Informal & Formal Proofs	G.G.52	0.52	0.41	0.3572	1.06
17	MC	1	2	Transformational Geometry	G.G.54	0.48	0.46	0.4000	1.01
18	MC	1	2	Informal & Formal Proofs	G.G.36	0.44	0.45	0.4957	1.02
19	MC	1	2	Informal & Formal Proofs	G.G.43	0.44	0.49	0.6399	0.95
20	MC	1	2	Coordinate Geometry	G.G.65	0.42	0.32	0.7780	1.13
21	MC	1	2	Geometric Relationships	G.G.10	0.47	0.47	0.6874	1.00
22	MC	1	2	Transformational Geometry	G.G.60	0.43	0.41	0.8292	1.08
23	MC	1	2	Geometric Relationships	G.G.15	0.41	0.36	0.8367	1.08
24	MC	1	2	Informal & Formal Proofs	G.G.29	0.75	0.41	-1.1017	1.06
25	MC	1	2	Informal & Formal Proofs	G.G.33	0.39	0.33	0.8966	1.13
26	MC	1	2	Informal & Formal Proofs	G.G.46	0.35	0.27	0.9100	1.15
27	MC	1	2	Informal & Formal Proofs	G.G.28	0.48	0.48	0.4228	1.00
28	MC	1	2	Constructions	G.G.18	0.37	0.42	0.9997	1.02
29	CR	2	1	Informal & Formal Proofs	G.G.45	1.38	0.63	-0.3998	0.84
30	CR	2	1	Transformational Geometry	G.G.61	1.36	0.65	-0.3311	0.94
31	CR	2	1	Informal & Formal Proofs	G.G.38	1.24	0.60	-0.3033	0.93
32	CR	2	1	Coordinate Geometry	G.G.67	1.02	0.67	0.3162	0.95
33	CR	2	1	Geometric Relationships	G.G.12	0.83	0.52	0.6794	1.09
34	CR	2	1	Constructions	G.G.17	0.81	0.61	0.7143	0.96
35	CR	4	1	Transformational Geometry	G.G.58	2.89	0.68	-0.5753	1.00
36	CR	4	1	Coordinate Geometry	G.G.70	2.00	0.75	0.4767	0.86
37	CR	4	1	Coordinate Geometry	G.G.69	0.87	0.77	1.4350	0.80
38	CR	6	1	Informal & Formal Proofs	G.G.47	0.90	0.79	1.6243	0.77

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

Table B.1 Score Table for August 2015 Administration

Raw Score	Ability	Scale Score
0	−5.4412	0.000
1	−4.2242	3.117
2	−3.5123	6.005
3	−3.0881	8.755
4	−2.7814	11.409
5	−2.5391	14.021
6	−2.3376	16.511
7	−2.1640	18.878
8	−2.0110	21.180
9	−1.8735	23.425
10	−1.7484	25.614
11	−1.6332	27.729
12	−1.5262	29.719
13	−1.4260	31.665
14	−1.3317	33.547
15	−1.2423	35.303
16	−1.1573	37.025
17	−1.0761	38.736
18	−0.9983	40.378
19	−0.9234	41.891
20	−0.8513	43.381
21	−0.7816	44.804
22	−0.7141	46.190
23	−0.6485	47.570
24	−0.5849	48.878
25	−0.5228	50.079
26	−0.4624	51.258
27	−0.4033	52.433
28	−0.3457	53.546
29	−0.2892	54.630
30	−0.2339	55.722
31	−0.1797	56.753
32	−0.1265	57.751
33	−0.0743	58.748
34	−0.0230	59.683
35	0.0275	60.596
36	0.0772	61.509
37	0.1261	62.363
38	0.1742	63.182
39	0.2217	64.004
40	0.2686	64.833
41	0.3148	65.654
42	0.3605	66.410
43	0.4056	67.130
44	0.4502	67.854
45	0.4944	68.589
46	0.5381	69.301
47	0.5814	69.923
48	0.6243	70.633
49	0.6670	71.241
50	0.7093	71.838
51	0.7515	72.535
52	0.7935	73.120
53	0.8354	73.692
54	0.8773	74.263
55	0.9193	74.835
56	0.9614	75.387
57	1.0037	76.013
58	1.0465	76.589
59	1.0897	77.133
60	1.1335	77.670
61	1.1781	78.234
62	1.2235	78.831
63	1.2701	79.365
64	1.3179	79.973
65	1.3672	80.506
66	1.4182	81.101
67	1.4711	81.708
68	1.5263	82.320
69	1.5840	82.970
70	1.6446	83.653
71	1.7084	84.290
72	1.7761	85.011
73	1.8481	85.731
74	1.9254	86.463
75	2.0088	87.273
76	2.0996	88.052
77	2.1996	88.880
78	2.3110	89.757
79	2.4373	90.721
80	2.5834	91.751
81	2.7569	92.821
82	2.9708	94.026
83	3.2491	95.228
84	3.6454	96.654
85	4.3302	98.150
86	5.5291	100.000

Table B.2 Score Table for January 2016 Administration

Raw Score	Ability	Scale Score
0	−5.4012	0.000
1	−4.1822	3.276
2	−3.4682	6.274
3	−3.0427	9.122
4	−2.7356	11.879
5	−2.4938	14.556
6	−2.2935	17.092
7	−2.1220	19.475
8	−1.9716	21.808
9	−1.8374	24.027
10	−1.7161	26.200
11	−1.6050	28.252
12	−1.5025	30.161
13	−1.4070	32.041
14	−1.3176	33.828
15	−1.2332	35.481
16	−1.1532	37.108
17	−1.0770	38.718
18	−1.0040	40.258
19	−0.9338	41.683
20	−0.8662	43.071
21	−0.8008	44.415
22	−0.7372	45.714
23	−0.6754	47.001
24	−0.6151	48.265
25	−0.5561	49.445
26	−0.4984	50.549
27	−0.4418	51.663
28	−0.3863	52.775
29	−0.3317	53.812
30	−0.2779	54.848
31	−0.2250	55.900
32	−0.1728	56.882
33	−0.1214	57.848
34	−0.0706	58.819
35	−0.0205	59.726
36	0.0290	60.624
37	0.0779	61.522
38	0.1261	62.364
39	0.1739	63.177
40	0.2211	63.993

