New York State Regents Examination in Integrated Algebra

June 2009 Administration

Technical Report on Reliability and Validity

Technical Report
Prepared for the New York State Education Department
by Pearson

Copyright

Table of Contents

List of Tables and Figures

List of Tables and Figures, Continued

Introduction

In March 2005, the Board of Regents adopted a new Learning Standard for Mathematics and issued a revised Mathematics Core Curriculum, resulting in the need for the development and phasing in of three new mathematics Regents examinations: Integrated Algebra, Geometry, and Algebra 2/Trigonometry. These new Regents examinations in mathematics will replace the Regents Examinations in Mathematics A and Mathematics B. To fulfill the mathematics Regents examination requirement for graduation, students must pass any one of these new commencement-level Regents examinations. The first administration of the Regents Examination in Integrated Algebra took place in June 2008. The first administration of the Regents Examination in Geometry took place in June 2009. The first administration of the Regents Examination in Algebra 2/Trigonometry will take place in June 2010.

Integrated Algebra is based on the content contained in the Mathematics Core Curriculum (Revised 2005). The first administration took place in June 2008 and the new standards were set. The same standards have been maintained through the use of equating for the subsequent administrations: August 2008, January 2009, and June 2009. In June 2009, a score collection effort was conducted, where a representative sample of students identified across the New York State and their answer sheets were sent back to Pearson for processing. Through the data collected, further reliability and validity evidence can be examined. This technical document provides such details based on the data collected from the June 2009 administration of the Regents Examination in Integrated Algebra.

First, discussions on reliability are presented, including classical test theory based reliability evidence, the Item Response Theory (IRT) based reliability evidence, evidence related to subpopulations, and reliability evidence on classification accuracy for three achievement levels. Next, validity evidence is described, including evidence in internal structure validity, content validity, and construct validity. Equating, scaling, and scoring approaches used for the Regents Examination in Integrated Algebra are then described. Contrasts between the pre-equating and the post-equating analyses are presented. Finally, scale score distributions for the entire state and for subpopulations are presented.

The analysis was based on data collected after the June 2009 administration. This technical report includes reliability and validity evidence for the tests, as well as summary statistics for the administration. The table below describes the distribution of public schools (Needs/Resource Capacity (N/RC) categories) and nonpublic schools. Based on the distribution, the sample resembles the characteristics of the population data collected from the June 2008 administration and can be considered representative. All the analysis in this report, therefore, is based on this representative data.

Table 1. Distribution of Needs/Resource Capacity (N/RC) Categories

Need/Resource Capacity Index	Number of Schools	Number of Students	Percent
New York City	39	4,012	26.96
Large Cities	6	831	5.59
Urban-Suburban High Need/ Resource Capacity Index	8	1,009	6.78
Rural	16	1,011	6.79
Average Need/Resource Capacity Index Districts	46	4,567	30.69
Low Need/Resource Capacity Index Districts	16	1,987	13.35
Charter Schools	2	114	0.77
Non-Public Schools	21	1,348	9.06
Total	154	14,879	

Table 2. Test Configuration by Item Type

Item Type	Number of Items	Number of Credits	Percent of Credits
Multiple-Choice	30	60	68.96
Constructed-Response	9	27	31.03
Total	39	87	

Table 3. Test Blueprint by Content Strand

Content Strands	Number of Items	Number of Credits	2009 Percent of Credits	Target Percent of Credits
Number Sense and Operations (1)	4	8	9.20	6–10%
Algebra (2)	21	45	51.72	50–55%
Geometry (3)	5	13	14.94	14–19%
Measurement (4)	3	6	6.90	3–8%
Statistics and Probability (5)	6	15	17.24	14–19%

Table 4. Test Map by Standard and Content Strand

Test Part	Item Number	Item Type	Maximum Credit	Content Strand
I	1	Multiple-Choice	2	Measurement
I	2	Multiple-Choice	2	Algebra
I	3	Multiple-Choice	2	Algebra
I	4	Multiple-Choice	2	Algebra
I	5	Multiple-Choice	2	Statistics and Probability
I	6	Multiple-Choice	2	Algebra
I	7	Multiple-Choice	2	Algebra
I	8	Multiple-Choice	2	Statistics and Probability
I	9	Multiple-Choice	2	Algebra
I	10	Multiple-Choice	2	Number Sense and Operations
I	11	Multiple-Choice	2	Measurement
I	12	Multiple-Choice	2	Algebra
I	13	Multiple-Choice	2	Algebra
I	14	Multiple-Choice	2	Algebra
I	15	Multiple-Choice	2	Statistics and Probability
I	16	Multiple-Choice	2	Algebra
I	17	Multiple-Choice	2	Algebra
I	18	Multiple-Choice	2	Algebra
I	19	Multiple-Choice	2	Geometry
I	20	Multiple-Choice	2	Geometry
I	21	Multiple-Choice	2	Algebra
I	22	Multiple-Choice	2	Algebra
I	23	Multiple-Choice	2	Algebra
I	24	Multiple-Choice	2	Geometry
I	25	Multiple-Choice	2	Algebra
I	26	Multiple-Choice	2	Number Sense and Operations
I	27	Multiple-Choice	2	Number Sense and Operations
I	28	Multiple-Choice	2	Measurement
I	29	Multiple-Choice	2	Algebra
I	30	Multiple-Choice	2	Algebra
II	31	Constructed-Response	2	Number Sense and Operations
II	32	Constructed-Response	2	Algebra
II	33	Constructed-Response	2	Statistics and Probability
III	34	Constructed-Response	3	Geometry
III	35	Constructed-Response	3	Algebra
III	36	Constructed-Response	3	Statistics and Probability
IV	37	Constructed-Response	4	Statistics and Probability
IV	38	Constructed-Response	4	Algebra
IV	39	Constructed-Response	4	Geometry

The scale scores range from 0 to 100 for all Regents examinations. The three achievement levels on the exams are Level 1 with a scale score from 0 to 64, Level 2 with a scale score from 65 to 84, and Level 3 with a scale score from 85 to 100.

The Regents examinations typically consist of some number of multiple-choice (MC) items, some number of constructed-response (CR) items, and sometimes essay questions. Table 2 shows how many MC and CR items were on the Regents Examination in Integrated Algebra, as well as the number and percentage of credits for both item types. Table 3 reports item information by content strand. All items on the Regents Examination in Integrated Algebra were classified based on the mathematical standard.

Each form of the examination must adhere to strict rules indicating how many items per standard and content strand should be placed on a single form. In this way, the examinations can claim to measure the same concepts and standards from administration to administration, as long as the standards remain constant. Table 4 provides a test map by standard and content strand, indicating the required number of items associated with each standard and content strand.

There are 30 MC items, each worth 2 credits, and nine CR items, worth from 2 to 4 credits each. Table 5 presents a summary of raw score means for the total number of MC items, the total number of CR items, and all questions combined. The standard deviation is also reported.

Table 5. Raw Score Mean and Standard Deviation Summary

Item Type	Raw Score Mean	Standard Deviation
Multiple-Choice	34.83	13.40
Constructed-Response	12.41	7.36
Total	47.24	20.01

Table 6 reports the empirical statistics per item. The table includes item position on the test, item type, maximum item score value, content strand, the number of students included in the data who responded to the item, point biserial, and weighted item mean.

Table 6. Empirical Statistics for the Regents Examination in Integrated Algebra, June 2009 Administration

Item Position	Item Type	Max. Item Score	Content Strand	Number of Students	Point Biserial	Item Mean	Weighted Item Mean
1	Multiple-Choice	2	4	14,861	0.27	1.77	0.89
2	Multiple-Choice	2	2	14,839	0.47	1.19	0.60
3	Multiple-Choice	2	2	14,859	0.43	1.64	0.82
4	Multiple-Choice	2	2	14,863	0.52	1.43	0.72
5	Multiple-Choice	2	5	14,836	0.43	1.22	0.61
6	Multiple-Choice	2	2	14,865	0.45	1.50	0.75
7	Multiple-Choice	2	2	14,853	0.44	1.53	0.77
8	Multiple-Choice	2	5	14,850	0.32	1.52	0.76
9	Multiple-Choice	2	2	14,854	0.57	1.40	0.70
10	Multiple-Choice	2	1	14,858	0.51	1.21	0.61
11	Multiple-Choice	2	4	14,857	0.48	1.17	0.59
12	Multiple-Choice	2	2	14,860	0.50	1.19	0.60
13	Multiple-Choice	2	2	14,813	0.25	1.00	0.50
14	Multiple-Choice	2	2	14,851	0.39	1.21	0.61
15	Multiple-Choice	2	5	14,853	0.36	1.14	0.57
16	Multiple-Choice	2	2	14,853	0.58	1.13	0.57
17	Multiple-Choice	2	2	14,829	0.44	0.97	0.49
18	Multiple-Choice	2	2	14,843	0.46	1.40	0.70
19	Multiple-Choice	2	3	14,856	0.23	0.95	0.48
20	Multiple-Choice	2	3	14,844	0.41	0.93	0.47
21	Multiple-Choice	2	2	14,858	0.43	0.99	0.50
22	Multiple-Choice	2	2	14,850	0.48	1.26	0.63
23	Multiple-Choice	2	2	14,855	0.47	0.92	0.46
24	Multiple-Choice	2	3	14,856	0.40	1.47	0.74
25	Multiple-Choice	2	2	14,828	0.45	0.96	0.48
26	Multiple-Choice	2	1	14,859	0.38	0.90	0.45
27	Multiple-Choice	2	1	14,857	0.56	0.92	0.46
28	Multiple-Choice	2	4	14,805	0.28	0.85	0.43
29	Multiple-Choice	2	2	14,853	0.37	0.48	0.24
30	Multiple-Choice	2	2	14,850	0.43	0.65	0.33
31	Constructed-response	2	1	14,879	0.62	0.95	0.48
32	Constructed-response	2	2	14,879	0.67	0.82	0.41
33	Constructed-response	2	5	14,879	0.41	1.60	0.80
34	Constructed-response	3	3	14,879	0.69	1.10	0.37
35	Constructed-response	3	2	14,879	0.63	0.74	0.25
36	Constructed-response	3	5	14,879	0.61	1.30	0.43
37	Constructed-response	4	5	14,879	0.72	1.52	0.38
38	Constructed-response	4	2	14,879	0.55	2.65	0.66
39	Constructed-response	4	3	14,879	0.70	1.72	0.43

The item mean score is a measure of the item difficulty, ranging from 0 to the item maximum score. The higher the item mean score relative to the maximum score attainable, the easier the item is. The following formula is used to calculate this index for both MC and CR items:

$$M_i = c_i / n_i ,$$

where

M_i = the mean score for item i,

c_i = the total credits students obtained on item i, and

n_i = the maximum credits students could have obtained on item i.

The weighted item mean score is the item mean score divided by the max item score, ranging from 0 to 1. The point biserial coefficient is a measure of the relationship between a student's performance on the given item (correct or incorrect for MC items and raw score points for CR items) and the student's score on the overall test. Conceptually, if an item has a high point biserial (i.e., 0.30 or above), it indicates that students who performed well on the test also performed relatively well on the given item and students who performed poorly on the test also performed relatively poorly on the given item. If the point biserial value is high, it is typically stated that the item did a good job discriminating between high performing and low performing students. Assuming the total test score represents the extent to which a student possesses the construct being measured by the test, high item total correlations indicate the items on the test require this construct to be answered correctly if it is an MC item, or a relatively high score out of the maximum credits possible if it is a CR item. The point biserial correlation coefficient was computed between the item score and the total score on the test with the target item score excluded (also called corrected point biserial correlation coefficient).

The possible range of the point biserial coefficient is -1.0 to 1.0. In general, relatively high point biserials are desirable. A negative point biserial suggests that overall the most proficient students are getting the item wrong (if it is an MC item) or scoring low on the item (if it is a CR item) and the least proficient students are getting the item correct or scoring high. Any item with a point biserial that is near zero or negative should be carefully reviewed.

On the basis of the values reported in Table 6, the item means ranged from 0.48 to 2.65, while the maximum credits ranged from 2 to 4 for the 39 items on the test. The point biserial correlations were reasonably high, ranging from 0.23 to 0.72, suggesting good discriminative power on the items to differentiate students who scored high on the test from students who scored low on the test. Altogether, there were 4 items that had point biserial correlations lower than 0.30. In Appendix A, Tables A1 and A2 report the percentage of students at each of the possible score points for all items.

Reliability

Internal Consistency

Reliability is the consistency of the results obtained from a measurement. The focus of reliability should be on the results obtained from a measurement and the extent to which they remain consistent over time or among items or subtests that constitute the test. The ability to consistently measure students' performance is a necessary prerequisite to making appropriate score interpretations.

As stated above, test score reliability refers to the consistency of the results of a measurement. This consistency can be seen in the degree of agreement between two measures on two occasions, or it can be viewed as the degree of agreement between the components and the overall measurement. Operationally, such comparisons are the essence of mathematically defined reliability indices.

