New York State Regents Examination in Geometry (Common Core)

2015 Technical Report

Prepared for the New York State Department of Education

by

Data Recognition Corporation

April 2016

Developed and published under contract with the New York State Education Department by Data Recognition Corporation.

Contents

List of Tables

List of Figures

Chapter 1: Introduction and History

1.1 Introduction

This technical report for the Regents Examination in Geometry (Common Core) will provide the state of New York 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. 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.

1.2 History

The Board of Regents adopted the Common Core State Standards (CCSS) for English Language Arts & Literacy and Mathematics at its July 2010 meeting and incorporated New York State-specific additions, creating the Common Core Learning Standards (CCLS), at its January 2011 meeting. In order to ensure adequate notice and time for students to be prepared to take the new Regents Exams measuring the CCLS, and based on feedback from the field, the Department provided an overlap in the administration of the Regents Exams measuring the 2005 Learning Standards with the Regents Exams measuring the CCLS and a phased-in sequence.

Students who took the old Regents Exam in addition to the new Regents Exam were allowed to use the higher of the two scores for local transcript purposes, and, similarly, the higher of the two scores was used for institutional accountability for the 2013–14 and 2014–15 school year results. Such students were able to meet the Geometry exam requirement for graduation by passing either of these exams. The complete memo detailing transition to the Common Core examinations can be located at http://www.p12.nysed.gov/assessment/commoncore/archive/transitionccregents1113rev-arc2.pdf.

1.3 Purposes of the Exam (Standard 12.1)

The Regents Examination in Geometry (Common Core) measures examinee achievement against the New York State (NYS) learning standards. The exam is prepared by educator examination committees and New York Department of Education subject 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 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 (Common Core) is intended for use in satisfying State testing requirements towards a NYS diploma for students who have finished a course of instruction 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:

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

http://www.p12.nysed.gov/ciai/gradreq/2015GradReq11-15.pdf. Results of the Regents Examination in Geometry (Common Core) may also be used to satisfy various locally established requirements throughout the State.

1.4 Target Population (Standard 7.2)

The examinee population for the Regents Examination in Geometry (Common Core) is composed of students who have completed a course of study in Geometry. Any student, regardless of grade level or cohort, who began their first commencement-level Geometry course in fall 2013 or later was provided with instruction aligned with the NYS P–12 Common Core Learning Standards for Geometry and therefore took or will take the Regents Examination in Geometry (Common Core). More information about testing requirements can be found at http://www.p12.nysed.gov/assessment/commoncore/transitionccregents1113rev.pdf.

Table 1 provides a demographic breakdown of all students who took the June 2015 Regents Examination in Geometry (Common Core). All analyses in this report are based on the population described in Table 1. Annual Regents Examination results in the New York State Report Cards are those reported in the Student Information Repository System (SIRS) as of the reporting deadline (see http://data.nysed.gov/). If a student takes the same exam multiple times in the year, 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 publically reported counts of student test-takers.

Table 1 Total Examinee Population: Regents Examination in Geometry (Common Core)

Demographics	Number	Percent
All Students*	112768	100
Race/Ethnicity		
American Indian or Alaska Native	524	0.46
Asian/Native Hawaiian/Other Pacific Islander	14443	12.81
Black or African American	16015	14.20
Hispanic or Latino	20436	18.12
Multiracial	1604	1.42
White	59744	52.98
English Language Learners		
No	109474	97.08
Yes	3294	2.92
Economically Disadvantaged		
No	66814	59.25
Yes	45954	40.75
Gender		
Female	59126	52.43
Male	53640	47.57
Student with Disabilities		
No	106001	94.00
Yes	6767	6.00

*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 (Common Core) 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).

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

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 a multiple-choice item has a *p*-value of 0.89, it means that 89 percent of the students answered the item correctly. This value might suggest that the item was relatively easy and/or the students who attempted the item were relatively high achievers. For constructed-response (CR) items, mean scores can range from the minimum possible score (usually zero) to the maximum possible score. To facilitate average score comparability across MC and CR items, mean item performance for CR items is divided by the maximum score possible so that the *p*-values for all items are reported as a ratio from 0.0 to 1.0.

Although the *p*-value statistic does not consider individual student ability in its computation, it provides a useful view of overall item difficulty and can provide an early and simple indication of items that are too difficult for the population of students taking the examination. Items with very high or very low *p*-values receive added scrutiny during all follow-up analyses, including item response theory analyses that factor student ability into estimates of item difficulty. Such items may be removed from the item pool during the test development process as field testing may reveal that they add insufficient measurement information. Items for the 2014 Regents Examination in Geometry (Common Core) show a range of *p*-values consistent with the targeted exam difficulty. Item *p*-values range from .30 to .85, with a mean of .63. Further, the point biserial values (discussed in the following section) for these items indicate that they are generally discriminating performance on the test well.

Refer to Tables 2 and 3 for item-by-item *p*-values for multiple-choice and constructed-response items respectively.

2.2 Item Discrimination

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

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

poorly on the exam overall) would be more likely to answer the same item incorrectly. Pearson's product-moment correlation coefficient (also commonly referred to as a point biserial correlation) between item scores and test scores is used to indicate discrimination (Pearson, 1896). The correlation coefficient can range from −1.0 to +1.0. If high-scoring students tend to get the item right while low-scoring students do not, the correlation between the item score and the total test score will be both positive and noticeably large in its magnitude (i.e., above zero), meaning the item is likely discriminating well between high- and low-performing students. Point biserials are computed for each answer option, including correct and incorrect options (commonly referred to as "distractors").Point biserial values for each distractor are an important part of test analysis. Point biserial values on distractors are typically negative. Positive point biserial values can indicate that higher performing students are selecting an incorrect answer or that the item key for the correct answer should be checked.