Raw Score	Ability	Scale Score
41	0.2678	64.818
42	0.3140	65.639
43	0.3597	66.397
44	0.4050	67.121
45	0.4499	67.850
46	0.4945	68.591
47	0.5387	69.310
48	0.5826	69.943
49	0.6262	70.660
50	0.6697	71.279
51	0.7129	71.898
52	0.7561	72.599
53	0.7992	73.198
54	0.8423	73.785
55	0.8855	74.375
56	0.9289	74.963
57	0.9725	75.537
58	1.0165	76.206
59	1.0609	76.773
60	1.1060	77.338
61	1.1518	77.890
62	1.1986	78.516
63	1.2464	79.089
64	1.2956	79.689
65	1.3463	80.281
66	1.3988	80.871
67	1.4535	81.509
68	1.5106	82.149
69	1.5706	82.813
70	1.6339	83.549
71	1.7011	84.212
72	1.7729	84.978
73	1.8500	85.750
74	1.9333	86.537
75	2.0241	87.423
76	2.1238	88.250
77	2.2343	89.169
78	2.3579	90.115
79	2.4983	91.176
80	2.6601	92.249
81	2.8511	93.362

Raw Score	Ability	Scale Score
82	3.0838	94.541
83	3.3818	95.750
84	3.7984	97.049
85	4.5032	98.465
86	5.7157	100.000

Appendix C: Item Writing Guidelines

General Rules for Writing Multiple-Choice Items

1. ***Use either a direct question or an incomplete statement as the item stem, whichever seems more appropriate to effective presentation of the item.***
   Some item ideas can be expressed more simply and clearly in the incomplete statement style of question. On the other hand, some items seem to require direct question stems for the most effective expression. Teachers should use the item style that seems most appropriate.

2. ***Items should be written in clear and simple language, with vocabulary kept as simple as possible.***
   Like any other item, the multiple-choice item should be perfectly clear. Difficult and technical vocabulary should be avoided unless essential for the purpose of the question. The important elements should generally appear early in the statement of the item, with qualifications and explanations following.

3. ***Each item should have one and only one correct answer.***
   While this requirement is obvious, it is not always fulfilled. Sometimes, writers produce items involving issues so controversial and debatable that even experts are unable to agree on one correct answer. More often, the trouble is failure to consider the full implications of each response.

4. ***Base each item on a single central problem.***
   A multiple-choice item functions most effectively when the student is required to compare directly the relative merits of a number of specific responses to a definite problem. An item consisting merely of a series of unrelated true-false statements, all of which happen to begin with the same phrase, is unacceptable.

5. ***State the central problem of the item clearly and completely in the stem.***
   The stem should be meaningful by itself. It should be clear and should convey the central problem of the item. It should not be necessary for the student to read and reread all the responses before he/she can understand the basis upon which he/she is to make a choice.

6. ***In general, include in the stem any words that must otherwise be repeated in each response.***
   The stem should contain everything the answers have in common or as much as possible of their common content. This practice serves to make the item shorter, so that it can be read and grasped more quickly.

7. ***Avoid negative statements.***
   Negative statements in multiple-choice items lead to unnecessary difficulties and confusion. Special care must be exercised against the double negative.

8. ***Avoid excessive "window dressing."***

The item should contain only material relevant to its solution, unless selection of what is relevant is part of the problem.

9. ***Make the responses grammatically consistent with the stem and parallel with one another in form.***

10. ***Make all responses plausible and attractive to students who lack the information or ability tested by the item.***

The incorrect responses should be plausible answers. So far as possible, each response should be designed specifically to attract students who have certain misconceptions or who tend to make certain common errors.

11. ***Arrange the responses in logical order, if one exists.***

Where the responses consist of numbers or letters, they should ordinarily be arranged in ascending order. Events should be listed in the order in which they occurred, from earliest to most recent, except when this order would clue the answer. This practice helps insure the student will mark the answer correctly.

12. ***Make the responses independent and mutually exclusive.***

Responses should not be interrelated in meaning. Responses that are not mutually exclusive, aid the student in eliminating wrong answers and reduce the reliability of the item by decreasing the number of effective, functioning responses.

13. ***Avoid extraneous clues.***

Since the student is required to associate one of several alternative responses with the stem, any aspect of the question that provides an extraneous basis for correctly associating the right answer or for eliminating a wrong response constitutes an undesirable clue.

14. ***Avoid using "all of the above" and "none of the above" as alternatives.***

15. ***Avoid using the phrase "of the following" in the stem.***

CHECKLIST OF TEST CONSTRUCTION PRINCIPLES
(Multiple-Choice Items)

		YES	NO
1.	Is the item significant?		
2.	Does the item have curricular validity?		
3.	Is the item presented in clear and simple language, with vocabulary kept as simple as possible?		
4.	Does the item have one and only one correct answer?		
5.	Does the item state one single central problem completely in the stem? (See Helpful Hint below.)		
6.	Does the stem include any extraneous material (“window dressing”)?		
7.	Are all responses grammatically consistent with the stem and parallel with one another in form?		
8.	Are all responses plausible (attractive to students who lack the information tested by the item)?		
9.	Are all responses independent and mutually exclusive?		
10.	Are there any extraneous clues due to grammatical inconsistencies, verbal associations, length of response, etc.?		
11.	Were the principles of Universal Design used in constructing the item?		