All measures consist of an accurate, or true, score component and an inaccurate, or error, score component. Errors occur as a natural part of the measurement process and can never be entirely eliminated. For example, uncontrollable factors, such as differences in the physical world and changes in examinee disposition, may work to increase error and decrease reliability. This is the fundamental premise of classical reliability analysis and classical measurement theory. Stated explicitly, this relationship can be represented with the following equation:

$$Observed\ Score = True\ Score + Error\ Score\ .$$

To facilitate a mathematical definition of reliability, these components can be rearranged to form the following ratio:

$$Reliability = \frac{\sigma^2_{True\ Score}}{\sigma^2_{Observed\ Score}} = \frac{\sigma^2_{True\ Score}}{\sigma^2_{True\ Score} + \sigma^2_{Error\ Score}}\ .$$

When there is no error, the reliability is the true score variance divided by true score variance, which is unity. However, as more error influences the measure, the error component in the denominator of the ratio increases. As a result, the reliability decreases.

Coefficient alpha (Cronbach, 1951), one of these internal consistency reliability indices, is provided for the entire test, for MC items only, for CR items only, for each of the content strands on the test, and for gender and ethnicity groups. Coefficient alpha is a more general version of the common Kuder-Richardson reliability coefficient and can accommodate both dichotomous and polytomous items. The formula for coefficient alpha is

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum\left(SD_i\right)^2}{\left(SD_x\right)^2}\right),$$

where

k = the number of items,

SD_i = the standard deviation of the set of scores associated with item i, and

SD_x = the standard deviation of the set of total scores.

Table 7. Reliability Estimates for Total Test, MC Items Only, CR Items Only and by Content Strands

	Number of Items	Raw Score Mean	Standard Deviation	Reliability[1]
Total Test	39	47.24	20.01	0.93
MC Items Only	30	34.83	13.40	0.88
CR Items Only	9	12.41	7.36	0.87
Number Sense and Operations	4	3.98	2.64	0.61
Algebra	21	25.02	10.75	0.87
Geometry	5	6.16	3.68	0.64
Measurement	3	3.79	1.70	0.28
Statistics and Probability	6	8.30	3.83	0.65

Table 7 reports reliability estimates for the entire test, for MC items only, for CR items only, and by content strands measured by the test. Notably, the reliability estimate is a statistic, and, like all other statistics, it is affected by the number of items, or test length. When the reliability estimate is calculated for content strands, sometimes there can be as few as three items within a given content strand, and so it is unlikely the alpha coefficient will be high. On the basis of the Spearman-Brown formula (Feldt & Brennan, 1988), with all other things being equal, the longer the test or the greater the number of items, the higher the reliability coefficient estimate is likely to be. Intuitively, the more items the students are tested on, the more information can be collected and the more reliable the achievement measure tends to be. The reliability coefficient estimates for the entire test, MC items only, and CR items only were all

[1] When the number of items is small, the calculated reliability tends to be low, because as a statistic, reliability is sample size sensitive.

reasonably high. Because the number of items per content strand tends to be small, the reliability coefficient for content strands tended not to be as high. This was especially true for the measurement content strand, which featured only three items.

Standard Error of Measurement

The standard error of measurement (SEM) uses the information from the test along with an estimate of reliability to make statements about the degree to which error is influencing individual scores. The SEM is based on the premise that underlying traits, such as academic achievement, cannot be measured exactly without a precise measuring instrument. The standard error expresses unreliability in terms of the reported score metric. The two kinds of standard errors of measurement are the SEM for the overall test and the SEM for individual scores. The second kind of SEM is sometimes called Conditional Standard Error of Measurement (CSEM). Through the use of CSEM, an error band can be placed around an individual score, indicating the degree to which error might have an impact on that score. The total test SEM is calculated using the following formula:

$$SEM = \sigma_x \sqrt{1 - \rho_{XX'}},$$

where

σ_x = the standard deviation of the total test (standard deviation of the raw scores), and

$\rho_{xx'}$ = the reliability estimate of the total test scores.

Through the use of an Item Response Theory (IRT) model, CSEM can be computed with the information function. The information function for the number correct score x is defined as

$$I(\theta, x) = \frac{(\sum_i P_i')^2}{\sum_i P_i Q_i},$$

where

P_i = the probability of a correct response to the item *i*,

P_i' = the derivative of P_i, and

$Q_i = 1 - P_i$.

For CSEM, it is the inversion of the square root of the test information function for a given proficiency score:

$$SEM(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}.$$

When IRT is used to model item responses and test scores, there is usually some kind of transformation used to convert ability estimates (θ) to scale scores. Similarly, CSEMs are converted using the same transformation function to scale scores so that they are reported on the same metric and the test users can interpret test scores together with the associated amount of measurement error.

Table 8 reports reliability estimate and SEM (on raw score metric) for different testing populations: all examinees, ethnicity groups (white, Hispanic, and African American), gender groups (male and female), English Language Learners (ELL), ELL Using Accommodations (ELL/SUA), ELL Using Translated Editions, Students with Disabilities (SWD), and SWD Using Accommodations (SWD/SUA). The number of students for each group is also provided. As can be observed from the table, the reliability estimates for total group and subgroups were all reasonably high compared with the industry standards, ranging from 0.86 to 0.93.

Table 8. Reliability Estimates and SEM for Total Population and Subpopulations

	Number of Students	Raw Score Mean	Standard Deviation	Reliability	SEM
All Students	14,879	47.24	20.01	0.93	5.45
White	7,249	53.71	17.83	0.91	5.37
Hispanic	2,196	39.01	17.25	0.90	5.55
African American	2,036	33.73	16.73	0.89	5.45
Male	6,924	46.92	19.90	0.92	5.47
Female	7,604	48.28	19.94	0.93	5.41
ELL	419	32.70	16.86	0.89	5.54
ELL/SUA	318	31.99	16.46	0.89	5.53
ELL/Translated Editions	160	33.85	14.87	0.86	5.60
SWD	1,366	30.65	15.35	0.87	5.48
SWD/SUA	1,277	30.60	15.52	0.88	5.47

Table 9 reports the raw scores, scale scores, Rasch proficiency estimates (θ), and corresponding CSEMs.

Table 9. Raw-to-Scale-Score Conversion Table and Conditional SEM for the Regents Examination in Integrated Algebra

Raw Score	Scale Score	Theta	CSEM	Raw Score	Scale Score	Theta	CSEM	Raw Score	Scale Score	Theta	CSEM
0	0	-5.503	1.834	30	65	-0.618	0.244	60	82	0.908	0.221
1	4	-4.780	1.015	31	66	-0.559	0.242	61	83	0.957	0.222
2	7	-4.057	0.728	32	67	-0.501	0.240	62	83	1.007	0.224
3	11	-3.622	0.602	33	68	-0.444	0.238	63	84	1.057	0.226
4	14	-3.306	0.528	34	69	-0.388	0.237	64	84	1.109	0.229
5	17	-3.055	0.477	35	70	-0.332	0.235	65	84	1.162	0.232
6	20	-2.845	0.441	36	71	-0.277	0.234	66	84	1.217	0.235
7	23	-2.663	0.412	37	71	-0.223	0.232	67	85	1.273	0.239
8	26	-2.503	0.390	38	72	-0.169	0.231	68	86	1.331	0.243
9	29	-2.358	0.371	39	73	-0.116	0.230	69	86	1.392	0.249
10	31	-2.227	0.355	40	74	-0.063	0.228	70	86	1.455	0.254
11	33	-2.105	0.342	41	74	-0.012	0.227	71	87	1.521	0.261
12	36	-1.992	0.330	42	75	0.040	0.226	72	87	1.592	0.269
13	38	-1.887	0.320	43	75	0.090	0.225	73	88	1.666	0.278
14	40	-1.787	0.311	44	76	0.141	0.224	74	88	1.746	0.288
15	42	-1.693	0.303	45	77	0.190	0.223	75	89	1.832	0.300
16	44	-1.603	0.296	46	77	0.240	0.221	76	89	1.926	0.313
17	46	-1.517	0.290	47	78	0.289	0.221	77	90	2.029	0.329
18	48	-1.435	0.284	48	78	0.337	0.220	78	91	2.144	0.347
19	49	-1.355	0.279	49	78	0.385	0.219	79	92	2.272	0.369
20	51	-1.279	0.274	50	79	0.433	0.218	80	92	2.417	0.395
21	53	-1.205	0.270	51	79	0.480	0.218	81	93	2.586	0.426
22	54	-1.133	0.266	52	80	0.528	0.217	82	94	2.784	0.466
23	56	-1.063	0.262	53	80	0.575	0.217	83	95	3.026	0.519
24	57	-0.996	0.259	54	80	0.622	0.217	84	96	3.335	0.596
25	59	-0.929	0.256	55	81	0.669	0.217	85	98	3.764	0.724
26	60	-0.865	0.253	56	81	0.716	0.217	86	99	4.482	1.013
27	61	-0.801	0.250	57	81	0.764	0.218	87	100	5.200	1.833
28	62	-0.739	0.248	58	82	0.811	0.219				
29	64	-0.678	0.246	59	82	0.859	0.220				

Classification Accuracy

Every test administration will result in some examinee classification error because of the limitations of educational measurement. Several elements used in test construction and in the roles for establishing cut scores can assist in minimizing these errors. However, it is still important to investigate reliability of classification.

The Rasch model was the IRT model used to carry out the item parameter estimation and examinee proficiency estimation for the Regents examinations. Some advantages of this IRT model include treating examinee proficiency as continuous, rather than discrete, and producing a 1-to-1 correspondence between raw score and proficiency estimate. When the Rasch model is applied to calibrate test data, a proficiency estimate will be assigned to a given examinee on the basis of the items the examinee got correct. The estimation of proficiency is also prone to error, which is the Conditional Standard Error of Measurement. Because of the CSEM, examinees whose proficiency estimates are near a cut score may be prone to misclassification. The classification reliability index calculated in the following section is a way to accommodate the measurement error and how that may affect examinee classification. This classification reliability index is based on the errors related to measurement limitations.

As can be observed in Table 9, the CSEMs tend to be relatively large at the two extremes of the distribution and relatively small in the middle. Because there are two cut scores associated with this 87 raw score point test, the cut scores are likely to be in the middle of the raw score distributions, as were cut scores for scale scores 65 and 85, where the CSEMs tend to be relatively small.

To calculate the classification reliability index under the Rasch model for a given ability score θ, the observed score $\hat{\theta}$ is expected to be normally distributed with a mean of θ and a standard deviation of $SE(\theta)$ (the SEM associated with the given θ). The expected proportion of examinees with true scores in any particular level is

$$\text{PropLevel}_k = \sum_{\theta=cut_{\theta_c}}^{cut_{\theta_d}} \left(\phi\left(\frac{cut_{\theta_b} - \theta}{SE(\theta)} \right) - \phi\left(\frac{cut_{\theta_a} - \theta}{SE(\theta)} \right) \right) \varphi\left(\frac{\theta - \mu}{\sigma} \right),$$

where cut_{θ_a} and cut_{θ_b} are Rasch scale points representing the score boundaries for levels of observed scores, cut_{θ_c} and cut_{θ_d} are the Rasch scale points representing score boundaries for levels of true scores, ϕ is the cumulative distribution function of the achievement level boundaries, and φ is the normal density function associated with the true scores (Rudner, 2005).

Because Rasch preserves the shape of the raw score distribution, which may not necessarily be normal, Pearson recommends that φ be replaced with the observed relative frequency distribution of θ. Some of the score boundaries may be unobserved. For example, the theoretical lower bound of Level 1 is $-\infty$. For practical purposes, boundaries with unobserved values can be substituted with reasonable theoretical values (–10.00 for lower bound of Level 1 and +10 for upper bound of Level 3).

To compute classification reliability, the proportions were computed for all the cells of a 3-by-3 classification table for the test, with the rows representing theoretical true percentages of examinees in each achievement Level and the columns representing the observed percentages.

		Observed		
True	1	2	3
	4	5	6
	7	8	9

For example, suppose the cut scores are 0.5 and 1.2 on the θ scale for the 3 levels. To compute the proportion in cell 4 (observed Level 1, with scores from –10 to 0.5; true Level 2, with scores from 0.5 to 1.2), the following formula will be used:

$$\text{PropLevel}_k = \sum_{\theta=.5}^{1.2}\left(\phi\left(\frac{0.5-\theta}{SE(\theta)}\right)-\phi\left(\frac{-10-\theta}{SE(\theta)}\right)\right)\varphi\left(\frac{\theta-\mu}{\sigma}\right).$$

Table 10 reports the percentages of students in each of the categories. The sum of the diagonal entries (cells 1, 5, and 9, shaded in the tables) represents the classification accuracy index for the test. The total percentage of students being classified accurately, on the basis of the model, was therefore 89.4%. At the proficiency cut (65), the false positive rate was 3.9% and the false negative rate was 1.6%, according to the model used.

Table 10. Classification Accuracy Table.[2]

Score Range	0–64	65–84	85–100	True
0–64	20.3%	3.9%	0.0%	24.1%
65–84	1.6%	50.7%	2.8%	55.1%
85–100	0.0%	2.0%	18.4%	20.4%
Observed	21.9%	56.6%	21.2%	99.7%

[2] Because of the calculation and the use of +10 and − 10 as cutoffs at the two extremes, the overall sum of true and observed was not always 100%, but it should be very close to 100%.

Validity

Validity is the process of collecting evidence to support inferences made with assessment results. In the case of the Regents examinations, the score use is applied to knowledge and understanding of the New York State content standards. Any correct use of the test scores is evidence of test validity.