Refer to Tables 2 and 3 for point biserial values on the correct response and three distractors (Table 2 only). The point biserial values for correct answers are all greater than .15, indicating acceptable discrimination between high- and low-performing examinees. Point biserials for all but two distractor (items 20 and 23) are negative; however, the positive value for these items is very small (at 0.00 and 0.02). This indicates that examinees are responding to the items as expected during item and rubric development.

Table 2 Multiple-Choice Item Analysis Summary: Regents Examination in Geometry (Common Core)

Item	Number	*p*-Value	SD	Point Biserial	Point Biserial Distractor 1	Point Biserial Distractor 2	Point Biserial Distractor 3
1	112768	.47	.50	.49	-.10	-.17	-.35
2	112768	.90	.30	.33	-.19	-.20	-.16
3	112768	.61	.49	.26	-.04	-.21	-.13
4	112768	.88	.32	.21	-.16	-.09	-.10
5	112768	.64	.48	.43	-.22	-.26	-.13
6	112768	.93	.26	.18	-.06	-.16	-.07
7	112768	.60	.49	.34	-.16	-.22	-.13
8	112768	.62	.49	.45	-.21	-.23	-.22
9	112768	.62	.48	.46	-.19	-.25	-.22
10	112768	.60	.49	.40	-.11	-.22	-.24
11	112768	.47	.50	.33	.03	-.24	-.28
12	112768	.69	.46	.47	-.23	-.21	-.27
13	112768	.18	.38	.25	-.10	-.04	-.07
14	112768	.49	.50	.26	-.10	-.16	-.08
15	112768	.68	.47	.27	-.19	-.13	-.11
16	112768	.56	.50	.43	-.15	-.28	-.19
17	112768	.75	.44	.42	-.15	-.22	-.27

Item	Number	*p*-Value	SD	Point Biserial	Point Biserial Distractor 1	Point Biserial Distractor 2	Point Biserial Distractor 3
18	112768	.43	.50	.18	-.19	-.19	.03
19	112768	.42	.49	.39	-.18	-.31	-.03
20	112768	.37	.48	.36	-.17	.00	-.25
21	112768	.60	.49	.46	-.13	-.22	-.30
22	112768	.50	.50	.37	-.22	-.14	-.14
23	112768	.30	.46	.43	.02	-.30	-.15
24	112768	.37	.48	.36	-.22	-.18	-.04

Table 3 Constructed-Response Item Analysis Summary: Regents Examination in Geometry (Common Core)

Item	Min. score	Max. score	Number of Students	Mean	SD	*p*-Value	Point Biserial
25	0	2	112768	0.84	0.96	0.42	0.61
26	0	2	112768	1.25	0.80	0.63	0.56
27	0	2	112768	0.90	0.90	0.45	0.54
28	0	2	112768	0.87	0.90	0.44	0.69
29	0	2	112768	0.57	0.85	0.29	0.65
30	0	2	112768	1.44	0.78	0.72	0.47
31	0	2	112768	0.60	0.82	0.30	0.58
32	0	4	112768	1.62	1.61	0.40	0.70
33	0	4	112768	1.77	1.46	0.44	0.74
34	0	4	112768	1.63	1.43	0.41	0.70
35	0	6	112768	1.52	2.00	0.25	0.74
36	0	6	112768	1.90	2.17	0.32	0.76

2.3 Discrimination on Difficulty Scatterplots

Figure 1 shows a scatterplot of item difficulty values (*x*-axis) and item discrimination values (*y*-axis). The distributions of *p*-value and point biserials are also included in the graphic to illustrate the mean, median, total range, and quartile ranges for each.

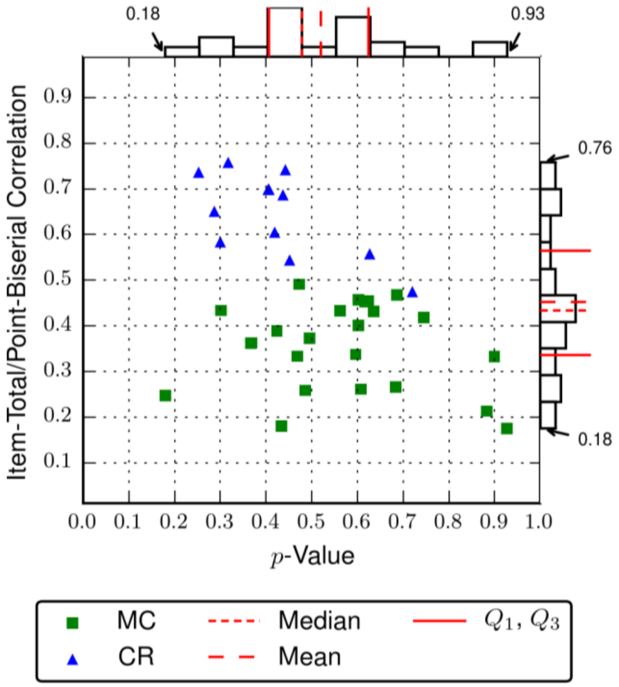


Figure 1 Scatterplot: Regents Examination in Geometry (Common Core)

2.4 Observations and Interpretations

The *p*-values for the MC items ranged from 0.18 to 0.93, with an average of 0.52 while proportion-correct values for the constructed response items (Table 3) were 0.25 and 0.72. The difficulty distribution illustrated in Figure 1 shows a wide range of item difficulties on the exam. This is consistent with general test development practice which seeks to measure student ability along a full range of difficulty.