HELPFUL HINT

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

1. Cover up the responses and read just the stem.

2. Try to turn the stem into a short-answer question by drawing a line after the last word. (If it would not be a good-short answer item you may have a problem with the stem.)

3. The stem must consist of a statement that contains a verb.

Appendix D: Tables and Figures for August 2015 Administration

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

Item	Number of Students	*p*-Value	SD	Point-Biserial	Point-Biserial Distractor 1	Point-Biserial Distractor 2	Point-Biserial Distractor 3
1	11,120	0.64	0.48	0.25	−0.04	−0.14	−0.20
2	11,120	0.71	0.45	0.38	−0.17	−0.21	−0.22
3	11,120	0.60	0.49	0.45	−0.15	−0.35	−0.11
4	11,120	0.59	0.49	0.37	−0.23	−0.19	−0.11
5	11,120	0.65	0.48	0.33	−0.16	−0.18	−0.17
6	11,120	0.74	0.44	0.39	−0.18	−0.18	−0.24
7	11,120	0.66	0.48	0.31	−0.16	−0.15	−0.15
8	11,120	0.65	0.48	0.20	−0.07	−0.15	−0.13
9	11,120	0.57	0.50	0.42	−0.22	−0.24	−0.13
10	11,120	0.45	0.50	0.41	−0.09	−0.19	−0.30
11	11,120	0.52	0.50	0.35	−0.16	−0.19	−0.15
12	11,120	0.57	0.50	0.33	−0.10	−0.17	−0.21
13	11,120	0.45	0.50	0.33	−0.14	−0.14	−0.15
14	11,120	0.64	0.48	0.38	−0.19	−0.16	−0.21
15	11,120	0.54	0.50	0.36	−0.18	−0.15	−0.20
16	11,120	0.61	0.49	0.39	−0.15	−0.21	−0.20
17	11,120	0.41	0.49	0.37	−0.10	−0.18	−0.17
18	11,120	0.46	0.50	0.24	−0.03	−0.23	−0.02
19	11,120	0.48	0.50	0.27	−0.13	−0.16	−0.10
20	11,120	0.48	0.50	0.44	−0.31	−0.24	−0.04
21	11,120	0.52	0.50	0.39	−0.24	−0.18	−0.11
22	11,120	0.28	0.45	0.40	−0.22	−0.15	−0.06
23	11,120	0.60	0.49	0.11	−0.09	−0.08	0.01
24	11,120	0.41	0.49	0.37	−0.10	−0.18	−0.18
25	11,120	0.43	0.50	0.27	0.00	−0.17	−0.19
26	11,120	0.53	0.50	0.32	−0.12	−0.23	−0.08
27	11,120	0.37	0.48	0.35	−0.03	−0.27	−0.09
28	11,120	0.37	0.48	0.28	−0.11	−0.08	−0.14

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
29	0	2	11,120	0.82	0.79	0.41	0.38
30	0	2	11,120	0.77	0.81	0.38	0.48
31	0	2	11,120	0.61	0.85	0.31	0.55
32	0	2	11,120	0.86	0.90	0.43	0.46
33	0	2	11,120	0.67	0.81	0.34	0.62
34	0	2	11,120	0.49	0.77	0.24	0.59
35	0	4	11,120	0.95	1.24	0.24	0.61
36	0	4	11,120	0.68	1.10	0.17	0.64
37	0	4	11,120	1.17	1.44	0.29	0.71
38	0	6	11,120	0.58	1.23	0.10	0.59

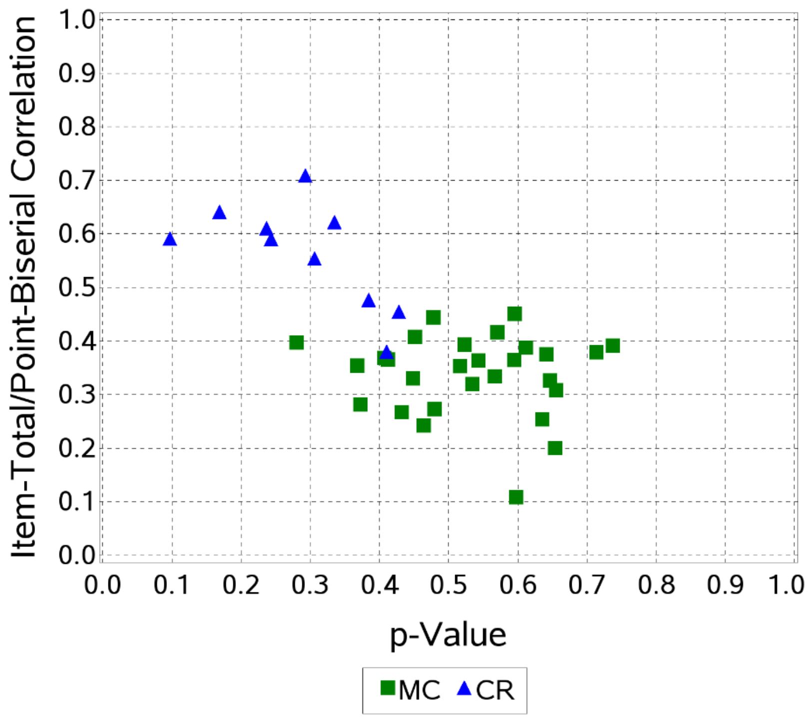

Figure D.1 Scatter Plot: Regents Examination in Geometry

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	38	0.47	0.10	0.37	0.47	0.60	0.74
Point-Biserial	38	0.40	0.11	0.33	0.38	0.45	0.71