Content and Curricular Validity

The Regents examinations are criterion-referenced assessments. That is, each assessment is based on an extensive definition of the content it assesses and its match to the content standards. Therefore, the Regents examinations are content-based and directly aligned to the statewide content standards. Consequently, the Regents examinations demonstrate good content validity. Content validity is a type of test validity that addresses whether the test adequately samples the relevant material it purports to cover.

Relation to Statewide Content Standards

The development of the Regents Examination in Integrated Algebra includes committees of educators from across the New York State, NYSED assessment and curriculum specialists and content developers from its test development contractor, Riverside. A sequential review process has been put in place by assessment and curriculum experts at NYSED and Riverside. Such an iterative process provides many opportunities for these assessment professionals to offer and implement suggestions for improving or eliminating items and to offer insights into the interpretation of the statewide content standards for the Regents Examination in Integrated Algebra. These review committees participate in this process to ensure the test content validity of the Regents examinations and the quality of the assessment.

In addition to providing information on the difficulty, appropriateness, and fairness of these items, committee members provide a needed check on the alignment between the items and the content standards they are intended to measure. When items are judged to be relevant—that is, representative of the content defined by the standards—this provides evidence to support the validity of inferences made (regarding knowledge of this content) with Regents examination results. When items are judged to be inappropriate for any reason, the committee can suggest either revisions (e.g., reclassification or rewording) or elimination of the item from the item pool. Items that are approved by the content review committee are later field-tested to allow for the collection of performance data. In essence, these committees review and verify the alignment of the test items with the objectives and measurement specifications to ensure that the items measure appropriate content. They also provide insights into the quality of the items, including making sure the items are well-written, ensuring the accuracy of answer keys, providing evaluation criteria for CR items, etc. The nature and

specificity of these review procedures provide strong evidence for the content validity of the Regents Examination in Integrated Algebra.

Educator Input

New York State educators provide valuable input on the content and the match between the items and the statewide content standards. In addition, many current and former New York State educators work as independent contractors to write items specifically to measure the objectives and specifications of the content standards for the Regents examinations. Using varied item writers provides a system of checks and balances for item development and review that reduces single-source bias. Because many people with different backgrounds write the items, it is less likely that items will suffer from a bias that might occur if items were written by a single author. This direct input from educators provides confirmation of the content validity of the Regents examinations.

Test Developer Input

The assessment experts at NYSED and their test development contractor, Riverside, provide a history of test-building experience, including content-related expertise. The input and review by these assessment professionals provide further support that the item is an accurate measure of the intended objective. As can be observed from Table 6, items are selected not only on the basis of their statistical properties, but also on the basis of their representation of the content standards. The same content specification and coverage are followed across all forms of the same assessment. These reviews and special efforts in test development offer additional evidence for the content validity of the Regents examinations.

Construct Validity

The term construct validity refers to the degree to which the test score is a measure of the characteristic (i.e., construct) of interest. A construct is an individual characteristic that is assumed to exist to explain some aspect of behavior (Linn & Gronlund, 1995). When a particular individual characteristic is inferred from an assessment result, a generalization or interpretation in terms of a construct is being made. For example, problem solving is a construct. An inference that students who master the mathematical reasoning portion of an assessment are good problem solvers implies an interpretation of the results of the assessment in terms of a construct. To make such an inference, it is important to demonstrate that this is a reasonable and valid use of the results.

The American Psychological Association provides the following list of possible sources for internal structure validity evidence (AERA, APA, & NCME, 1999):

- High intercorrelations among assessment items or tasks, attesting that the items are measuring the same trait, such as a content objective, sub-domain, or construct

- Substantial relationships between the assessment results and other measures of the same defined construct
- Little or no relationship between the assessment results and other measures that are clearly not those of the defined construct
- Substantial relationships between different methods of measurement regarding the same defined construct
- Relationships to non-assessment measures of the same defined construct

As previously mentioned, internal consistency also provides evidence of construct validity. The higher the internal consistency, or the reliability of the test scores, the more consistent the items are toward measuring a common underlying construct. In the previous chapter, it can be observed that the reliability estimates for the assessment were reasonably high, providing positive evidence for the construct validity of the assessment.

The collection of construct-related evidence is a continuous process. Five current metrics of construct validity for the Regents examinations are the item point biserial correlations, Rasch fit statistics, intercorrelation among content strands, principal component analysis of the underlying construct, and differential item functioning (DIF) check. Validity evidence in each of these metrics is described and presented below.

Item-Total Correlation

Item-total correlations provide a measure of the congruence between the way an item functions and our expectations. Typically, we expect students with relatively high ability (i.e., those who perform well on the Regents examinations overall) to answer items correctly, and students with relatively low ability (i.e., those who perform poorly on the Regents examinations overall) to answer items incorrectly. If these expectations are accurate, the point biserial (i.e., item-total) correlation between the item and the total test score will be high and positive, indicating that the item is a good discriminator between high-performing and low-performing students. A correlation value above 0.20 is considered acceptable; a correlation value above 0.30 is considered moderately good, and values closer to 1.00 indicate superb discrimination. A test consisting of maximally discriminating items will maximize internal consistency reliability. Correlation is a mathematical concept; therefore, it is not free from misinterpretation. Often, when an item is very easy or very difficult, the point biserial correlation will be artificially deflated. For example, an item with a p-value of 99 may have a correlation of only 0.05. This does not mean that this is a bad item. The low correlation can simply be a side effect of the item difficulty. Since the item is extremely easy for everyone, not just for high-scoring students, the item is not differentiating high-performing students from low-performing students; hence, it has low discriminating power. Because of these potential misinterpretations of the correlation, it is important to remember that the point biserial should not be used *alone* to determine the quality of an item.

Assuming that the total test score represents the extent to which a student possesses the construct being measured by the test, high point biserial correlations indicate that the tasks on the test require this construct to be answered correctly. Table 11 reports the point biserial correlation values for each of the items on the test. As can be observed from this table, all the items had point biserial values of at least 0.20. Overall, it seems that all the items on the test were performing well in terms of differentiating high-ability students from low-ability students and measuring toward a common underlying construct.

Rasch Fit Statistics

In addition to item point biserials, Rasch fit statistics also provide evidence of construct validity. The Rasch model is unidimensional. Therefore, statistics showing the model-to-data fit also provide evidence that each item is measuring the same unidimensional construct. The mean square fit (MNSQ) statistics are used to determine whether items are functioning in a way that is congruent with the assumptions of the Rasch mathematical model. Under these assumptions, how a student will respond to an item depends on the proficiency of the student and the difficulty of the item, both of which are on the same measurement scale. If an item is as difficult as a student is able, the student will have a 50% chance of getting the item correct. If a student is more able than an item is difficult, under the assumptions of the Rasch model, that student has a greater than 50% chance of correctly answering the item. On the other hand, if the item is more difficult than the student is able, he or she has a less than 50% chance of correctly responding to the item. Rasch fit statistics estimate the extent to which an item is functioning in this predicted manner. Items showing a poor fit with the Rasch model typically have values outside the range of –1.3 to 1.3.

Items may not fit the Rasch model for several reasons, all of which relate to students responding to items in unexpected ways. For example, if an item appears to be easy, but consistently solicits an incorrect response from high-scoring students, the fit value will likely be outside the range. Similarly, if a difficult item is answered correctly by many low-performing students, the fit statistics will not perform well. In most cases, the reason that students respond in unexpected ways to a particular item is unclear. However, it is occasionally possible to determine the cause of an item's misfit values by reexamining the item and its distracters. For example, if several high-performing students miss an easy item, reexamination of the item may show that it actually has more than one correct response. Two response types of MNSQ values are presented in Table 11, OUTFIT and INFIT. MNSQ OUTFIT values are sensitive to outlying observations. Consequently, OUTFIT values will be outside the range when students perform unexpectedly on items that are far from their ability level—for example, easy items for which high-performing students answer incorrectly and difficult items for which low-performing students answer correctly.

Table 11. Rasch Fit Statistics for All Items on Test

Item Position	Item Type	INFIT MNSQ	OUTFIT MNSQ	Point Biserial	Item Mean
1	Multiple-Choice	0.99	1.03	0.27	1.77
2	Multiple-Choice	0.95	0.91	0.47	1.19
3	Multiple-Choice	0.88	0.76	0.43	1.64
4	Multiple-Choice	0.84	0.73	0.52	1.43
5	Multiple-Choice	0.99	0.98	0.43	1.22
6	Multiple-Choice	0.90	0.81	0.45	1.50
7	Multiple-Choice	0.92	0.80	0.44	1.53
8	Multiple-Choice	1.05	1.18	0.32	1.52
9	Multiple-Choice	0.80	0.68	0.57	1.40
10	Multiple-Choice	0.90	0.83	0.51	1.21
11	Multiple-Choice	0.94	0.90	0.48	1.17
12	Multiple-Choice	0.92	0.86	0.50	1.19
13	Multiple-Choice	1.23	1.33	0.25	1.00
14	Multiple-Choice	1.03	1.04	0.39	1.21
15	Multiple-Choice	1.08	1.13	0.36	1.14
16	Multiple-Choice	0.83	0.76	0.58	1.13
17	Multiple-Choice	1.00	0.99	0.44	0.97
18	Multiple-Choice	0.93	0.84	0.46	1.40
19	Multiple-Choice	1.25	1.34	0.23	0.95
20	Multiple-Choice	1.04	1.06	0.41	0.93
21	Multiple-Choice	1.01	1.00	0.43	0.99
22	Multiple-Choice	0.94	0.86	0.48	1.26
23	Multiple-Choice	0.97	0.97	0.47	0.92
24	Multiple-Choice	1.00	0.89	0.40	1.47
25	Multiple-Choice	0.99	0.99	0.45	0.96
26	Multiple-Choice	1.08	1.11	0.38	0.90
27	Multiple-Choice	0.86	0.82	0.56	0.92
28	Multiple-Choice	1.18	1.30	0.28	0.85
29	Multiple-Choice	1.02	1.17	0.37	0.48
30	Multiple-Choice	0.97	1.07	0.43	0.65
31	Constructed-Response	1.03	1.05	0.62	0.95
32	Constructed-Response	0.90	0.86	0.67	0.82
33	Constructed-Response	1.16	1.61	0.41	1.60
34	Constructed-Response	1.02	1.04	0.69	1.10
35	Constructed-Response	1.08	0.98	0.63	0.74
36	Constructed-Response	1.01	1.00	0.61	1.30
37	Constructed-Response	1.13	1.11	0.72	1.52
38	Constructed-Response	1.43	1.72	0.55	2.65
39	Constructed-Response	1.16	1.22	0.70	1.72

MNSQ INFIT values are sensitive to behaviors that affect students' performance on items near their ability estimates. Therefore, high INFIT values would occur if a group of students of similar ability consistently responded incorrectly to an item at or around their estimated ability. For example, under the Rasch model, the probability of a student with an ability estimate of 1.00 responding correctly to an item with a difficulty of 1.00 is 50%. If several students at or around the 1.00 ability level consistently miss this item such that only 20% get the item correct, the fit statistics for these items are likely to be outside the typical range. Mis-keyed items or items that contain cues to the correct response (i.e., students get the item correct regardless of their ability) may elicit high INFIT values as well. In addition, tricky items, or items that may be interpreted to have double meaning, may elicit high INFIT values.

On the basis of the results reported in Table 11, items 13, 19, 28, 33 and 38 had relatively high INFIT/OUTFIT statistics. The fit statistics for the rest of the items were all reasonably good. It appears that the fit of the Rasch model was good for this test.

Correlation among Content Strands

There are five content strands within the core curriculum to which items are aligned on this examination. The number of items associated with the content strands ranged from three items to 21 items. Content judgment was made when classifying items into each of the content strands. To assess the extent to which all items aligned with the content strands are assessing the same underlying construct, a correlation matrix was computed. First, the total raw scores were computed for each content strand by summing up the items within the strand. Next, correlations were computed. Table 12 presents the results.

Table 12. Correlations among Content Strands

	Number Sense and Operations	Algebra	Geometry	Measurement	Statistics And Probability
Number Sense and Operations	1.00	0.73	0.64	0.45	0.65
Algebra		1.00	0.77	0.53	0.77
Geometry			1.00	0.45	0.69
Measurement				1.00	0.47
Statistics and Probability					1.00

As can be observed from Table 12, the correlations between the five content strands ranged from 0.45 (between number sense and operations and measurement) to 0.77 (between algebra and geometry, and between algebra and statistics and probability). This is another empirical piece of evidence suggesting that the content strands are measuring a common underlying construct.

Correlation among Item Types

Two types of items were used on the Regents Examination in Integrated Algebra: multiple-choice and constructed-response. Table 13 presents a correlation matrix (based on raw scores within each item type) to show the extent to which these two item types assessed the same underlying construct. As can be observed from the table, the correlations seem reasonably high. The high correlations between these two item types as well as between each item type and the total test is an indication of construct validity.