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

The item response theory (IRT) model used for the Regents Examination in Geometry (Common Core) is based on the work of George 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 through 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 (Common Core) items. Generally, item calibration is the process of assigning a difficulty or item "location" estimate to each item in an assessment so that all items are placed onto a common scale. This chapter briefly introduces the Rasch model, reports the results from evaluations of the adequacy of the Rasch assumptions, and summarizes the Rasch item statistics.

3.1 Description of the Rasch Model

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

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

where θ_n represents examinee ability, and D_{ij} is the step difficulty of the j^{th} step on item *i*. 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:

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

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

3.2 Software and Estimation Algorithm

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

3.3 Characteristics of the Testing Population

The data analyses reported here are based on all students who took the Regents Examination in Geometry (Common Core) in June 2014. The characteristics of this population are provided in Table 1 Total Examinee Population: Regents Examination in Geometry (Common Core).

3.4. Item Difficulty-Student Performance Maps

The distributions of the Rasch item logits (item difficulty estimates) and student performance are shown on the item difficulty–student performance map presented in Figure 2. This graphic illustrates the location of student performance and item difficulty on the same scale, along with their respective distributions and cut scores (indicated by the horizontal dotted lines). The figure shows more difficult items and higher examinee performance at the top and lower performance and easier items at the bottom. Figure 2 also demonstrates that measurement precision tends to be higher at the critical cut scores based on a concentration of items and students at these locations.

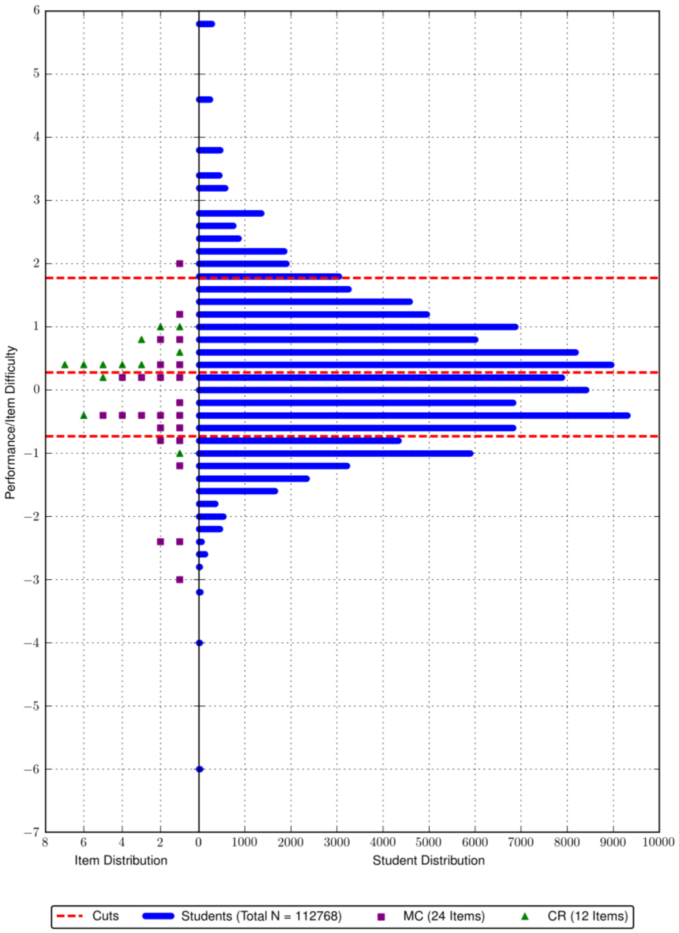


Figure 2 Student Performance Map: Regents Examination in Geometry (Common Core)

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 (Common Core), the validity of the inferences from these results depends on the degree to which the assumptions of the model were met and how well the model fits the test data. Therefore, it is important to check these assumptions. This section evaluates the dimensionality of the data, local item independence, and item fit. It should be noted that only operational items were analyzed, since they are the basis of student scores.

Unidimensionality

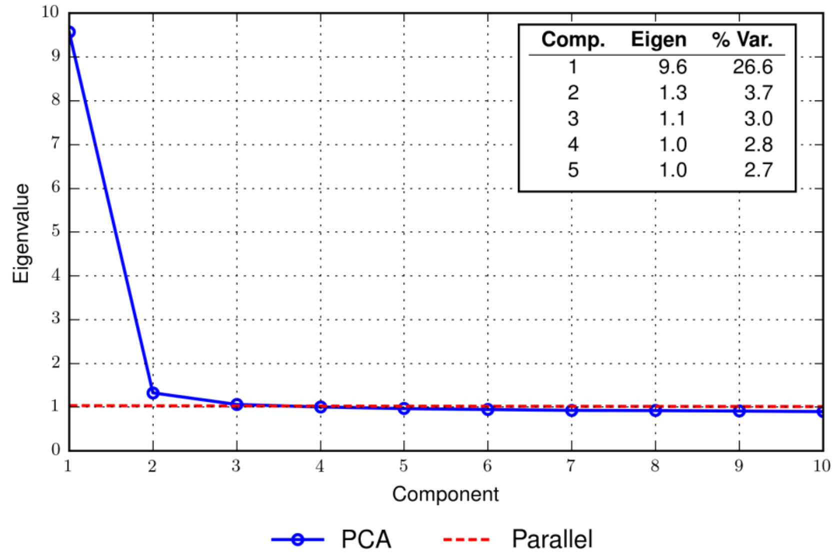
Rasch models assume that one dominant dimension determines the differences in student performance. Principal Components Analysis (PCA) can be used to assess the unidimensionality assumption. The purpose of the analysis is to verify whether any other dominant components exist among the items. If any other dimensions are found, the unidimensionality of test content assumption may be violated.