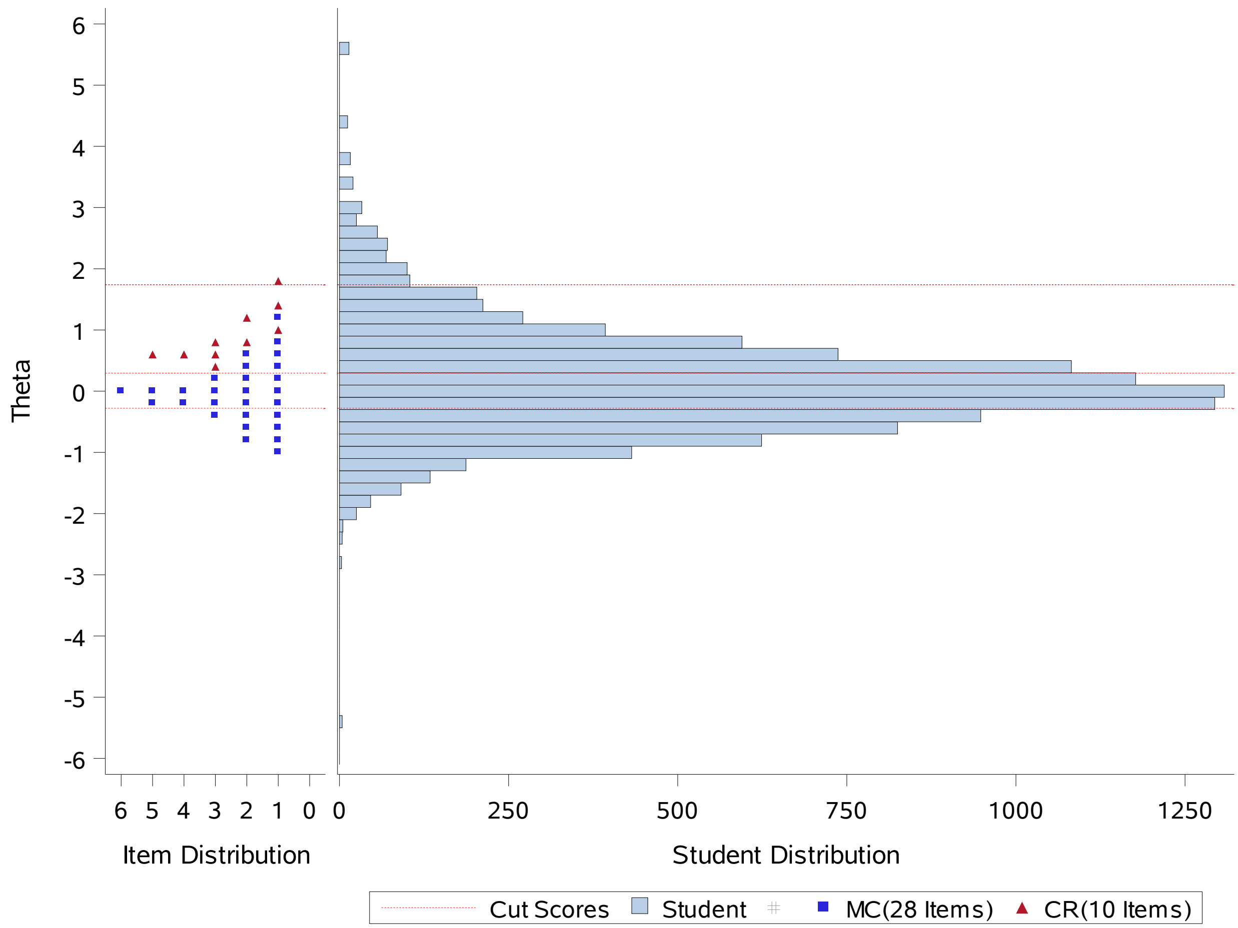

Figure D.2 Student Performance Map: Regents Examination in Geometry

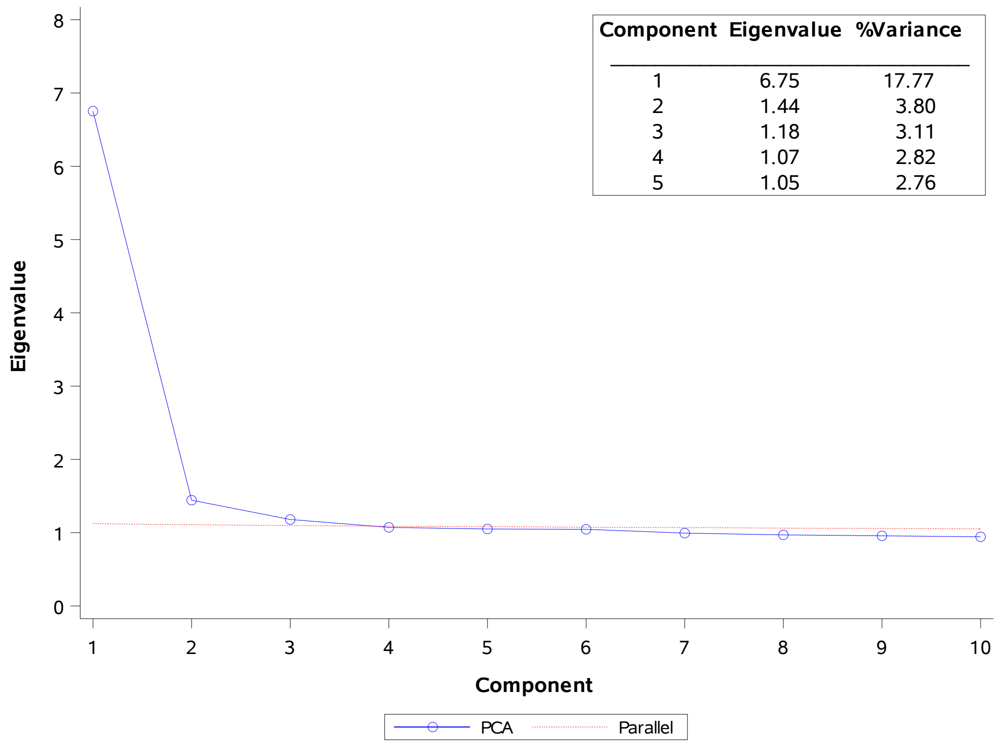

Figure D.3 Scree Plots: Regents Examination in Geometry

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

Statistic Type	Value
N	703
Mean	−0.02
SD	0.03
Minimum	−0.15
P_{10}	−0.06
P_{25}	−0.05
P_{50}	−0.02
P_{75}	0.00
P_{90}	0.01
Maximum	0.14
>|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	38	1.00	0.09	0.79	1.19	[38/38]

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

Subject	Coefficient Alpha	SEM
Geometry	0.87	5.45

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

Statistic	1/2	2/3	3/4
Consistency	0.86	0.84	0.97
Accuracy	0.90	0.89	0.98

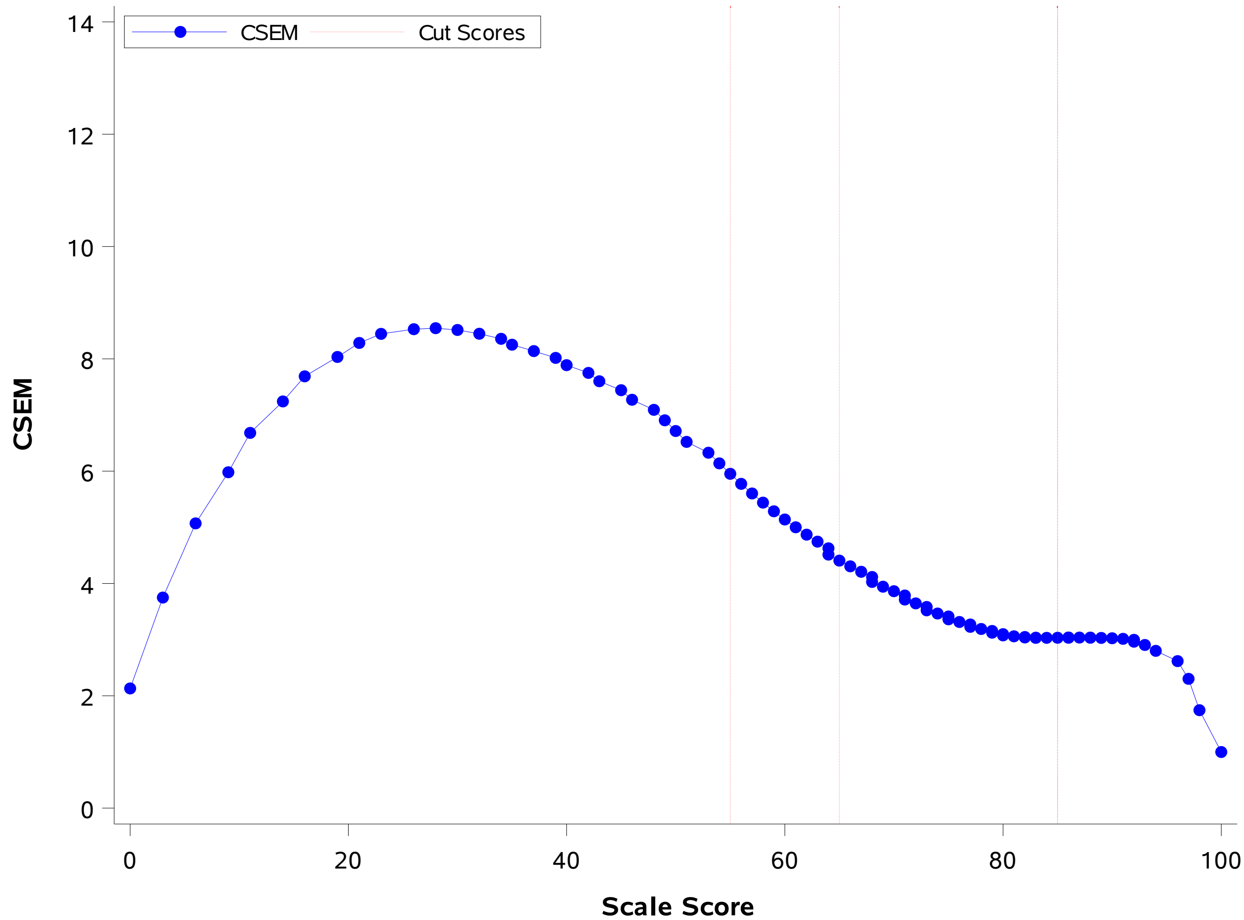

Figure D.4 Conditional Standard Error Plots: Regents Examination in Geometry

Table D.8 Group Means: Regents Examination in Geometry

Demographics	Number	Mean Scale Score	SD Scale Score
All Students*	11,120	60.66	13.05
Ethnicity			
American Indian/Alaska Native	63	60.40	12.93
Asian/Native Hawaiian/Other Pacific Islander	905	65.91	14.77
Black/African American	3,091	56.33	12.23
Hispanic/Latino	2,790	58.12	12.16
Multiracial	132	61.16	12.27
White	4,131	64.46	12.36
English Language Learner			
No	10,778	60.84	12.92
Yes	342	54.82	15.73
Economically Disadvantaged			
No	5,380	62.87	13.32
Yes	5,740	58.58	12.45
Gender			
Female	5,864	60.76	13.12
Male	5,248	60.55	12.98
Student with Disabilities			
No	9,967	61.39	12.93
Yes	1,153	54.32	12.38