Table 13. Correlations among Item Types and Total Test

	Multiple-Choice	Constructed-Response	Total
Multiple-Choice	1.00	0.85	0.98
Constructed-Response		1.00	0.93
Total			1.00

Principal Component Analysis

As previously mentioned, the Rasch model (Partial Credit Model, or PCM, for CR items) was used to conduct calibration for the Regents examinations. The Rasch model is a unidimensional IRT model. Under this model, only one underlying construct is assumed to influence students' responses to items. To check whether only one dominant dimension exists in the assessment, exploratory principal component analysis was conducted on the students' item responses to further observe the underlying structure. Factor analysis was conducted on the item response matrix for different testing populations: all examinees, ethnicity groups (white, Hispanic, and African American), gender groups (male and female), ELL, ELL Using Accommodations (ELL/SUA), ELL Using Translated Editions, SWD, and SWD Using Accommodations (SWD/SUA). Only factors with eigenvalues greater than 1 were retained, a criteria proposed by Kaiser (1960). A scree plot was also developed (Cattell, 1966) to graphically display the relationship between factors with eigenvalues exceeding 1. Cattell suggests that when the scree plot appears to level off it is an indication that the number of significant factors has been reached. Table 14 reports the eigenvalues computed for each of the factors (only factors with eigenvalues exceeding 1 were kept and included in the table). Figure 1 shows the scree plot.

Table 14. Factors and Their Eigenvalues

	Eigenvalue													
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Factor 11	Factor 12	Factor 13	Total
All Students	10.65	1.40	1.21											13.26
White	9.15	1.40	1.23	1.02	1.00									13.81
Hispanic	8.49	1.59	1.17	1.15	1.10	1.08	1.03							15.61
African American	8.46	1.53	1.23	1.11	1.07	1.04	1.02							15.46
Male	10.48	1.47	1.17											13.13
Female	10.73	1.33	1.25											13.31
ELL	8.24	1.88	1.47	1.36	1.26	1.26	1.13	1.10	1.06	1.05				19.81
ELL/SUA	7.96	1.94	1.53	1.48	1.29	1.27	1.16	1.14	1.11	1.08	1.06			21.02
ELL/Translated Editions	6.99	2.19	1.72	1.64	1.58	1.49	1.44	1.39	1.25	1.17	1.17	1.10	1.01	24.14
SWD	7.46	1.48	1.45	1.16	1.12	1.09	1.07	1.04	1.03	1.01				17.90
SWD/SUA	7.62	1.47	1.43	1.18	1.12	1.09	1.08	1.04	1.04	1.01				18.07
	Proportion													
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Factor 11	Factor 12	Factor 13	
All Students	0.80	0.11	0.09											
White	0.66	0.10	0.09	0.07	0.07									
Hispanic	0.54	0.10	0.07	0.07	0.07	0.07	0.07							
African American	0.55	0.10	0.08	0.07	0.07	0.07	0.07							
Male	0.80	0.11	0.09											
Female	0.81	0.10	0.09											
ELL	0.42	0.09	0.07	0.07	0.06	0.06	0.06	0.06	0.05	0.05				
ELL/SUA	0.38	0.09	0.07	0.07	0.06	0.06	0.06	0.05	0.05	0.05	0.05			
ELL/Translated Editions	0.29	0.09	0.07	0.07	0.07	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.04	
SWD	0.42	0.08	0.08	0.06	0.06	0.06	0.06	0.06	0.06	0.06				
SWD/SUA	0.42	0.08	0.08	0.07	0.06	0.06	0.06	0.06	0.06	0.06				

In Table 14, there are up to 13 factors with eigenvalues exceeding 1. For all students, the dominant factor has an eigenvalue of 10.65, accounting for 80% of the variance among factors with loadings exceeding 1, whereas the other factors had eigenvalues around 1. After the first and second factors, the scree plot leveled off. The scree plot also demonstrates the large magnitude of the first factor, indicating that the items on the test are measuring toward one dominant common factor. This is another piece of empirical evidence that the test has 1 dominant underlying construct and the IRT unidimensionality assumption is met. Also, the single dominant factor for each student subgroup can be observed from Table 14. Note that for some subgroups, such as ELL groups, the sample size is much smaller compared to the rest of the groups of interest and therefore, the results may contain more error. But still, the dominancy of the first factor is apparent based on the results.

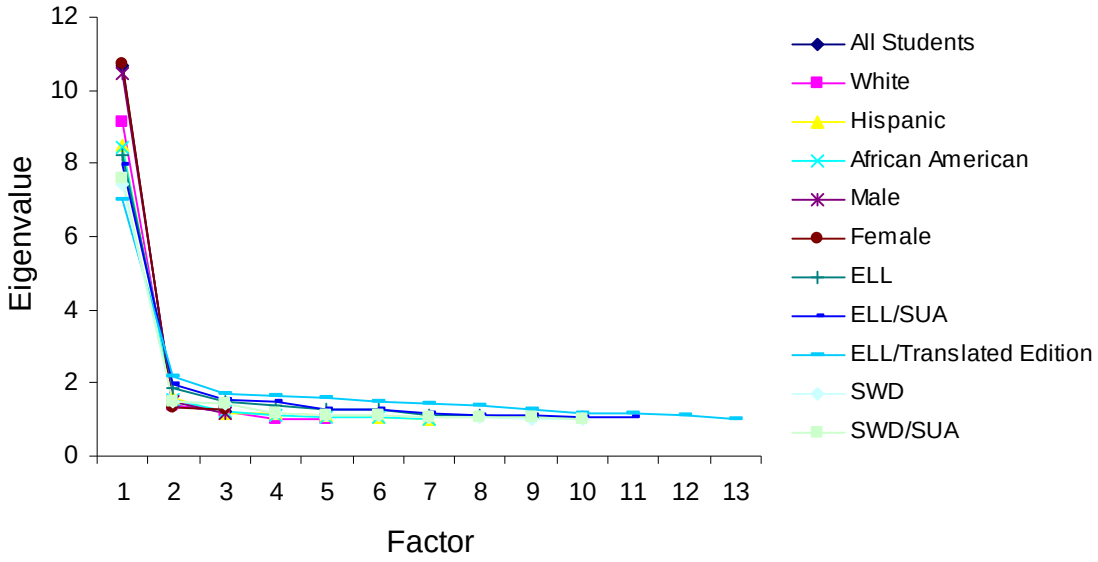

Figure 1. Scree Plot for Principal Component Analysis of Items on the Regents Examination in Integrated Algebra

Validity Evidence for Different Student Populations

The primary evidence for the validity of the Regents examinations lies in the content being measured. Since the test assesses the statewide content standards that are recommended to be taught to all students, the test is not more valid or less valid for use with one subpopulation of students relative to another. Because the Regents examinations measure what is recommended be taught to all students and are given under the same standardized conditions to all students, the tests have the same validity for all students. Moreover, great care has been taken to ensure that the items that make up the Regents examinations are fair and representative of the content domain expressed in the content standards.

Additionally, much scrutiny is applied to the items and their possible impact on minority or subpopulations in New York State. Every effort is made to eliminate items that may have ethnic or cultural biases. For example, content review and bias review are routinely conducted as part of the item review process, to eliminate any potential elements in the items that may unfairly advantage subpopulations of students.

Besides these content-based efforts that are routinely put forth in the test development process, statistical procedures are employed to observe whether, on the basis of data, there exists possibly unfair treatment of different populations. The differential item functioning (DIF) analysis was carried out on the data collected from the June 2009 administration. DIF statistics are used to identify items for which members of a focal group have a different probability of getting the items correct than members of a reference group, after the groups have been matched on ability level on the test. In the DIF analyses, the total raw score on the operational items is used as an ability-matching variable. Four comparisons were made for each item because the same DIF analyses are typically conducted for the other New York State assessments:

- males versus females
- white versus African American
- white versus Hispanic
- high need versus low need

For the MC items, the Mantel-Haenszel Delta (MHD) DIF statistics were computed (Dorans and Holland, 1992). The DIF null hypothesis for the Mantel-Haenszel method can be expressed as

$$H_0: \text{MH } \alpha=(P_{rm}/Q_{rm})/(P_{fm}/Q_{fm})=1,\ m=1,...,M\ ,$$

where P_{rm} refers to the proportion of students correctly answering the item in the reference group at proficiency level m and Q_{rm} refers to the proportion of students incorrectly answering the item in the reference group at proficiency level m. P_{fm} and Q_{fm} are defined similarly for the focal group. Holland and Thayer (1985) converted α into a difference in deltas via the following formula:

$$MHD=-2.35\ln(\text{MH } \alpha)\ ,$$

The following three categories were used to classify test items in three levels of DIF for each comparison: negligible DIF (A), moderate DIF (B), and large DIF (C). An item is flagged if it exhibits category B or C of DIF, using the following rules derived from the National Assessment of Educational Progress (NAEP) guidelines (Allen, Carlson, and Zalanak 1999):

Rules	Descriptions	Category
Rule 1	• MHD[3] not significant from 0 **or** • |MHD| < 1.0	A
Rule 2	• MHD is significantly different from 0 and {|MHD| ≥ 1.0 and < 1.5} **or** • MHD is not significantly different from 0 and |MHD| ≥ 1.0	B
Rule 3	• |MHD| ≥ 1.5 and is significantly different from 0	C

The effect size of the standardized mean difference (SMD) was used to flag the DIF for the CR items. The SMD reflects the size of the differences in performance on CR items between student groups matched on the total score. The following equation defines SMD:

$$SMD = \sum_k w_{Fk} m_{Fk} - \sum_k w_{Fk} m_{Rk} \ ,$$

where $w_{Fk} = n_{F+k} / n_{F++}$ is the proportion of focal group members who are at the k th stratification variable, $m_{Fk} = (1/n_{F+k})F_k$ is the mean item score for the focal group in the k th stratum, and $m_{Rk} = (1/n_{R+k})R_k$ is the analogous value for the reference group. In words, the SMD is the difference between the unweighted item mean of the focal group and the weighted item mean of the reference group. The weights applied to the reference group are applied so that the weighted number of reference group students is the same as the weighted number of focal group students (within the same ability group). The SMD is divided by the total group item standard deviation to get a measure of the effect size for the SMD using the following equation:

$$\text{Effect Size} = \frac{\text{SMD}}{SD} \ .$$

The SMD effect size allows each item to be placed into one of three categories: negligible DIF (AA), moderate DIF (BB), or large DIF (CC). The following rules are

[3] Note: The MHD is the ETS delta scale for item difficulty, where the natural logarithm of the common odds ratio is multiplied by –(4/1.7).

applied for the classification. Only categories BB and CC were flagged in the results.

Rules	Descriptions	Category
Rule 1	• If the probability is >0.05 or |Effect Size| is ≤ 0.17	AA
Rule 2	• If the probability is < 0.05 and if 0.17<|Effect Size|≤0.25	BB
Rule 3	• If the probability is <0.05 and if |Effect Size| is >0.25	CC

For MC and CR items, the favored group is indicated if an item was flagged. Tables 15–18 report DIF analysis for gender, ethnicity, and social economic status subpopulations. The sample sizes used for each of the subpopulations are reported in Table 8. When MHD values are positive, the focal group had a better odds ratio against the reference group; when the MHD values are negative, the reference group had a better odds ratio against the focal group. Similarly, when the SMD effect size values are positive, it is an indication that, at the same proficiency level, the focal group is performing better than the reference group on the item; when the SMD effect size values are negative, the reference group is performing better when the proficiency of the students is controlled.

Table 15 reports the DIF analysis for gender groups. The male group was treated as the reference group, and the female group was treated as the focal group. As can be observed from the table, one item was flagged for moderate DIF values (item 15) and one item was flagged for significant DIF values (item 11). Both favored the reference group, the male group.

Tables 16 and 17 report the DIF analyses for ethnicity groups. The white student group was treated as the reference group. In Table 16, the Hispanic student group was treated as the focal group and the DIF statistics reported; in Table 17, the African American student group was treated as the focal group and the DIF statistics reported. No item was flagged for the DIF in Tables 16 and 17.

Table 18 reports the DIF analysis for the high need category versus the low need category. N/RC based on the schools was used as the identification variable. The focal group is the low need group, with N/RC values being 5 and 6; the reference group is the high need group, with N/RC values being 1–4. The sample size for the high need group was 6,863 and for the low need group was 6,554. On the basis of the results presented in Table 18, one item (item 7) was flagged for moderate DIF favoring the high need group.