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

Figure 3 shows the PCA and parallel analysis results for the Regents Examination in Geometry (Common Core). The results include the eigenvalues and the percentage of variance explained for the first five components as well as the scree plots. The scree plots show the eigenvalues plotted by component number and the results from a parallel analysis. Although the total number of components in PCA is same as the total number of items in a test, Figure 3 shows only 10 components. This view is sufficient for interpretation because components are listed in descending eigenvalue order. The lower eigenvalues from components 2 through 10 demonstrates that components beyond 1 are not individually contributing to the explanation of variance in the data.

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

As shown in Figure 3, the primary dimension explained 26.6 percent of the total variance for the Regents Examination in Geometry (Common Core). The eigenvalue of the second dimension less than one third of the first at 1.3, and the second value is not significantly different from the parallel analysis. Overall, the PCA suggests that the test is reasonably unidimensional.



Comp.	Eigen	% Var.
1	9.6	26.6
2	1.3	3.7
3	1.1	3.0
4	1.0	2.8
5	1.0	2.7

Figure 3 Scree Plots: Regents Examination in Geometry (Common Core)

Local Independence

Local independence (LI) is a fundamental assumption of IRT. This means simply, 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 (cf. 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 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:

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

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 violation occurs when responses to an item depend on responses to another item. This is a violation of local 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 among the Regents Examination in Geometry (Common Core) items. In general, these residuals are computed as follows. First, expected item performance based on the Rasch model is determined using (θ) and item parameter estimates. Next, deviations (residuals) between the examinees' expected and observed performance are determined for each item. Finally, for each item pair, a correlation between the respective deviations is computed.

Three types of residual correlations are available in WINSTEPS: raw, standardized, and logit. It is noted that the raw score residual correlation essentially corresponds to Yen's Q_3 index, a popular statistic used to assess local independence. The expected value for the Q_3 statistic is approximately $-1/(k-1)$ when no local dependence exists, where k is test length (Yen, 1993). Thus, the expected Q_3 values should be approximately -0.03 for the items on the exam. Index values that are greater than

0.20 indicate a degree of local dependence that probably should be examined by test developers (Chen & Thissen, 1997).

Since the three residual correlations are very similar, the default "standardized residual correlation" in WINSTEPS was used for these analyses. Table 4 shows the summary statistics—mean, standard deviation, minimum, maximum, and several percentiles (P10, P25, P50, P75, P90) — for all the residual correlations for each test. The total number of item pairs (N) and the number of pairs with the residual correlations greater than 0.20 are also reported in this table. There were no item pairs with residual correlations greater than 0.20. The mean residual correlations were slightly negative and the values were close to −0.03. The vast majority of the correlations were very small, suggesting local item independence generally holds for the Regents Examination in Geometry (Common Core).

Table 4 Summary of Item Residual Correlations: Geometry (Common core)

Statistic Type	Value
N	630
Mean	−0.02
SD	0.03
Minimum	−0.12
P_{10}	−0.06
P_{25}	−0.04
P_{50}	−0.02
P_{75}	0.00
P_{90}	0.01
Maximum	0.14
>|0.20|	0

Item Fit

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

The expected MnSq value is 1.0 and can range from 0.0 to infinity. Deviation in excess of the expected value can be interpreted as either noise or lack of fit between the items and the model. Values lower than the expected value can be interpreted as item redundancy or overfitting items (too predictable, too much redundancy), and values greater than the expected value indicate underfitting items (too unpredictable, too much noise). Rules of thumb regarding "practically significant" MnSq values vary. Table 5 presents the summary statistics of infit mean square statistics for the Regents Examination in Geometry (Common Core), including the mean, standard deviation, and minimum and maximum values. The number of items within a targeted range of [0.7, 1.3] is also reported in Table 5. The mean infit value is 1.00, with all items falling in a targeted range of [0.7, 1.3]. As the range of [0.7, 1.3] is used as guide for ideal fit, fit values outside of the range are considered individually. A finding of 36 out of 36 items falling in the ideal fit range indicates that the Rasch model fits the Regents Examination in Geometry (Common Core) item data well.

Table 5 Summary of Infit Mean Square Statistics: Geometry (Common Core)

	Infit Mean Square				
	Mean	**SD**	**Min**	**Max**	**[0.7, 1.3]**
Geometry	1.00	0.09	0.80	1.25	36/36

Items for the Regents Examination in Geometry (Common Core) were field tested in 2012-2014, and separate technical reports for each year were produced to document the full test development, scoring, scaling, and data analysis conducted. Please refer to http://www.p12.nysed.gov/assessment/reports for details.

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.

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 may exist. 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). Examples of such factors that can influence reliability estimates include test length and the variability of observed scores. Tests with fewer items or with a lack of heterogeneity in scores tend to produce lower reliability estimates.

The remainder of this chapter discusses reliability results for Regents Examination in Geometry (Common Core) 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. It is defined as the proportion of true score variance contained in the observed scores. The total variance contains two components: 1) the variance in true scores and 2) the variance due to the imperfections in the measurement process. 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).