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

Appendix E: Tables and Figures for January 2016 Administration

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

Item	Number of Students	*p*-Value	SD	Point-Biserial	Point-Biserial Distractor 1	Point-Biserial Distractor 2	Point-Biserial Distractor 3
1	10,221	0.62	0.49	0.42	−0.13	−0.30	−0.17
2	10,221	0.48	0.50	0.37	−0.04	−0.28	−0.14
3	10,221	0.47	0.50	0.29	−0.11	−0.14	−0.11
4	10,221	0.49	0.50	0.38	−0.11	−0.21	−0.18
5	10,221	0.52	0.50	0.27	−0.19	−0.12	−0.06
6	10,221	0.54	0.50	0.41	−0.10	−0.24	−0.21
7	10,221	0.49	0.50	0.38	−0.17	−0.17	−0.16
8	10,221	0.49	0.50	0.32	−0.16	−0.15	−0.12
9	10,221	0.46	0.50	0.33	−0.08	−0.25	−0.07
10	10,221	0.45	0.50	0.41	−0.13	−0.19	−0.21
11	10,221	0.51	0.50	0.47	−0.19	−0.21	−0.23
12	10,221	0.56	0.50	0.28	−0.11	−0.14	−0.14
13	10,221	0.53	0.50	0.39	−0.18	−0.14	−0.21
14	10,221	0.52	0.50	0.44	−0.20	−0.21	−0.19
15	10,221	0.43	0.49	0.37	−0.20	−0.16	−0.09
16	10,221	0.40	0.49	0.30	−0.14	−0.14	−0.10
17	10,221	0.39	0.49	0.39	−0.15	−0.19	−0.20
18	10,221	0.39	0.49	0.30	−0.12	−0.15	−0.11
19	10,221	0.39	0.49	0.40	−0.13	−0.27	−0.07
20	10,221	0.44	0.50	0.30	−0.11	−0.17	−0.12
21	10,221	0.32	0.47	0.35	−0.12	−0.13	−0.14
22	10,221	0.32	0.47	0.38	−0.18	−0.13	−0.13
23	10,221	0.36	0.48	0.23	−0.11	−0.04	−0.11
24	10,221	0.73	0.44	0.37	−0.18	−0.21	−0.17
25	10,221	0.38	0.49	0.25	−0.07	−0.11	−0.13
26	10,221	0.30	0.46	0.22	−0.12	0.04	−0.18
27	10,221	0.33	0.47	0.29	−0.14	−0.13	−0.03
28	10,221	0.26	0.44	0.31	−0.11	−0.08	−0.11

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
29	0	2	10,221	0.94	0.96	0.47	0.47
30	0	2	10,221	1.10	0.90	0.55	0.47
31	0	2	10,221	0.91	0.71	0.45	0.53
32	0	2	10,221	0.67	0.89	0.34	0.57
33	0	2	10,221	0.64	0.81	0.32	0.31
34	0	2	10,221	0.43	0.78	0.21	0.45
35	0	4	10,221	2.43	1.31	0.61	0.48
36	0	4	10,221	1.41	1.61	0.35	0.56
37	0	4	10,221	0.35	1.01	0.09	0.63
38	0	6	10,221	0.43	1.28	0.07	0.57

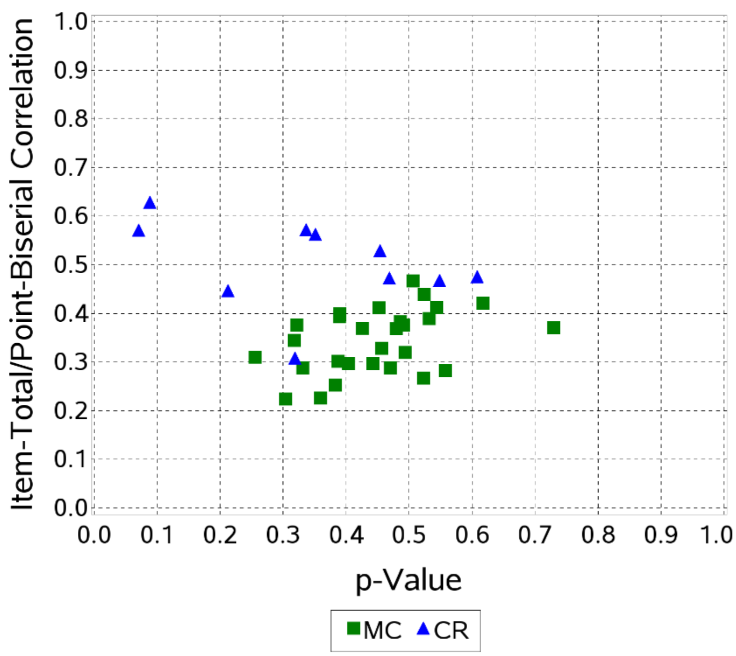

Figure E.1 Scatter Plot: Regents Examination in Geometry

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	38	0.42	0.07	0.34	0.45	0.51	0.73
Point-Biserial	38	0.39	0.22	0.30	0.38	0.45	0.63