Table 15. DIF Statistics for the Regents Examination in Integrated Algebra, Focal Group: Female; Reference Group: Male

Item Position	Item Type	MH Delta	Effect Size	DIF Category	Favored Group
1	Multiple-Choice	-0.37	-0.02		
2	Multiple-Choice	0.39	0.03		
3	Multiple-Choice	0.43	0.03		
4	Multiple-Choice	-0.45	-0.03		
5	Multiple-Choice	0.33	0.03		
6	Multiple-Choice	-0.59	-0.04		
7	Multiple-Choice	0.17	0.01		
8	Multiple-Choice	0.22	0.02		
9	Multiple-Choice	0.63	0.04		
10	Multiple-Choice	0.30	0.02		
11	Multiple-Choice	-1.53	-0.12	C	Male
12	Multiple-Choice	-0.53	-0.04		
13	Multiple-Choice	0.17	0.02		
14	Multiple-Choice	-0.37	-0.03		
15	Multiple-Choice	-1.01	-0.09	B	Male
16	Multiple-Choice	-0.15	-0.01		
17	Multiple-Choice	-0.36	-0.03		
18	Multiple-Choice	-0.13	-0.01		
19	Multiple-Choice	-0.15	-0.02		
20	Multiple-Choice	0.32	0.03		
21	Multiple-Choice	0.32	0.03		
22	Multiple-Choice	-0.31	-0.02		
23	Multiple-Choice	0.28	0.02		
24	Multiple-Choice	0.30	0.02		
25	Multiple-Choice	0.06	0.01		
26	Multiple-Choice	-0.21	-0.02		
27	Multiple-Choice	-0.17	-0.01		
28	Multiple-Choice	0.07	0.01		
29	Multiple-Choice	-0.12	-0.01		

Table 15. DIF Statistics for the Regents Examination in Integrated Algebra, Focal Group: Female; Reference Group: Male, Continued

Item Position	Item Type	MH Delta	Effect Size	DIF Category	Favored Group
30	Multiple-Choice	-0.27	-0.02		
31	Constructed-Response	N/A	0.02		
32	Constructed-Response	N/A	0.07		
33	Constructed-Response	N/A	-0.02		
34	Constructed-Response	N/A	-0.01		
35	Constructed-Response	N/A	-0.06		
36	Constructed-Response	N/A	0.04		
37	Constructed-Response	N/A	0.05		
38	Constructed-Response	N/A	0.08		
39	Constructed-Response	N/A	-0.01		

Table 16. DIF Statistics for Regents Examination in Integrated Algebra, Focal Group: Hispanic; Reference Group: White

Item Position	Item Type	MH Delta	Effect Size	DIF Category	Favored Group
1	Multiple-Choice	0.48	0.04		
2	Multiple-Choice	0.38	0.04		
3	Multiple-Choice	-0.33	-0.04		
4	Multiple-Choice	0.58	0.05		
5	Multiple-Choice	-0.56	-0.05		
6	Multiple-Choice	-0.26	-0.02		
7	Multiple-Choice	0.55	0.04		
8	Multiple-Choice	-0.47	-0.05		
9	Multiple-Choice	0.47	0.04		
10	Multiple-Choice	0.58	0.05		
11	Multiple-Choice	0.46	0.04		
12	Multiple-Choice	-0.05	-0.01		
13	Multiple-Choice	0.10	0.01		
14	Multiple-Choice	0.13	0.02		
15	Multiple-Choice	-0.60	-0.05		
16	Multiple-Choice	0.01	0.00		
17	Multiple-Choice	-0.10	-0.01		
18	Multiple-Choice	-0.09	-0.00		
19	Multiple-Choice	0.18	0.02		
20	Multiple-Choice	0.17	0.01		
21	Multiple-Choice	0.06	0.01		
22	Multiple-Choice	0.74	0.06		
23	Multiple-Choice	-0.01	-0.00		
24	Multiple-Choice	-0.19	-0.01		
25	Multiple-Choice	0.52	0.04		
26	Multiple-Choice	-0.22	-0.02		
27	Multiple-Choice	0.58	0.04		
28	Multiple-Choice	0.13	0.01		
29	Multiple-Choice	0.34	0.03		

Table 16. DIF Statistics for Regents Examination in Integrated Algebra, Focal Group: Hispanic; Reference Group: White, Continued

Item Position	Item Type	MH Delta	Effect Size	DIF Category	Favored Group
30	Multiple-Choice	0.34	0.03		
31	Constructed-Response	N/A	-0.02		
32	Constructed-Response	N/A	-0.03		
33	Constructed-Response	N/A	-0.07		
34	Constructed-Response	N/A	-0.02		
35	Constructed-Response	N/A	-0.04		
36	Constructed-Response	N/A	-0.02		
37	Constructed-Response	N/A	-0.02		
38	Constructed-Response	N/A	-0.04		
39	Constructed-Response	N/A	-0.01		

Table 17. DIF Statistics for Regents Examination in Integrated Algebra, Focal Group: African American; Reference Group: White

Item Position	Item Type	MH Delta	Effect Size	DIF Category	Favored Group
1	Multiple-Choice	-0.02	-0.00		
2	Multiple-Choice	0.38	0.03		
3	Multiple-Choice	0.33	0.01		
4	Multiple-Choice	0.24	0.02		
5	Multiple-Choice	0.05	0.02		
6	Multiple-Choice	-0.17	-0.00		
7	Multiple-Choice	0.53	0.04		
8	Multiple-Choice	-0.27	-0.02		
9	Multiple-Choice	0.74	0.05		
10	Multiple-Choice	0.44	0.03		
11	Multiple-Choice	-0.22	-0.02		
12	Multiple-Choice	0.43	0.04		
13	Multiple-Choice	0.01	0.00		
14	Multiple-Choice	0.26	0.03		
15	Multiple-Choice	-0.46	-0.03		
16	Multiple-Choice	0.20	0.02		
17	Multiple-Choice	0.15	0.01		
18	Multiple-Choice	-0.15	-0.01		
19	Multiple-Choice	0.40	0.04		
20	Multiple-Choice	0.23	0.01		
21	Multiple-Choice	0.39	0.04		
22	Multiple-Choice	0.63	0.05		
23	Multiple-Choice	0.41	0.04		
24	Multiple-Choice	-0.01	0.00		
25	Multiple-Choice	0.54	0.04		
26	Multiple-Choice	-0.34	-0.03		
27	Multiple-Choice	0.32	0.02		
28	Multiple-Choice	0.44	0.05		
29	Multiple-Choice	0.26	0.01		

Table 17. DIF Statistics for Regents Examination in Integrated Algebra, Focal Group: African American; Reference Group: White, Continued

Item Position	Item Type	MH Delta	Effect Size	DIF Category	Favored Group
30	Multiple-Choice	0.46	0.03		
31	Constructed-Response	N/A	-0.05		
32	Constructed-Response	N/A	-0.02		
33	Constructed-Response	N/A	-0.12		
34	Constructed-Response	N/A	-0.04		
35	Constructed-Response	N/A	-0.04		
36	Constructed-Response	N/A	-0.03		
37	Constructed-Response	N/A	-0.02		
38	Constructed-Response	N/A	-0.08		
39	Constructed-Response	N/A	-0.05		

Table 18. DIF Statistics for Regents Examination in Integrated Algebra, Focal Group: High Need; Reference Group: Low Need

Item Position	Item Type	MH Delta	Effect Size	DIF Category	Favored Group
1	Multiple-Choice	0.07	-0.00		
2	Multiple-Choice	-0.04	0.01		
3	Multiple-Choice	-0.90	-0.09		
4	Multiple-Choice	0.82	0.05		
5	Multiple-Choice	-0.78	-0.05		
6	Multiple-Choice	0.06	0.01		
7	Multiple-Choice	1.22	0.09	B	High Need
8	Multiple-Choice	-0.15	-0.02		
9	Multiple-Choice	0.00	0.00		
10	Multiple-Choice	0.17	0.01		
11	Multiple-Choice	0.66	0.04		
12	Multiple-Choice	0.76	0.06		
13	Multiple-Choice	0.09	0.01		
14	Multiple-Choice	0.84	0.08		
15	Multiple-Choice	-0.65	-0.05		
16	Multiple-Choice	-0.01	-0.00		
17	Multiple-Choice	0.27	0.02		
18	Multiple-Choice	-0.09	0.00		
19	Multiple-Choice	0.46	0.05		
20	Multiple-Choice	0.05	0.00		
21	Multiple-Choice	-0.51	-0.04		
22	Multiple-Choice	0.34	0.03		
23	Multiple-Choice	0.18	0.01		
24	Multiple-Choice	-0.48	-0.03		
25	Multiple-Choice	0.64	0.05		
26	Multiple-Choice	0.20	0.01		
27	Multiple-Choice	0.77	0.04		
28	Multiple-Choice	0.01	0.01		
29	Multiple-Choice	0.61	0.03		

Table 18. DIF Statistics for Regents Examination in Integrated Algebra, Focal Group: High Need; Reference Group: Low Need, Continued

Item Position	Item Type	MH Delta	Effect Size	DIF Category	Favored Group
30	Multiple-Choice	0.53	0.04		
31	Constructed-Response	N/A	-0.01		
32	Constructed-Response	N/A	-0.06		
33	Constructed-Response	N/A	-0.02		
34	Constructed-Response	N/A	-0.03		
35	Constructed-Response	N/A	-0.00		
36	Constructed-Response	N/A	-0.03		
37	Constructed-Response	N/A	-0.06		
38	Constructed-Response	N/A	-0.03		
39	Constructed-Response	N/A	-0.06		

Equating, Scaling, and Scoring

To maintain the same performance standards across different administrations, the statistical procedure of *equating* is used with the Regents examinations so that the same scale scores, even though based on a different set of items, carry the same meaning over various administrations.

There are two main kinds of equating models: the pre-equating model and the post-equating model. For regular Regents examinations, NYSED uses the pre-equating model to construct test forms of similar difficulty.

Pre-equating results were available for the items that appeared on the June 2009 Regents Examination in Integrated Algebra. These items were field-tested in the spring of 2008, together with many other items in the item bank. In these stand-alone field test sessions, the number of students taking the field test forms ranged from 700 to 800. The field test forms typically contained 10–12 items, to lighten students' testing load in these sessions.

It has been speculated that the motivation of the students who participate in the field-testing may be lower than students who take the operational assessment, given the limited consequences of the field test and the lack of feedback (i.e., score reports) pertaining to their performance. Despite this possible lack of motivation, NYSED requirements regarding the availability of the raw score-to-scale score conversion chart of the recent administration dictates that a pre-equating model be employed for regular administrations of the Regents examinations. The rationale for these requirements is based primarily on the need to allow for the local scoring of the Regents examinations in the field and prompt knowledge of test results.

In this section, procedures employed in equating, scaling, and scoring for the Regents examinations are described. Furthermore, a contrast between the pre-equating results (based on the 2008 field testing and were the operational results) and the post-equating results for the operational items on the June 2009 administration of Regents Examination in Integrated Algebra is also presented.

Equating Procedures

Under the pre-equating model, the field test forms were equated by using two designs for the Regents examinations: Equivalent Groups and Common Item. A brief description of each method follows.

Equivalent Groups. For those field test forms without common items, it is assumed that the field test forms are administered to equivalent groups of students. This makes it possible to equate these forms using an equivalent group design. This is accomplished using the following steps:

Step 1: Calibrate all the field test forms allowing the item difficulties to center at a mean value of zero. This calibration produces three valuable components for the equating and scaling process. First, this produces item parameter estimates (item difficulties and step values) for MC and CR items. Second, this produces raw-score-to-theta tables for each form. Third, this produces a mean and standard deviation of the students who take the test form.

Step 2: Using the mean-ability estimate of one of the field test forms determines an equating constant for each of the other field test forms, which will produce a mean-ability estimate equal to that of the first form. Assuming that the samples of students who take each form are randomly equivalent, this will place the item parameters for the field test forms onto a common scale.

Step 3: Add the equating constant found in step 2 to the item difficulties and recalibrate each test form fixing the item parameters. This will provide a check to determine whether the equating constant actually produces student ability estimates that are equal to those found in the base field test form.

Step 4: Using the item parameter estimates from the field test forms, produce a raw-score-to-theta table for all complete forms. This will provide the tables needed to do the final scaling. Because the raw-score-to-theta tables for each form will be on the same scale, it will be possible to calculate the comparable scaled score for each raw score on the new tests.

Common Item Equating. For field test forms that contain common items, the equating is conducted in the following manner:

Step 1: Calibrate one form allowing the item difficulties to center at a mean value of 0, or use previously calibrated difficulty values if available. For the base test form, the calibration produces three valuable components for the equating and scaling process. First, this produces item parameter estimates (item difficulties and step values) for MC and CR items. Second, it produces raw-score-to-theta tables for each form. Third, this will produce a mean and standard deviation of the students who take the test form.

Step 2: Calibrate the other field test forms fixing the common item parameters to those found in step 1. This will place the item parameters for the mini-forms onto a common scale. (Before this step, an analysis of the stability of the item-difficulty estimates for the anchor items will be performed. Items demonstrating unstable estimates will not be used as anchors.)

Step 3: Repeat steps 2 and 3 for the other field test forms. This will place the item parameters for all the field test forms onto a common scale.

Step 4: Using the item parameter estimates from step 3, produce a raw-score-to-theta table for all complete forms. This will provide the tables needed to do the final scaling. Because the raw-score-to-theta tables will be on the same scale for each form, it will be possible to calculate the comparable scale score for each raw score on the new tests.

The stability of the anchor was evaluated before being used as an anchor in the equating. This stability check involved the examination of the displacement values provided in the BIGSTEPS/WINSTEPS output. Anchor items with displacements larger than 0.30 were "freed" in the calibration process.

The administration of the field test forms for the Regents Examination in Integrated Algebra used a spiraled design. In this design, equivalent groups of students were administered the various mini field test forms, including two anchor forms. With this design, the forms can be calibrated using the Rasch and PCM models and equated using an equivalent groups equating design as mentioned above.

The field testing was conducted in the spring of 2008. These items were then calibrated and placed onto the same scale. Operational test forms were constructed for June and August administrations in 2009 and January administration in 2010. Pre-equating procedure was employed to make the test forms as parallel as possible so that the cut scores will maintain similar magnitude across other forms.