Reliability coefficients range from 0.0 to 1.0. The index will be 0.0 if none of the test score variances are true. Such scores would be pure random noise (i.e., all measurement error). If all test score variances were true, the index would equal 1.0. If the index achieved a value of 1.0, scores would be perfectly consistent (i.e., contain no measurement error). Although values of 1.0 are never achieved in

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

practice, it is clear that larger coefficients are more desirable because they indicate that test scores are less influenced by random error.

Coefficient Alpha

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

A general computational formula for 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*. Table 6 in section 4.2 displays the coefficient alpha for the in Regents Examination in Geometry (Common Core), along with the standard error of measurement (SEM).

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 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 (SEM) 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 an SEM of 3 on a 10-point test would be very different than an 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 raw 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 (Common Core) are provided in Table 6.

Table 6 Reliabilities and Standard Errors of Measurement: Regents Examination in Geometry (Common Core)

Subject	Coefficient Alpha	*SEM*
Geometry	0.91	5.82

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

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

Conditional Standard Error of Measurement

Every time an assessment is administered, the score the student receives contains some error. If the same exam were administered an infinite number of times to the same student, the mean of the distribution of the student's raw scores would be equal to their true score (θ, the score obtained with no error), and the standard deviation of the distribution of their raw scores would be the conditional standard error. Since there is a one-to-one correspondence between the raw score and θ in the Rasch model, we can apply this concept more generally to all students who obtained a particular raw score, and calculate the probability of obtaining each possible raw score given the 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 along with the nonlinear relationship between the expected raw and θ scores. In addition, as the exam is equated from year to year, the relationship between the raw and scale scores moves away from the original third degree polynomial relationship to one that is also no longer expressible in simple mathematical form. In the absence of a simple mathematical relationship between θ and the scale scores, the CSEMs that are available for each θ score via Rasch IRT cannot be converted directly to the scale score metric.

[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.

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 (Common Core) that do not have a one-to-one relationship between raw 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 towards the middle of the distribution.

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

Results and Observations

The CSEMs for the Regents Exams can be expected to have inverted-U shaped patterns, with some variations. The relationship between raw and scale scores for the Regents Exams tends to be roughly linear from scale scores of 0 to 79 and then concave down from about 79 to 100. In other words, the scale scores track linearly with the raw scores for about the lower 80 percent of the scale score range and then are compressed relative to the raw scores for about the remaining 20 percent of the range, though there are variations.

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

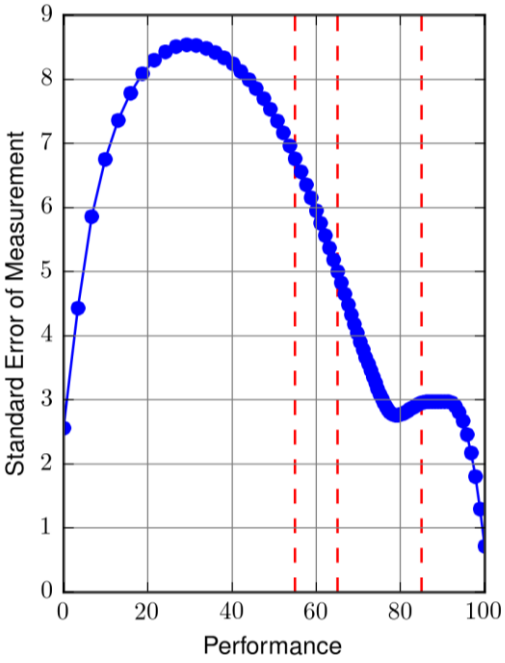


Figure 4 Conditional Standard Error Plots: Regents Examination in Geometry (Common Core)

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 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 made from 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 5 Pseudo-Decision Table for Two Hypothetical Categories

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

Figure 6 Pseudo-Decision Table for Four Hypothetical Categories

If a student is classified as being in one category based on Test One's score, how probable would it be that the student would be reclassified as being in the same category if he or she took Test Two (a non-overlapping, equally difficult form of the test)? This proportion is a measure of decision consistency.

The proportions of correct decisions, φ, for two and four categories are computed by the following two formulas, respectively:

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

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

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

Classification accuracy refers to the agreement of the observed classifications of students with the classifications made on the basis of their true scores. As discussed above, an observed score contains measurement error while a true score is theoretically free of measurement error. A student's observed score can be formulated by the sum of his or her true score plus measurement error, or $Observed = True + Error$. Decision accuracy is an index to determine the extent to which measurement error causes a classification different than 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 (Common Core), 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 for four performance levels should be lower than those based on two categories. This is not surprising, since classification and accuracy using four levels would allow more opportunity to change achievement levels. Hence, there would be more classification errors and less accuracy with four achievement levels, resulting in lower consistency indices.

Results and Observations The results for the dichotomies created by the four cut scores, are presented in Table 7. The tabled values are derived with the program *BB-Class* (Brennan, 2004) using the Livingston and Lewis method. The overall decision consistency ranged from 0.87 to 0.94, and the decision accuracy ranged from 0.91 to .96. Both decision consistency and accuracy values indicate good consistency and accuracy of examinee classifications.

Table 7 Decision Consistency and Accuracy Results: Regents Examination in Geometry (Common Core)

Statistic	1/2	2/3	3/4
Consistency	0.87	0.87	0.94
Accuracy	0.90	0.91	0.96

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 are reported in Table 8.