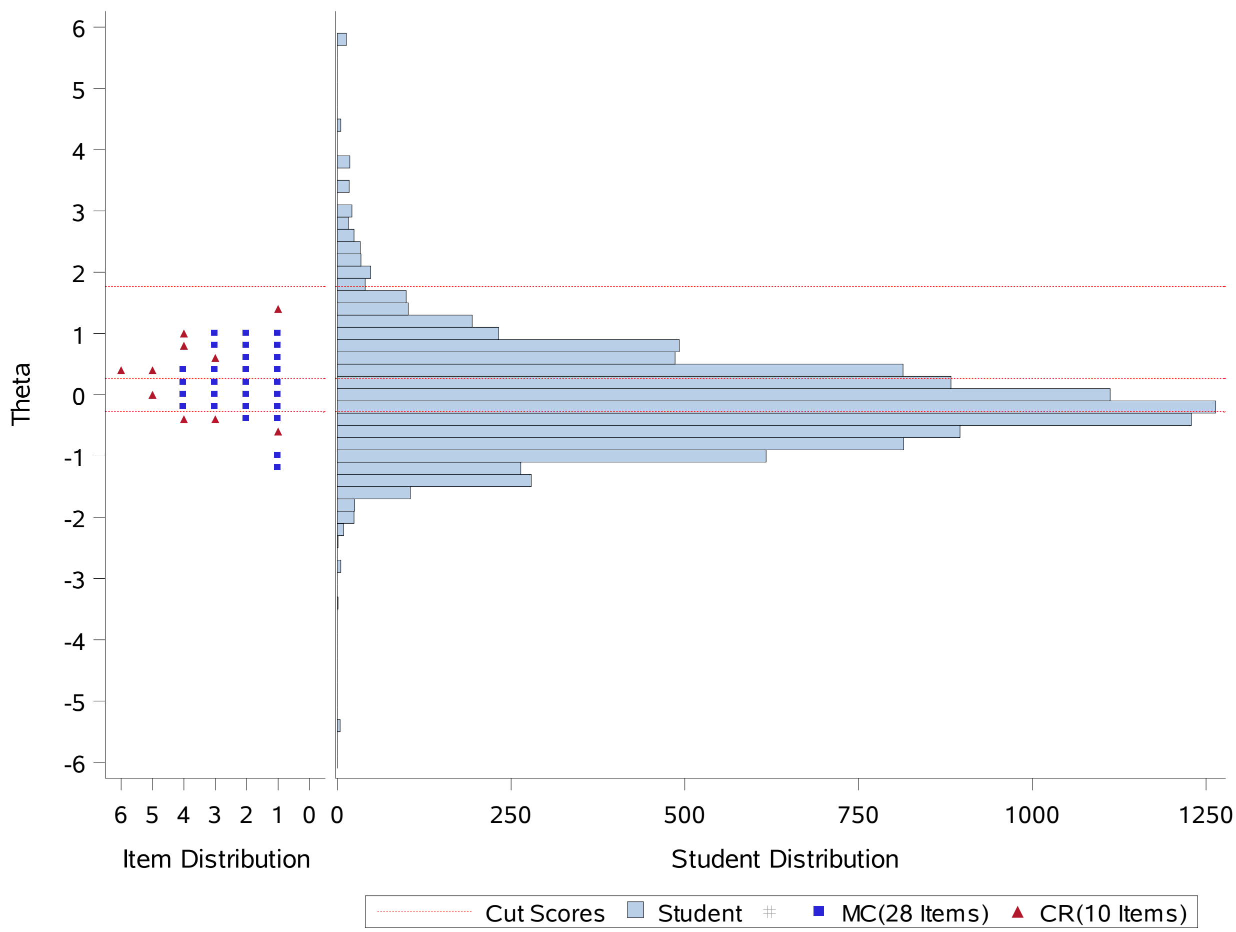

Figure E.2 Student Performance Map: Regents Examination in Geometry

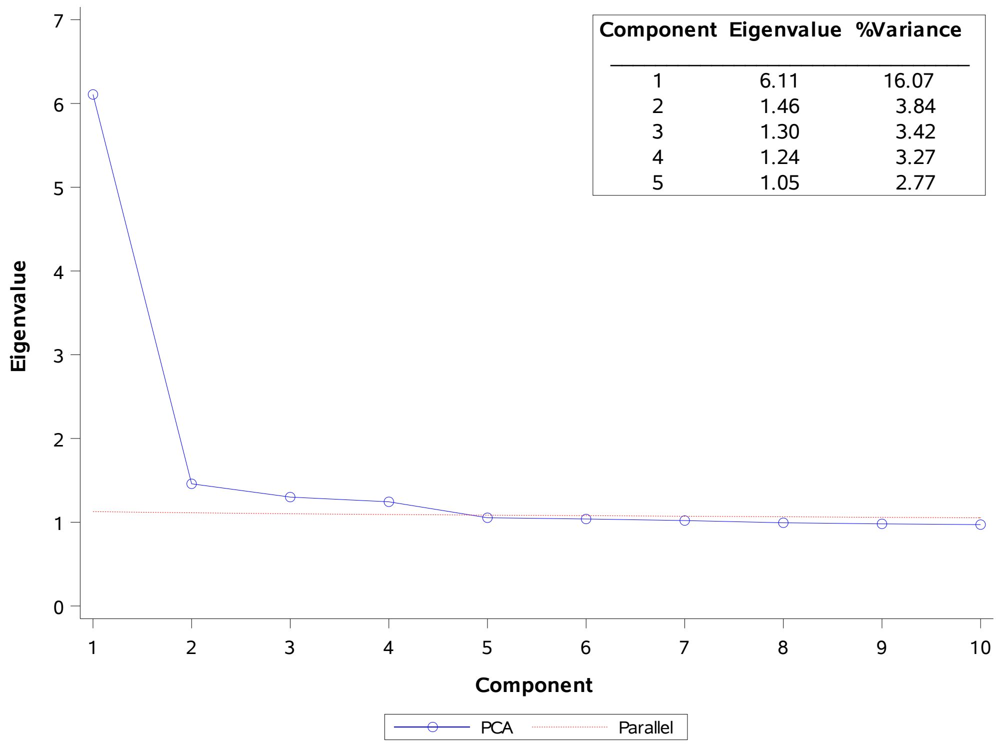

Figure E.3 Scree Plots: Regents Examination in Geometry

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

Statistic Type	Value
N	703
Mean	−0.02
SD	0.04
Minimum	−0.15
P_{10}	−0.07
P_{25}	−0.04
P_{50}	−0.02
P_{75}	0.00
P_{90}	0.02
Maximum	0.22
>|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	38	1	0.08	0.78	1.19	[38/38]