Scoring Tables

As a result from the item analysis, each item in the bank has a Rasch difficulty that is on the same scale as all other items in the bank. As items are selected for use on the operational test forms, the average item difficulty of the test forms indicates the difficulty of the test relative to previous test forms. This relationship influences the resulting raw scores associated with the scale scores of 65 and 85.

Using equated Rasch difficulties and step values, a raw-score-to-theta (e.g., student ability) relationship is developed. Each theta represents a level of student performance needed to attain each raw score that can be compared across test forms. Using this relationship, the level of student performance needed to attain each scale score (e.g., 65 and 85) is held constant across test forms. That means that if a particular test form is more difficult than another, students are not penalized. This process of equating the scoring tables will cause an adjustment on the more difficult form and the 65 and/or 85 scale score will be assigned to a lower raw score. If a particular test form is easier than another, students are not unfairly advantaged either. This process of equating the scoring tables will also cause an adjustment on the much easier form and assign the 65 and/or 85 scale score to a higher raw score. With this adjustment, a constant expectation of student performance is maintained across test forms.

Pre-equating and Post-equating Contrast

Post-equating also was conducted, using samples collected after the June 2009 administrations of the New York State Regents Examination in Integrated Algebra, to observe how robust the pre-equating procedure was and the impact that it had on student-achievement level classifications.

The following steps are used to conduct post-equating:

1. Conduct a free calibration on the operational data to obtain item parameter estimates and raw score to theta conversion table.
2. The mean of the item parameters obtained from step 1 was computed
3. Equating constant (-0.08) was obtained by subtracting the mean obtained at step 2 from the corresponding mean value based on pre-equating item parameters.
4. Add the equating constant to all item parameter estimates and theta estimates obtained from operational data.
5. After step 4, post equating results were equated to the operational reporting scale.

Table 19 presents the Rasch Item Difficulties (RIDs) for the pre-equating model, the post-equating model, and the differences between the 2 models. As can be observed from this table, the p values tended to be higher when based on operational data compared with their corresponding p values based on field-testing data. Such observation may be partly due to the separate field-testing session, in which a student's motivation tends not to be ideal.

The average absolute difference between the RIDs was 0.28 for the 2009 administration items. The correlation between pre-equating and post-operational RIDs was 0.90. In most applications of the Rasch model, correlations between RIDs obtained between the two administrations are expected to be above 0.90 and average absolute differences are expected to be below 0.20. Thus, these results suggest some degree of dissimilarities in terms of item parameter estimates between the pre-equating and post-equating RIDs for the 2009 administration.

Scoring tables display the relationship between the raw score and theta (e.g., student ability). Specifically, the field test equated item parameters (e.g., RIDs) were used to develop the scoring table for the pre-equating model. On the other hand, the scaling constants (-0.08) were added to the scoring table created by post-equating RIDs in order to place those on the same scale as was used for pre-equating.

Table 19. Contrasts between Pre-equated and Post-operational Item Parameter Estimates

Item	Pre-equated Item Mean	Post - operational Item Mean	Pre-equated Item Parameters	Post - operational Item Parameters	Pre-Post Difference
1	1.70	1.77	-2.33	-2.26	-0.07
2	0.88	1.19	-0.33	-0.25	-0.08
3	1.54	1.64	-1.75	-1.64	-0.11
4	1.34	1.42	-1.18	-0.91	-0.27
5	0.82	1.21	0.02	-0.32	0.34
6	1.40	1.50	-1.37	-1.15	-0.22
7	1.38	1.52	-1.32	-1.22	-0.10
8	1.36	1.52	-1.47	-1.20	-0.27
9	1.24	1.40	-0.97	-0.83	-0.14
10	1.00	1.21	-0.37	-0.31	-0.06
11	0.98	1.16	-0.51	-0.18	-0.33
12	0.98	1.19	-0.47	-0.25	-0.22
13	0.94	1.00	-0.23	0.25	-0.48
14	0.92	1.21	-0.17	-0.31	0.14
15	0.90	1.14	-0.18	-0.12	-0.06
16	0.82	1.13	0.02	-0.09	0.11
17	0.80	0.96	-0.07	0.34	-0.41
18	0.80	1.39	-0.11	-0.81	0.70
19	0.76	0.94	0.20	0.38	-0.18
20	0.74	0.92	0.23	0.43	-0.20
21	0.72	0.98	0.09	0.28	-0.19
22	0.72	1.26	0.14	-0.44	0.58
23	0.70	0.92	0.33	0.45	-0.12
24	1.28	1.46	-1.06	-1.04	-0.02
25	0.74	0.95	0.03	0.37	-0.34
26	0.62	0.90	0.52	0.05	0.02
27	0.58	0.91	0.65	0.46	0.19
28	0.56	0.85	0.73	0.63	0.10
29	0.36	0.48	1.09	1.72	-0.63
30	0.34	0.64	1.45	1.21	0.24
31	0.41	0.95	0.62	0.33	0.29
32	0.59	0.82	0.48	0.64	-0.16
33	1.35	1.59	-1.00	-1.17	0.17
34	0.35	1.09	1.15	0.75	0.40
35	0.23	0.74	1.30	1.34	-0.04
36	0.63	1.30	1.34	0.56	0.78
37	1.75	1.52	-0.11	0.70	-0.81
38	0.86	2.65	0.63	-0.50	1.13
39	0.77	1.71	0.74	0.52	0.22

Figure 2 presents the scoring tables based on the 2 equating models mentioned above. The horizontal axis represents the ability estimate, and the vertical axis represents raw scores. According to the figure, the scoring tables for pre-equating and post-equating models were quite similar.

To further observe the impact between the two equating models, Table 20 was constructed, reporting raw score cuts and percent of students in each of the achievement levels based on the sample used for the analysis. For the identification of the cut score, the theta cuts from the standard setting were used. Because of the discrete nature of the scoring table, it is unlikely to have the theta values on the scoring table that exactly match the standard setting thetas. Therefore, the closest theta values to the standard setting values without going over were identified, and their corresponding raw scores were assigned to be the cut scores based on post-equating. Appendix B provides the comparison between the scoring tables based on pre-equating and post-equating.

As can be observed from Table 20, the raw score cut corresponding to the scale score of 65 was one point lower for the pre-equating results, resulting in about 2.9% fewer students being classified into 0–64 and 2.9% more into 65–84 based on the pre-equating model. The raw score cut corresponding to a scale score of 85 for the post-equating model was the same under the two equating models.

There are some differences between the item parameter estimates as well as scoring tables between the pre- and post-equating models. The differences were noticeable, but the differences appeared more at the middle score level than other places across the entire distribution.

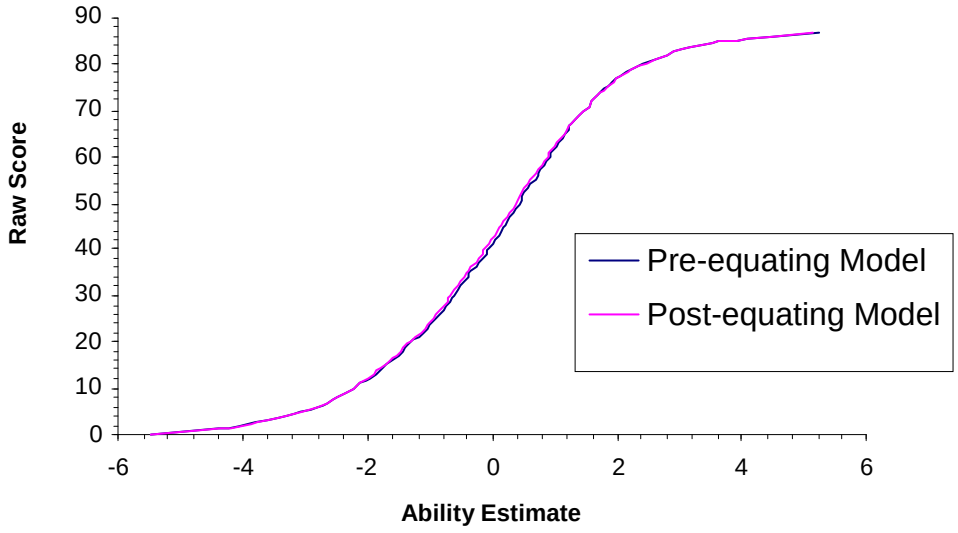

Figure 2. Comparison of Relationship between Raw Score and Ability Estimates between Pre-equating Model and Post-equating Model

Table 20. Comparisons of Raw Score Cuts and Percentages of Students in Each of the Achievement Levels between Pre-equating and Post-equating Models

	Pre-equating Model		Post-equating Model	
Scale Score	Raw Score Cut	Percent in Level	Raw Score Cut	Percent in Level
0–64		21.91		24.83
65–84	30	56.60	31	53.67
85–100	67	21.49	67	21.49

Scale Score Distribution

To observe how students performed on the Regents Examination in Integrated Algebra, the scale scores for the students included in the sample were analyzed. In Tables 21–29, frequency distributions are reported for all students, male and female students, white, Hispanic and African American students, students with ELL, ELL Using Accommodations, ELL using the translated version, students with low social economic status, SWD, SWD Using Accommodations, and grade levels. Mean and standard deviations of scale scores are also computed for all students and each of the subgroups in Table 30. Percentages of students in each of the achievement levels (0–64, 65–84, and 85–100) are reported in Table 31.

Table 21. Scale Score Distribution for All Students.

Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.
0	1	0.01	0.01	56	274	1.84	14.05	80	643	4.32	61.60
7	5	0.03	0.04	57	255	1.71	15.76	81	636	4.27	65.87
11	1	0.01	0.05	59	231	1.55	17.31	82	633	4.25	70.13
14	6	0.04	0.09	60	206	1.38	18.70	83	425	2.86	72.98
17	1	0.01	0.09	61	211	1.42	20.12	84	822	5.52	78.51
20	11	0.07	0.17	62	191	1.28	21.40	85	246	1.65	80.16
23	14	0.09	0.26	64	76	0.51	21.91	86	604	4.06	84.22
26	22	0.15	0.41	65	435	2.92	24.83	87	384	2.58	86.80
29	26	0.17	0.58	66	269	1.81	26.64	88	360	2.42	89.22
31	49	0.33	0.91	67	265	1.78	28.42	89	338	2.27	91.49
33	40	0.27	1.18	68	248	1.67	30.09	90	181	1.22	92.71
36	99	0.67	1.85	69	260	1.75	31.84	91	184	1.24	93.94
38	94	0.63	2.48	70	233	1.57	33.40	92	278	1.87	95.81
40	126	0.85	3.33	71	437	2.94	36.34	93	154	1.04	96.85
42	107	0.72	4.05	72	207	1.39	37.73	94	117	0.79	97.63
44	164	1.10	5.15	73	224	1.51	39.24	95	119	0.80	98.43
46	161	1.08	6.23	74	446	3.00	42.23	96	64	0.43	98.86
48	165	1.11	7.34	75	468	3.15	45.38	98	84	0.56	99.43
49	158	1.06	8.40	76	209	1.40	46.78	99	39	0.26	99.69
51	164	1.10	9.50	77	457	3.07	49.86	100	46	0.31	100.00
53	199	1.34	10.84	78	655	4.40	54.26				
54	203	1.36	12.21	79	449	3.02	57.28				

Table 22. Scale Score Distribution for Male Students

Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.
7	3	0.04	0.04	56	125	1.81	13.94	79	195	2.82	58.20
11	1	0.01	0.06	57	112	1.62	15.55	80	315	4.55	62.75
14	3	0.04	0.10	59	112	1.62	17.17	81	279	4.03	66.78
17	1	0.01	0.12	60	92	1.33	18.50	82	282	4.07	70.85
20	6	0.09	0.20	61	95	1.37	19.87	83	201	2.90	73.76
23	5	0.07	0.27	62	86	1.24	21.11	84	381	5.50	79.26
26	9	0.13	0.40	64	39	0.56	21.68	85	122	1.76	81.02
29	15	0.22	0.62	65	210	3.03	24.71	86	256	3.70	84.72
31	24	0.35	0.97	66	142	2.05	26.76	87	162	2.34	87.06
33	15	0.22	1.18	67	118	1.70	28.47	88	171	2.47	89.53
36	52	0.75	1.94	68	127	1.83	30.30	89	142	2.05	91.58
38	44	0.64	2.57	69	119	1.72	32.02	90	76	1.10	92.68
40	60	0.87	3.44	70	122	1.76	33.78	91	87	1.26	93.93
42	39	0.56	4.00	71	214	3.09	36.87	92	135	1.95	95.88
44	76	1.10	5.10	72	107	1.55	38.42	93	70	1.01	96.89
46	80	1.16	6.25	73	95	1.37	39.79	94	48	0.69	97.59
48	73	1.05	7.31	74	246	3.55	43.34	95	50	0.72	98.31
49	78	1.13	8.43	75	209	3.02	46.36	96	35	0.51	98.82
51	83	1.20	9.63	76	109	1.57	47.93	98	40	0.58	99.39
53	80	1.16	10.79	77	217	3.13	51.07	99	17	0.25	99.64
54	93	1.34	12.13	78	299	4.32	55.39	100	25	0.36	100.00