Table 8 Group Means: Regents Examination in Geometry (Common Core)

Demographics	Number	Mean Scale-score	Standard error of group
All Students*	112768	68.29	15.20
Ethnicity			
American Indian/Alaska Native	524	65.68	13.69
Asian/Native Hawaiian/Other Pacific Islander	14443	74.84	14.70
Black/African American	16015	57.77	14.37
Hispanic/Latino	20436	60.21	14.40
Multiracial	1604	69.52	14.97
White	59744	72.28	13.09
English Language Learner			
No	109474	68.65	14.96
Yes	3294	56.29	17.96
Economically Disadvantaged			
No	66814	72.05	13.96
Yes	45954	62.83	15.27
Gender			
Female	59126	68.03	15.12
Male	53640	68.58	15.28
Student with Disabilities			
No	106001	69.11	14.79
Yes	6767	55.40	15.64

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

4.5 State Percentile Rankings

State percentile rankings based on raw score distributions are noted in Table 9. The percentiles are based on the distribution of all students taking the Regents Examination in Geometry (Common Core). The percentile ranks are computed in the following manner:

- A student's assigned "State percentile rank" will be the cumulative percentage of students scoring at the immediate lower score plus half of the percentage of students obtaining the given score.
- Students who obtain the highest possible score will receive a percentile rank of 99.
- Students who obtain the lowest possible score (0) will not receive a percentile rank.

Table 9 State Percentile Ranking for Raw Score – Regents Examination in Geometry (Common Core)

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

Chapter 5: Validity (Standard 1)

Restating the purpose and uses of the Regents Examination in Geometry (Common Core), this exam measures examinee achievement against New York State's learning standards. The exam is prepared by teacher examination committees and New York Department of Education subject 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 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 (Common Core) is intended for use in satisfying State testing requirements for students who have finished a course of instruction 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.p12.nysed.gov/ciai/gradreq/2015GradReq11-15.pdf. Results of the Regents Examination in Geometry (Common Core) may also be used to satisfy various locally established requirements throughout the State.

The validity of score interpretations for the Regents Examination in Geometry (Common Core) 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 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. 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 content standards associated with Geometry are based on the New York State Common Core Learning Standards for Mathematics and the PARCC Model Content Framework for Geometry. The content standards define what students should understand and be able to do at the high school level; the Model Content Framework describes which content is included and emphasized within the

Geometry course, specifically. More information about the relationship between the NYS CCLS and the PARCC Model Content Frameworks can be found in this memo.

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 (Common Core) is essentially the design document for constructing the exam. It provides explicit definition of the construct 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 10 displays the Conceptual category, domains, and target percent of each for the Regents Examination in Geometry (Common Core). Geometry is associated with the high school content standards within the *conceptual category* of Geometry. This conceptual category contains *domains* of related *clusters* of standards. This chart shows the high school mathematics domains included in Geometry, as well as the corresponding percent of credits on the Geometry Regents Exam.

Table 10 Test Blueprint, Regents Examination in Geometry (Common Core)

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

Item Development Process

Test development for the Regents Examination in Geometry (Common Core) 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 Department of Education 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 (Common Core) under strict guidelines that leverage best practices (see https://www.engageny.org/resource/regents-exams-ela). 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 multiple documents, including the blueprint noted in Table 10 and Item Writing Guidelines noted in Appendix A. To facilitate the alignment of items during development with standards, Standards Interpretations are also provided to developers. These interpretations are noted in Appendix B. Both multiple-choice and constructed-response items are included in the Regents Examination in Geometry (Common Core) to ensure appropriate coverage of the construct domain.