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

Subject	Coefficient Alpha	SEM
Geometry	0.85	5.74

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

Statistic	1/2	2/3	3/4
Consistency	0.83	0.86	0.98
Accuracy	0.88	0.90	0.98

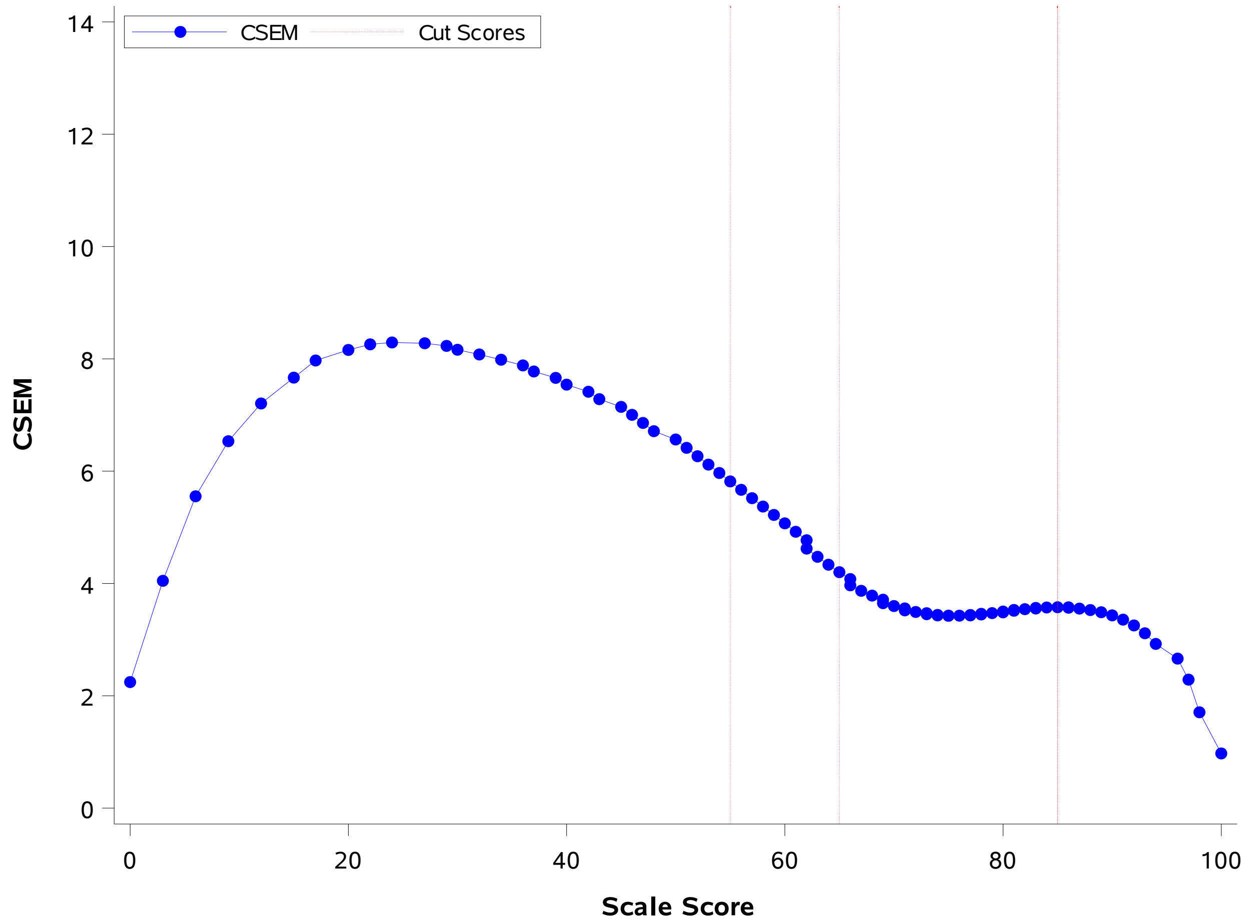

Figure E.4 Conditional Standard Error Plots: Regents Examination in Geometry

Table E.8 Group Means: Regents Examination in Geometry

Demographics	Number	Mean Scale Score	SD Scale Score
All Students*	10,221	57.20	13.42
Ethnicity			
American Indian/Alaska Native	84	60.81	14.36
Asian/Native Hawaiian/Other Pacific Islander	1,141	63.17	15.72
Black/African American	3,136	55.12	12.63
Hispanic/Latino	3,593	55.06	12.38
Multiracial	110	59.36	12.47
White	2,155	60.40	13.24
English Language Learner			
No	9,462	57.37	13.07
Yes	759	55.01	17.01
Economically Disadvantaged			
No	3,408	58.84	13.73
Yes	6,813	56.38	13.18
Gender			
Female	5,514	57.46	13.09
Male	4,705	56.90	13.79
Student with Disabilities			
No	9,066	57.94	13.32
Yes	1,155	51.39	12.75

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