Table 23. Scale Score Distribution for Female Students

Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.
0	1	0.01	0.01	59	108	1.42	16.14	81	352	4.63	63.90
14	3	0.04	0.05	60	99	1.30	17.44	82	350	4.60	68.50
20	5	0.07	0.12	61	109	1.43	18.87	83	221	2.91	71.41
23	6	0.08	0.20	62	94	1.24	20.11	84	435	5.72	77.13
26	11	0.14	0.34	64	36	0.47	20.58	85	121	1.59	78.72
29	11	0.14	0.49	65	206	2.71	23.29	86	341	4.48	83.21
31	23	0.30	0.79	66	124	1.63	24.92	87	219	2.88	86.09
33	20	0.26	1.05	67	139	1.83	26.75	88	186	2.45	88.53
36	42	0.55	1.60	68	111	1.46	28.21	89	194	2.55	91.08
38	48	0.63	2.24	69	134	1.76	29.97	90	104	1.37	92.45
40	57	0.75	2.99	70	107	1.41	31.38	91	96	1.26	93.71
42	59	0.78	3.76	71	205	2.70	34.07	92	142	1.87	95.58
44	77	1.01	4.77	72	95	1.25	35.32	93	84	1.10	96.69
46	72	0.95	5.72	73	124	1.63	36.95	94	69	0.91	97.59
48	75	0.99	6.71	74	193	2.54	39.49	95	67	0.88	98.47
49	69	0.91	7.61	75	251	3.30	42.79	96	29	0.38	98.86
51	64	0.84	8.46	76	97	1.28	44.07	98	44	0.58	99.43
53	105	1.38	9.84	77	234	3.08	47.15	99	22	0.29	99.72
54	98	1.29	11.13	78	350	4.60	51.75	100	21	0.28	100.00
56	142	1.87	12.99	79	251	3.30	55.05				
57	131	1.72	14.72	80	321	4.22	59.27				

Table 24. Scale Score Distribution for White Students

Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.
7	1	0.01	0.01	59	57	0.79	7.12	81	391	5.39	54.82
17	1	0.01	0.03	60	46	0.63	7.75	82	405	5.59	60.41
20	3	0.04	0.07	61	56	0.77	8.53	83	277	3.82	64.23
23	2	0.03	0.10	62	72	0.99	9.52	84	526	7.26	71.49
26	1	0.01	0.11	64	22	0.30	9.82	85	162	2.23	73.72
29	3	0.04	0.15	65	130	1.79	11.62	86	413	5.70	79.42
31	5	0.07	0.22	66	109	1.50	13.12	87	257	3.55	82.96
33	5	0.07	0.29	67	100	1.38	14.50	88	225	3.10	86.07
36	9	0.12	0.41	68	93	1.28	15.78	89	218	3.01	89.07
38	13	0.18	0.59	69	101	1.39	17.17	90	127	1.75	90.83
40	17	0.23	0.83	70	109	1.50	18.68	91	125	1.72	92.55
42	18	0.25	1.08	71	172	2.37	21.05	92	168	2.32	94.87
44	36	0.50	1.57	72	105	1.45	22.50	93	92	1.27	96.14
46	21	0.29	1.86	73	104	1.43	23.93	94	77	1.06	97.20
48	35	0.48	2.35	74	209	2.88	26.82	95	73	1.01	98.21
49	32	0.44	2.79	75	236	3.26	30.07	96	36	0.50	98.70
51	36	0.50	3.28	76	109	1.50	31.58	98	46	0.63	99.34
53	44	0.61	3.89	77	249	3.43	35.01	99	22	0.30	99.64
54	46	0.63	4.52	78	393	5.42	40.43	100	26	0.36	100.00
56	75	1.03	5.56	79	275	3.79	44.23				
57	56	0.77	6.33	80	377	5.20	49.43				

Table 25. Scale Score Distribution for Hispanic Students

Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.
14	1	0.05	0.05	59	49	2.23	24.68	80	65	2.96	79.51
20	1	0.05	0.09	60	54	2.46	27.14	81	60	2.73	82.24
23	2	0.09	0.18	61	45	2.05	29.19	82	77	3.51	85.75
26	5	0.23	0.41	62	37	1.68	30.87	83	44	2.00	87.75
29	5	0.23	0.64	64	14	0.64	31.51	84	72	3.28	91.03
31	11	0.50	1.14	65	117	5.33	36.84	85	17	0.77	91.80
33	7	0.32	1.46	66	66	3.01	39.85	86	49	2.23	94.03
36	15	0.68	2.14	67	53	2.41	42.26	87	37	1.68	95.72
38	13	0.59	2.73	68	41	1.87	44.13	88	23	1.05	96.77
40	31	1.41	4.14	69	56	2.55	46.68	89	15	0.68	97.45
42	27	1.23	5.37	70	51	2.32	49.00	90	5	0.23	97.68
44	29	1.32	6.69	71	103	4.69	53.69	91	8	0.36	98.04
46	35	1.59	8.29	72	40	1.82	55.51	92	14	0.64	98.68
48	42	1.91	10.20	73	45	2.05	57.56	93	13	0.59	99.27
49	26	1.18	11.38	74	88	4.01	61.57	94	4	0.18	99.45
51	40	1.82	13.21	75	85	3.87	65.44	95	6	0.27	99.73
53	37	1.68	14.89	76	32	1.46	66.89	96	3	0.14	99.86
54	45	2.05	16.94	77	69	3.14	70.04	98	3	0.14	100.00
56	65	2.96	19.90	78	88	4.01	74.04				
57	56	2.55	22.45	79	55	2.50	76.55				

Table 26. Scale Score Distribution for African American Students

Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.
7	2	0.10	0.10	56	73	3.59	32.07	78	55	2.70	81.78
11	1	0.05	0.15	57	65	3.19	35.27	79	37	1.82	83.60
14	3	0.15	0.29	59	59	2.90	38.16	80	65	3.19	86.79
20	6	0.29	0.59	60	55	2.70	40.86	81	48	2.36	89.15
23	3	0.15	0.74	61	60	2.95	43.81	82	43	2.11	91.26
26	8	0.39	1.13	62	36	1.77	45.58	83	20	0.98	92.24
29	13	0.64	1.77	64	12	0.59	46.17	84	43	2.11	94.35
31	18	0.88	2.65	65	87	4.27	50.44	85	15	0.74	95.09
33	13	0.64	3.29	66	54	2.65	53.09	86	26	1.28	96.37
36	39	1.92	5.21	67	47	2.31	55.40	87	8	0.39	96.76
38	39	1.92	7.12	68	45	2.21	57.61	88	25	1.23	97.99
40	42	2.06	9.18	69	59	2.90	60.51	89	13	0.64	98.62
42	38	1.87	11.05	70	33	1.62	62.13	90	4	0.20	98.82
44	49	2.41	13.46	71	67	3.29	65.42	91	7	0.34	99.17
46	60	2.95	16.40	72	25	1.23	66.65	92	6	0.29	99.46
48	46	2.26	18.66	73	37	1.82	68.47	93	5	0.25	99.71
49	47	2.31	20.97	74	71	3.49	71.95	94	1	0.05	99.75
51	46	2.26	23.23	75	62	3.05	75.00	95	3	0.15	99.90
53	52	2.55	25.79	76	28	1.38	76.38	96	1	0.05	99.95
54	55	2.70	28.49	77	55	2.70	79.08	98	1	0.05	100.00

Table 27. Scale Score Distribution for Students with Limited English Proficiency

Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.
20	4	0.95	0.95	57	14	3.34	37.23	77	10	2.39	82.34
23	1	0.24	1.19	59	16	3.82	41.05	78	7	1.67	84.01
26	4	0.95	2.15	60	8	1.91	42.96	79	8	1.91	85.92
29	3	0.72	2.86	61	14	3.34	46.30	80	7	1.67	87.59
31	1	0.24	3.10	62	10	2.39	48.69	81	6	1.43	89.02
33	2	0.48	3.58	64	3	0.72	49.40	82	6	1.43	90.45
36	6	1.43	5.01	65	24	5.73	55.13	83	9	2.15	92.60
38	4	0.95	5.97	66	14	3.34	58.47	84	6	1.43	94.03
40	7	1.67	7.64	67	11	2.63	61.10	86	6	1.43	95.47
42	4	0.95	8.59	68	11	2.63	63.72	87	1	0.24	95.70
44	18	4.30	12.89	69	11	2.63	66.35	88	6	1.43	97.14
46	12	2.86	15.75	70	8	1.91	68.26	89	3	0.72	97.85
48	16	3.82	19.57	71	12	2.86	71.12	91	1	0.24	98.09
49	11	2.63	22.20	72	4	0.95	72.08	92	2	0.48	98.57
51	8	1.91	24.11	73	5	1.19	73.27	93	1	0.24	98.81
53	12	2.86	26.97	74	17	4.06	77.33	94	1	0.24	99.05
54	12	2.86	29.83	75	9	2.15	79.47	95	2	0.48	99.52
56	17	4.06	33.89	76	2	0.48	79.95	98	2	0.48	100.00

Table 28. Scale Score Distribution for Students with Low Social Economic Status

Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.
0	1	0.01	0.01	56	199	2.90	23.09	80	217	3.16	73.93
7	5	0.07	0.09	57	174	2.54	25.63	81	208	3.03	76.96
11	1	0.01	0.10	59	161	2.35	27.98	82	171	2.49	79.46
14	4	0.06	0.16	60	133	1.94	29.91	83	132	1.92	81.38
17	1	0.01	0.17	61	144	2.10	32.01	84	247	3.60	84.98
20	7	0.10	0.28	62	121	1.76	33.78	85	67	0.98	85.95
23	13	0.19	0.47	64	48	0.70	34.47	86	171	2.49	88.45
26	19	0.28	0.74	65	282	4.11	38.58	87	110	1.60	90.05
29	23	0.34	1.08	66	151	2.20	40.78	88	108	1.57	91.62
31	42	0.61	1.69	67	148	2.16	42.94	89	107	1.56	93.18
33	31	0.45	2.14	68	135	1.97	44.91	90	52	0.76	93.94
36	85	1.24	3.38	69	137	2.00	46.90	91	59	0.86	94.80
38	74	1.08	4.46	70	114	1.66	48.56	92	102	1.49	96.28
40	100	1.46	5.92	71	241	3.51	52.08	93	52	0.76	97.04
42	88	1.28	7.20	72	94	1.37	53.45	94	36	0.52	97.57
44	128	1.87	9.06	73	116	1.69	55.14	95	55	0.80	98.37
46	124	1.81	10.87	74	214	3.12	58.25	96	24	0.35	98.72
48	122	1.78	12.65	75	201	2.93	61.18	98	45	0.66	99.37
49	109	1.59	14.24	76	92	1.34	62.52	99	20	0.29	99.66
51	127	1.85	16.09	77	175	2.55	65.07	100	23	0.34	100.00
53	145	2.11	18.20	78	235	3.42	68.50				
54	137	2.00	20.20	79	156	2.27	70.77				

Table 29. Scale Score Distribution for Students with Disabilities

Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.	Scale Score	Freq.	Percent	Cum. Per.
7	3	0.22	0.22	56	60	4.39	38.58	77	33	2.42	84.33
14	2	0.15	0.37	57	49	3.59	42.17	78	33	2.42	86.75
20	4	0.29	0.66	59	37	2.71	44.88	79	25	1.83	88.58
23	4	0.29	0.95	60	37	2.71	47.58	80	30	2.20	90.78
26	14	1.02	1.98	61	36	2.64	50.22	81	29	2.12	92.90
29	11	0.81	2.78	62	41	3.00	53.22	82	26	1.90	94.80
31	21	1.54	4.32	64	8	0.59	53.81	83	9	0.66	95.46
33	11	0.81	5.12	65	56	4.10	57.91	84	25	1.83	97.29
36	39	2.86	7.98	66	35	2.56	60.47	85	3	0.22	97.51
38	20	1.46	9.44	67	35	2.56	63.03	86	10	0.73	98.24
40	41	3.00	12.45	68	24	1.76	64.79	87	6	0.44	98.68
42	21	1.54	13.98	69	28	2.05	66.84	88	3	0.22	98.90
44	48	3.51	17.50	70	27	1.98	68.81	89	5	0.37	99.27
46	35	2.56	20.06	71	41	3.00	71.82	90	2	0.15	99.41
48	44	3.22	23.28	72	28	2.05	73.87	91	2	0.15	99.56
49	39	2.86	26.13	73	25	1.83	75.70	92	2	0.15	99.71
51	40	2.93	29.06	74	31	2.27	77.96	94	1	0.07	99.78
53	30	2.20	31.26	75	38	2.78	80.75	95	2	0.15	99.93
54	40	2.93	34.19	76	16	1.17	81.92	98	1	0.07	100.00

Table 30. Descriptive Statistics on Scale Scores for Various Student Groups

	Number of Students	Percent	Mean	Standard Deviation
All Students	14,879	N.A.	73.62	14.45
Female	7,604	51.11	74.33	14.17
Male	6,924	46.54	73.47	14.41
African American	2,036	13.68	63.51	15.37
Hispanic	2,196	14.76	68.49	13.78
White	7,249	48.72	78.42	11.14
ELL	419	2.82	62.65	15.23
ELL/SUA	318	2.14	62.07	15.08
ELL/Translated Editions	160	1.08	64.91	12.50
Low SES	6,863	46.13	68.72	15.92
SWD	1,366	9.18	60.69	15.52
SWD/SUA	1,277	8.58	60.55	15.63
Grade 8	4,471	30.05	84.61	7.55
Grade 9	7,244	48.69	71.69	13.21
Grade 10	2,018	13.56	62.96	13.20
Grade 11	507	3.41	59.90	12.98
Grade 12	198	1.33	57.28	13.78