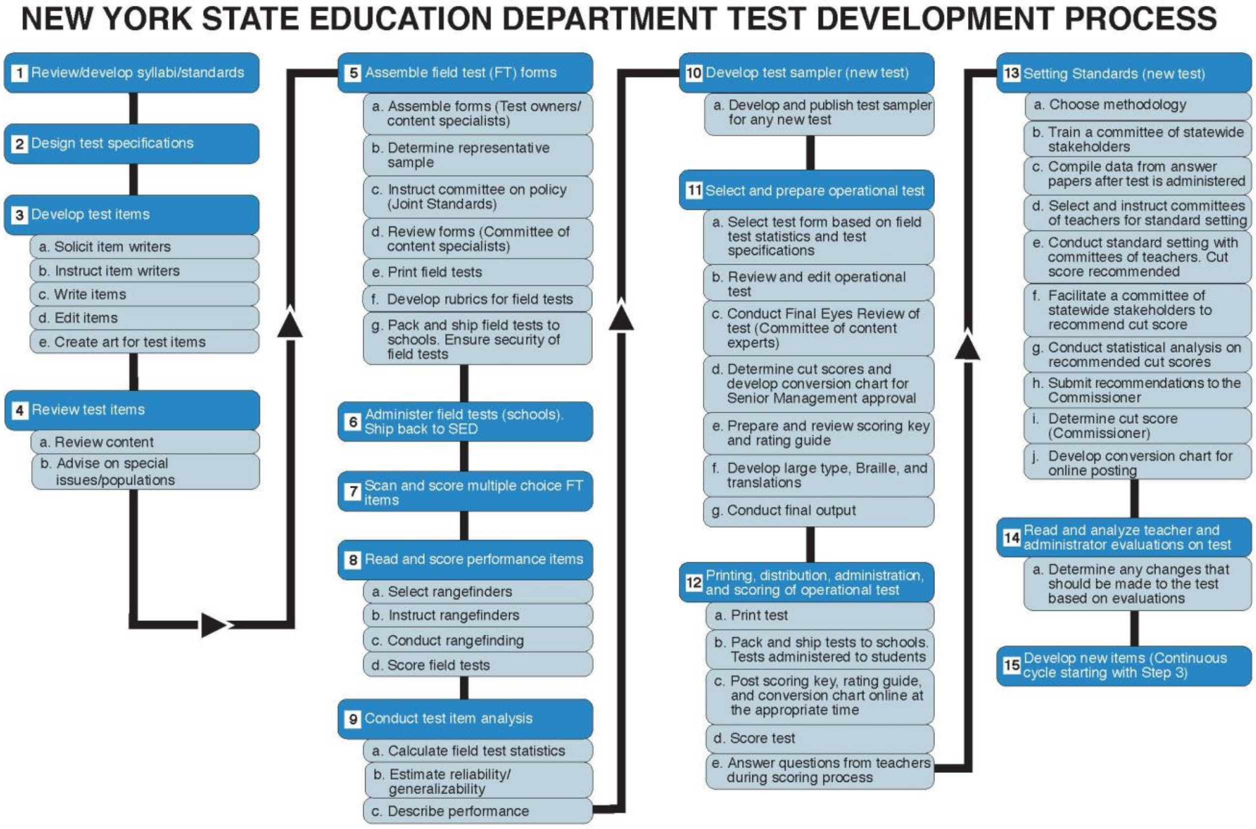

NEW YORK STATE EDUCATION DEPARTMENT TEST DEVELOPMENT PROCESS
1 Review/develop syllabi/standards
2 Design test specifications
3 Develop test items
a. Solicit item writers
b. Instruct item writers
c. Write items
d. Edit items
e. Create art for test items
4 Review test items
a. Review content
b. Advise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

Figure 7 New York State Education Department Test Development Process

Item Review Process

The item review process helps to ensure the consistent application of rigorous item reviews intended to assess the quality of the items developed and identify items that require edits or removal from the pool of items to be field tested. This process allows high quality items to be continually developed in a manner that is consistent with the test blueprint. Item review guidelines for multiple-choice items are included in Appendix C.

All reviewers participate in rigorous training designed to assist in a consistent interpretation of the standards throughout the item review process. This is a critical step in item development because consistency between the standards and what the items are asking examinees is a fundamental form of evidence of the validity of the intended score interpretations. Another integral component of this item review process is to review the scoring rules, or "rubrics," for their clarity and consistency in what the examinee is being asked to demonstrate by responding to each item. Each of these elements of the review process are 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 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 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. Likewise, 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 Department of Education (NYSED) is committed to ongoing standards review over time and periodically solicits thoughtful, specific responses from stakeholders about individual standards within the NY 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 consistent with the rubrics. Accordingly, it is important to control and monitor whether construct-irrelevant variance in response patterns has been introduced at any point in the test development, administration, or scoring processes.

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

Administration and Scoring

Adherence to standardized administration procedures is fundamental to the validity of test scores and their interpretation, as such procedures allow for adequate and consistently applied conditions for scoring the work of every student who takes the examination. For this reason, guidelines titled *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 Regents testing program. All secondary level Regents examinations are administered under these standard conditions to support valid inferences for all students. These standard procedures also cover testing students with disabilities that 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, Wang, 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 Conarro, 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 are 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. A consensus on a training for each score point of each item 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 set for the June 2015 administration of the Regents Examination in Geometry (Common Core) is located at http://www.nysedregents.org/geometrycc/615/geomcc62015-mrs.pdf.

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 occur 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 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 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 is also 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.

In support of the goal of valid score interpretations for each examinee, then, such scoring training procedures are implemented for the Regents Examination in Physical Setting/Earth Science. 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 the item developer intended them to. As documented in Table 2, the point biserial statistics for distractors in the multiple-choice items all appear to be very low (negative or very close to 0), indicating that examinees are not being drawn to an unintended construct. Infit statistics are provided in Table 5. Values for all 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 (Common Core):

- 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.5 Items for the 2014 Regents Examination in Geometry (Common Core) show a range of p-values consistent with the targeted exam difficulty. The p-values for the MC items ranged from 0.18 to 0.93, while proportion-correct values for the constructed response items (Table 3) were 0.25 and 0.72. The difficulty distribution illustrated in Figure 1 shows a wide range of item difficulties on the exam. This is consistent with general test development practice which seeks to measure student ability along a full range of difficulty. Refer to 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. Tables 2 and 3 provide point biserial values on the correct responses, and Table 2 also provides point biserial values on the three distractors for multiple choice items. The values for all items indicate that they are discriminating well between high- and low-performing examinees. Point biserials for all distractors are negative or very close to zero, indicating that examinees are responding to the items as expected during item and rubric development. Refer to 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 2012-2014. 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 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,” and “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. No items used on the New York Regents examination in Geometry (Common Core) displayed DIF during field test analyses.

Full differential item functioning results are reported in Appendix C of the 2012 technical report and in Appendix E of the 2013 and 2014 technical reports located at http://www.p12.nysed.gov/assessment/reports/2014/hselacc-tr14.pdf.

IRT Model Fit

Model fit for the Rasch method used to estimate location (difficulty) parameters for the items on the Regents Examination in Geometry (Common Core) provide important evidence that the internal structure of the test is of high technical quality. The Infit values for all items fall within a targeted range of [0.7, 1.3]. The mean infit value is 1.00. A finding of 36 out of 36 items falling in the ideal fit

[5] Refer to the field test report for details: http://www.p12.nysed.gov/assessment/reports.

range indicates that the Rasch model fits the Regents Examination in Geometry (Common Core) item data well. Refer to section 3 of this report for additional details.

Test Reliability

As discussed, test reliability is a measure of the internal consistency of a test (Cronbach, 1951). It is a measure of the extent to which the items on a test provide consistent information about student mastery of the domain. Reliability should ultimately demonstrate that examinee score estimates maximize consistency and therefore minimize error or, theoretically speaking, that examinees who take a test multiple times would get the same score each time. Assessments that include items with higher maximum possible score points may show slightly lower reliabilities than assessments with dichotomous and low maximum possible scores points. The Regents Examination in Geometry (Common Core) contains two constructed response items with maximum possible points of 4 and 6. The reliability estimate for the Regents Examination in Geometry (Common Core) is .78. Refer to section 4 of this report for additional details related to evaluating the standard errors of measurement, and the consistency and accuracy of examinee scores.

Classification Consistency and Accuracy

A decision consistency analysis measures the agreement between the classifications based on two non-overlapping, equally difficult forms of the test. If two parallel forms of the test were given to the same students, the consistency of the measure would be reflected by the extent that the classification decisions made from 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 provides strong evidence that the internal structure of a test is sound.

The results for the dichotomies created by the four cut scores, are presented in Table 7. The tabled values are derived with the program *BB-Class* (Brennan, 2004) using the Livingston and Lewis method. The overall decision consistency ranged from 0.87 to 0.94, and the decision accuracy ranged from 0.91 to 0.96. Both decision consistency and accuracy values indicate good consistency and accuracy of examinee classifications.

Dimensionality

In addition to model fit, a strong assumption of the Rasch model is that the construct measured by a test is unidimensional. Violation of this assumption might suggest that the test is measuring something other than the intended content and indicate that the quality of the test structure is compromised. A principal components analysis was conducted to test the assumption of unidimensionality, and the results provide strong evidence that a single dimension in the Regents Examination in Geometry (Common Core) 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 analyses on the internal structure of the Regents Examination in Geometry (Common Core), strong evidence exists that the exam is functioning as intended and is providing reasonably valid and reliable information about examinee performance.

5.4 Evidence Based on Relations to Other Variables

Another source of validity evidence is based on the relation of the test to other variables. This source commonly encompasses two validity categories prevalent in the literature and practice—concurrent and predictive validity. To make claims about the validity of a test that is to be used for high stakes purposes, such as the Regents Examination in Geometry (Common Core), these claims could be supported by providing evidence that performance on the Geometry test correlates well with other tests that measure the same or similar constructs. Although not absolute in its ability to offer evidence that concurrent test score validity exists, such correlations can be helpful for supporting a claim of concurrent validity if the correlation is high. To conduct such studies, matched examinee score data for other tests measuring the same content as the Regents Examination in Geometry (Common Core) is ideal, but the systematic acquisition of such data is complex and costly.

Importantly, a strong connection between classroom curriculum and test content may be inferred by the fact that New York State educators, deeply familiar with both the curriculum standards and their enactment in the classroom, develop all content for the Regents Examination in Geometry (Common Core).

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 Common Core Learning Standards is to prepare students for college and career, it will be important to gather evidence of this empirical relationship over time.

Currently, the predictive validity is supported by expert judgments gathered during the standard-setting process for Regents Examination in Geometry (Common Core). During this process, subject matter experts described the performance of examinees across five levels and made recommendations on the cut scores to be used in distinguishing such performance. The process reflected best psychometric practice as articulated in the Standards for Educational and Psychological Measurement (AERA et al., 2014) and proceeded according to the plans reviewed by the New York State Technical Advisory Committee as well as independent national expert. This effort inherently represents further expert review of the test content and its alignment with the objectives of the CCLS. Participating subject matter experts made explicit judgments about what each item was asking of examinees and what successful performance on the items means for the progress toward college and career readiness as defined by the standards.

After careful consideration of the nature of the new examinations including their goal of providing evidence to support readiness claims, the rigor of the new curricula, the transitional and aspirational aspects of the State policy directives, and the role of the assessment in student learning throughout high school and beyond, the standard setting committees made recommendations on the cut scores to the Commissioner of Education. The Commissioner accepted the recommendations of the standard setting panelists. More information is available in the Standard Setting technical report at http://www.p12.nysed.gov/assessment/reports/commoncore.

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, however, 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 for 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 the competencies demonstrated in the Regents Examination in Geometry (Common Core) 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 (Common Core) 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 (Common Core) 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.

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

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.

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*. McGraw-Hill Education CTB Innovative Research and Development Grant. Monterey: McGraw-Hill Education CTB.

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.

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.

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.

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 – Item Writing Guidelines

Guidelines for Writing Multiple-Choice Math Items

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

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

3. Include problems that come from a real-world context or problems that make use of multiple representations.
When using real-world problems, use real-world formulas and equations *(e.g., the kinetic energy,* k*, of an object with mass,* m*, and velocity,* v*, is* $k = ½ mv^2$*).* Use real-world statistics whenever possible.

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

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

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

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

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

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

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

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

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

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

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

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

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

Guidelines for Writing Constructed-Response Math Items

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

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

3. Include problems that come from a real-world context or problems that make use of multiple representations.
When using real-world problems, use real-world formulas and equations *(e.g., the kinetic energy,* k*, of an object with mass,* m*, and velocity,* v *is* $k = ½ mv^2$*).* Use real-world statistics whenever possible.

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

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

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

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

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

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

10. The item should be written to require a specific form of answer.
Phrases like “in terms of π,” “to the nearest tenth,” and “in simplest radical form” may simplify the writing of the rubric for these types of items.

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

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

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

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

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

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