Table 31. Performance Classification for Various Student Groups

	Number of Students	Percent (All Students)	Percent (0-64)	Percent (65-84)	Percent (85-100)
All Students	14,879	N.A.	21.91	56.60	21.49
Female	7,604	51.11	20.58	56.55	22.87
Male	6,924	46.54	21.68	57.58	20.74
African American	2,036	13.68	46.17	48.18	5.65
Hispanic	2,196	14.76	31.51	59.52	8.97
White	7,249	48.72	9.82	61.66	28.51
ELL	419	2.82	49.40	44.63	5.97
ELL/SUA	318	2.14	50.00	44.34	5.66
ELL/Translated Editions	160	1.08	41.25	53.75	5.00
Low SES	6,863	46.13	34.47	50.50	15.02
SWD	1,366	9.18	53.81	43.48	2.71
SWD/SUA	1,277	8.58	54.27	42.83	2.90
Grade 8	4,471	30.05	1.66	46.68	51.67
Grade 9	7,244	48.69	22.56	66.52	10.92
Grade 10	2,018	13.56	45.99	53.37	0.64
Grade 11	507	3.41	58.78	40.43	0.79
Grade 12	198	1.33	62.63	36.87	0.51

Quality Assurance

The Regents examinations program and its associated data play an important role in the State accountability system as well as in many local evaluation plans. Therefore, it is vital that quality control procedures, which ensure the accuracy of student, school and district-level data and reports, are implemented. A set of quality control procedures has been developed and refined to help ensure that the testing quality assurance requirements are met or exceeded. These quality control procedures are detailed in the paragraphs that follow.

Field Test

Before items are placed on an operational test form of the Regents examinations, they have to go through several phases of reviews and field-testing to ensure the quality of the test. During field testing, items are tried out to observe their statistical behaviors. After field testing, the answer sheets are collected from students, and scanned at NYSED. Quality control procedures and regular preventative maintenance ensure that the NYSED scanner is functioning properly at all times.

To score essay items, rangefinding is conducted first to define detailed rubrics for the items. Next, scorers are trained through a set of rigorous procedures to ensure consistent ratings. For each rangefinding session, Pearson sent a meeting coordinator, a recorder, and a scoring director (who had responsibility for each given content area) to work with each committee. The primary goal of the rangefinder meetings is to identify and select a representative sample of student responses for each item for use as exemplar papers for each of the content areas for each of the items. These responses accurately represent the range of student achievement levels described in the rubric for each item, as interpreted by the committee members in each session. Careful selection of papers during rangefinding and the subsequent compilation of anchor papers and other training materials are essential to ensuring that scoring can be conducted consistently and reliably. All the necessary steps were also taken to ensure the security of the materials. Scoring directors kept a formal log of all papers discussed, recording all scores assigned along with any recommendations for the placement of papers in training sets. In addition, scoring directors noted the comments of committee members on the scoring of particular papers, as these comments are useful in the training of scorers (helping to ensure that scorers understand and implement the committee's wishes), and to provide benchmark points for discussions in subsequent years to help ensure longitudinal consistency. This master list also serves as a tracking log, including information on the placement of each paper in training sets.

After rangefinding, scoring supervisors and scorers are selected and trained to score field test items. Scoring supervisors had college degrees in the subject area or a related area. Supervisors had experience in scoring the subject area

and demonstrated strong organizational abilities and communication skills. Each scorer possessed, at a minimum, a 4-year college degree. They were assigned to work in the most appropriate subject area based on their educational qualifications and their work or scoring experience. Scoring directors or supervisors began training by reviewing and discussing the scoring guides and anchor sets for items in a book. Scoring directors or supervisors then gave scorers practice sets, and scorers assigned scores to these sample responses. After scorers completed the set, scoring directors or supervisors reviewed and explained true scores for the practice papers. Subsequent practice sets were processed in the same manner. If scorer performance or discussion of the practice sets indicated a need for reviewing or retraining, it occurred at that time. Scorers were expected to meet quality standards during training and scoring. Scorers who failed to meet those quality standards were released from the project. Quality control steps taken during the project were:

- **Backreading (read behinds)** was one of the primary responsibilities of scoring directors and scoring supervisors. It was an immediate source of information on scoring accuracy and quickly alerted scoring directors and supervisors to misconceptions at the team level, indicating the need to review or retrain. Backreading continued throughout the project. Supervisors increased focus on scorers whose scoring accuracy, based on statistical reports or backreading records, was falling below expectations.

- **Second Scoring** began immediately with 10% of responses each receiving an independent scoring by a second scorer.

- **Reports** were available throughout the project and were monitored daily by the program manager and scoring directors. These reports included the inter-rater reliability and frequency distribution for individual scorers and for teams. To remain on the project, scorers whose statistics were not meeting quality expectations received retraining and had to demonstrate the ability to meet expectations.

Test Construction

Stringent quality assurance procedures are employed in the test construction phase. To select items for an operational test, content specifications are carefully followed to ensure a representative set of items for each of the standards. In the meantime, item statistics obtained from the field tests are also reviewed to make sure the statistical property of the test is taken into consideration. NYSED assessment specialists and research staff work closely with test development specialists at Riverside in this endeavor. A set of procedures is followed to obtain an operational test form with desired content and statistical properties. Item and form statistical characteristics from the baseline test are used as targets when constructing the current test form. Once a set of items has been selected,

psychometricians and content specialists from both NYSED and Riverside review and consider replacement items for a variety of reasons. Test maps are then created, with content specifications, answer keys, and field test statistics. Multiple reviews are conducted of the test map to ensure its accuracy. Test construction is an iterative process and goes through many phases before a final test form is constructed and printed for administration. To ensure test security, locked boxes are utilized whenever secure materials are transported between NYSED and their contractors, such as Riverside and Pearson.

Quality Control for Test Form Equating

Test form equating is the process that enables fair and equitable comparisons across test forms. A pre-equating model is typically used for Regents examinations. As mentioned in the Equating, Scaling, and Scoring section, various procedures are employed to ensure the quality of the equating procedure. Refer to that section for the detailed and specific procedures followed in this process. Periodically, samples of subjects were selected and their item responses collected. Post-equating then is employed to the representative sample to evaluate the pre-equating scaling tables, as was the case with the June 2009 administration.

References

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. Joint Technical Committee, (1999). *The Standards for Educational and Psychological Testing.*

Allen, Nancy L., James E. Carlson, and Christine A. Zalanak. 1999. *The NAEP 1996 Technical Report.* Washington, DC: National Center for Education Statistics.

Cattell, R. B. (1966). The Scree Test for the Number of Factors. *Multivariate Behavioral Research, 1*, 245–276.

Dorans, Neil J., and Holland, Paul W. 1992. DIF Detection and Description: Mantel–Haenszel and Standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 35–66). Hillsdale, NJ: Erlbaum.

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement, 20*, 141–151.

Linn, R. L., & Gronlund, N. E., (1995). *Measurement in Assessment and Teaching*, 7th edition. New Jersey: Prentice–Hall.

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

Rudner, L. M. (2005). Expected Classification Accuracy. *Practical Assessment, Research & Evaluation. Vol. 10, Number 13.*

Appendix A

Table A 1. Percentage of Students Included in Sample for Each Option (MC Only)

Item Position	Item Type	Key	Item Category Percentage			
			A	B	C	D
1	Multiple-Choice	D	2.93	3.94	4.48	88.65
2	Multiple-Choice	D	8.88	11.88	19.55	59.68
3	Multiple-Choice	A	82.00	12.20	4.66	1.14
4	Multiple-Choice	B	5.39	71.34	16.26	7.00
5	Multiple-Choice	C	7.84	18.01	61.00	13.15
6	Multiple-Choice	D	3.23	3.49	18.20	75.08
7	Multiple-Choice	A	76.29	6.59	12.37	4.75
8	Multiple-Choice	B	8.55	76.00	8.80	6.65
9	Multiple-Choice	C	8.36	16.39	69.99	5.27
10	Multiple-Choice	B	12.10	60.64	13.11	14.14
11	Multiple-Choice	D	33.41	3.72	4.58	58.29
12	Multiple-Choice	B	6.37	59.57	31.74	2.33
13	Multiple-Choice	C	20.97	18.82	50.18	10.04
14	Multiple-Choice	A	60.71	22.13	11.60	5.56
15	Multiple-Choice	C	9.47	10.38	57.05	23.10
16	Multiple-Choice	D	9.53	23.40	10.56	56.51
17	Multiple-Choice	A	48.41	11.59	19.34	20.66
18	Multiple-Choice	A	69.80	14.27	8.69	7.24
19	Multiple-Choice	C	7.45	8.57	47.39	36.59
20	Multiple-Choice	A	46.49	18.79	22.92	11.79
21	Multiple-Choice	B	5.20	49.34	15.35	30.12
22	Multiple-Choice	A	63.05	22.64	9.39	4.92
23	Multiple-Choice	B	20.87	46.14	16.02	16.97
24	Multiple-Choice	C	9.67	9.24	73.39	7.70
25	Multiple-Choice	B	17.96	47.86	24.01	10.17
26	Multiple-Choice	C	47.39	3.57	45.18	3.86
27	Multiple-Choice	D	39.73	5.40	8.99	45.88
28	Multiple-Choice	B	14.47	42.76	29.23	13.54
29	Multiple-Choice	B	38.88	24.14	6.18	30.80
30	Multiple-Choice	D	29.77	27.50	10.45	32.28

Table A 2. Percentage of Students Included in Sample at Each Possible Score Credit (CR only)

Item Position	Item Type	Max Credits	Item Category Percentage				
			0	1	2	3	4
31	Constructed-Response	2	43.81	17.00	39.18		
32	Constructed-Response	2	48.16	21.76	30.08		
33	Constructed-Response	2	13.62	13.19	73.18		
34	Constructed-Response	3	49.79	16.92	7.04	26.25	
35	Constructed-Response	3	60.99	17.53	7.47	14.01	
36	Constructed-Response	3	16.22	48.39	24.21	11.18	
37	Constructed-Response	4	46.30	9.69	12.44	8.39	23.17
38	Constructed-Response	4	10.69	11.60	16.84	23.54	37.33
39	Constructed-Response	4	30.63	18.53	19.33	11.12	20.40

Pre-equating and Post-equating Scoring Tables

Table B 1. Comparison of Pre-equating and Post-equating Scoring Tables.

Raw Score	Pre-Equating	Post-Equating		Raw Score	Pre-Equating	Post-Equating
0	-5.503	-5.473		44	0.141	0.083
1	-4.780	-4.753		45	0.190	0.133
2	-4.057	-4.033		46	0.240	0.182
3	-3.622	-3.602		47	0.289	0.232
4	-3.306	-3.290		48	0.337	0.281
5	-3.055	-3.042		49	0.385	0.330
6	-2.845	-2.835		50	0.433	0.379
7	-2.663	-2.657		51	0.480	0.428
8	-2.503	-2.500		52	0.528	0.476
9	-2.358	-2.359		53	0.575	0.525
10	-2.227	-2.231		54	0.622	0.574
11	-2.105	-2.113		55	0.669	0.623
12	-1.992	-2.003		56	0.716	0.672
13	-1.887	-1.900		57	0.764	0.722
14	-1.787	-1.804		58	0.811	0.772
15	-1.693	-1.712		59	0.859	0.823
16	-1.603	-1.625		60	0.908	0.874
17	-1.517	-1.542		61	0.957	0.927
18	-1.435	-1.462		62	1.007	0.980
19	-1.355	-1.385		63	1.057	1.034
20	-1.279	-1.311		64	1.109	1.089
21	-1.205	-1.239		65	1.162	1.146
22	-1.133	-1.169		66	1.217	1.205
23	-1.063	-1.101		67	1.273	1.265
24	-0.996	-1.035		68	1.331	1.328
25	-0.929	-0.971		69	1.392	1.393
26	-0.865	-0.908		70	1.455	1.460
27	-0.801	-0.846		71	1.521	1.531
28	-0.739	-0.785		72	1.592	1.606
29	-0.678	-0.726		73	1.666	1.684
30	-0.618	-0.667		74	1.746	1.768
31	-0.559	-0.609		75	1.832	1.857
32	-0.501	-0.552		76	1.926	1.953
33	-0.444	-0.496		77	2.029	2.057
34	-0.388	-0.441		78	2.144	2.171
35	-0.332	-0.386		79	2.272	2.298
36	-0.277	-0.332		80	2.417	2.441
37	-0.223	-0.278		81	2.586	2.604
38	-0.169	-0.225		82	2.784	2.797
39	-0.116	-0.173		83	3.026	3.030
40	-0.063	-0.121		84	3.335	3.329
41	-0.012	-0.069		85	3.764	3.747
42	0.040	-0.018		86	4.482	4.453
43	0.090	0.032		87	5.200	5.159