The New York State Examination Program
Department Review of August 2010
Regents Examination in Integrated Algebra

Technical Report

April 2011

Office of Assessment, Policy, Development and Administration
The New York State Education Department

I. Introduction

This report summarizes the results of the Department Review of the Regents Examination in Integrated Algebra administered in August 2010. Department Review is an internal audit process conducted by the New York State Education Department to ensure test score reliability of the Regents Examinations. Each year, the New York State Education Department (SED) conducts audits of New York school teachers' local scoring of a selected number of Regents Examinations. In the 2010-2011 school year, the August Regents Examination in Integrated Algebra was chosen for Department Review. A sample of 584 Integrated Algebra test papers from 67 high schools across New York State was rescored by an independent panel of sixteen scorers convened by the Department. The sixteen scorers were broken into five teams. The 2010 Department Review included the rescoring of all nine constructed response (CR) items and the mechanical review of the total test scores, as well as the subscores for each part of the examination.

The purpose of the Department Review is to provide important evidence for test reliability and inter-rater reliability of the Regents examinations. The audit process allows the Department to evaluate the extent to which teachers and committees of teachers are properly applying the scoring rubrics and scoring guides when scoring the CR items. Department Review also acts as a check on to schools and teachers to ensure that they score tests properly in accordance with overall SED directions and oversight. The process also provides feedback to schools, which can lead them to improve their scoring procedures and enhance compliance with the scoring rubrics. The process of Department Review is an essential element for maintaining overall test reliability.

II. Sample Section and Responses

A sample of 71 schools was selected for the Department Review from all the 626 middle and high schools in New York State that ordered Integrated Algebra Examinations for the August 2010 administration. Three of the schools were included because they were on department watch list. A total of 67 schools submitted test papers to the Department. Due to the limited resources for rescoring the test papers, up to 10 test papers were randomly selected from each responding school's total submitted papers for rescoring. This process yielded a total of 584 test papers for the Department Review. This number represented approximately 2.7% of the August Regents Integrated Algebra examination takers and approximately 6% of papers submitted from the total participating schools.

It is essential that the audit sample represents the test population. Stratified sampling design at the school level was adopted based on Need/Resource Category. Table 1 shows the distribution of test population, selected audit sample, and actual audit sample percentages of Need/Resource Categories. The results indicate that the distribution of the sample test-takers approximates the Integrated Algebra examination population. For example, approximately 36% of the August Regents Integrated Algebra examination takers were from New York City, compared to 38% selected and 40% actual audited test papers from New York City.

Table 1. Distribution of August 2010 Integrated Algebra Audit Sample Schools

Need/Resource Category	Population*		Selected Sample**		Actual Audit Sample	
	N-count	%	N-count	%	N-count	%
New York City	224	35.8%	24	35.3%	27	40.3%
Big 4 Cities	23	3.7%	3	4.4%	3	4.5%
High-Need Urban/Suburban	37	5.9%	3	4.4%	4	6.0%
High-Need Rural	45	7.2%	8	11.8%	6	9.0%
Average Need	146	23.3%	15	22.1%	13	19.4%
Low Need	45	7.2%	5	7.4%	6	9.0%
Charter	18	2.9%	3	4.3%	2	3.0%
Nonpublic	88	14.1%	7	10.4%	6	9.0%
Total	626	100%	68	100%	67	100%
* Based on August 2010 operational test data.						
** Ten papers were randomly selected from each of the sample schools.						

III. Rescoring Procedures:

Mechanical Review: The purpose of the mechanical review was to determine whether school scorers added student scores from the answer sheet correctly and recorded the correct total score and subtotal score, each of the four parts of the examination (i.e., Part I containing multiple-choice (MC) items and Parts II, III, and IV containing constructed response (CR) items). To conduct the mechanical review, SED clerks added the raw scores from the students' answer sheets and generated subscores for the four parts of the exam and the total raw score. The mechanical-review total raw scores were then converted to scale scores using the raw-to-scale conversion table for the August 2010 administration. The process yielded two sets of summary scores, the school scores and SED mechanical review scores, for each part of the examination and the total test raw score and scale score, for analysis.

Rescoring Constructed Response Items: Sixteen high school math teachers (SED raters) were convened in October 2010 as raters for the New York State Education Department Review. The task of these SED raters was to rescore a sample of student papers for the August examination in Integrated Algebra. The session was conducted by SED mathematics test owners who trained the SED raters in the procedures for rating. Some of the raters had participated in the development of the August Examination and all had a thorough understanding of the examination and scoring requirements. The raters were divided into five teams (four teams of three raters and one team of four raters). Each team was led by a team leader and was assigned to rescore a fixed number of items (totaling nine raw score points). No one team rated all items of a given student paper.

To ensure the rating consistency, each team was responsible for scoring the same questions (items) in the student papers (see Organization Chart in Appendix A). For example,

raters on Team 1 scored items 31, 34, and 37 and raters on Team 2 scored items 32, 35, and 38, etc. The team leader periodically checked the scored student papers for accuracy and consistency of the team members' work. Once a student's paper was rescored by one team, the team leader compared the credit(s) allowed by the SED raters to the credit(s) allowed by the school for each item. If the former agreed with the latter, the team leader recorded the "Final SED credit(s)" for the item and no further rescoring was done for that item.

If there were disagreements between the credit(s) allowed by SED raters and by the school, the team leader would pass this student paper to another team for a second round of rescoring (see Appendix A). The team leader of the second team compared the credit(s) allowed by SED rater 2 to both the credit(s) allowed by SED rater 1 and the credit(s) allowed by the school. If two of the three parties agreed on the item, the team leader recorded the credit(s) agreed on by two of the three parties as the "Final SED credit(s)" for that item and no further rescoring was done for that item. If no agreement was reached by two of the three parties, the team leader did a final rescoring of the item and recorded the credit(s) as the "Final SED credit(s)." The "Final SED credit(s)" and other data for all items on the test were then compiled for analysis.

Scoring data were entered into a spreadsheet together with the school packing code, student ID, and mechanical review results. Data were checked by test owners and education specialist to ensure accuracy. Originally entered data were then converted to Microsoft ACCESS or SPSS for analysis and reports generation.

IV. Data Analysis

Three sets of scores were generated and used to assess the scoring reliability: local school scores, audit scores, and mechanic review scores. The inter-rater reliability of the 2010 August Regents Examinations in Integrated Algebra was examined at multiple levels. First, at the item level, the inter-rater agreement between the school score and audit score for each CR item was examined. Second, at the total-score level, the school and audit total scores and subscores for all four parts of the exam (i.e., Part I with 30 MC items, Part II with three CR items, Part III with three CR items, and Part IV with three CR items) were compared to determine the overall inter-rater reliability. Finally, the total scores from the school scorers and the mechanic reviewers were examined for accuracy.

It is believed that no single method is adequate to determine scoring reliability. Therefore, multiple statistical methods were employed to assess the degree of agreement between school raters and SED audit raters:

1. **Item raw score agreement** as a measure of consensus between school scores and SED audit scores were examined. In this method, the percentages were calculated for exact agreement, adjacent agreement, and nonadjacent agreement.
2. **Intraclass correlation** was calculated as a measure of inter-rater reliability estimate by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects.

3. **Item mean and standard deviation** between the local school scores and SED rescores were calculated and compared as measures of average agreement/difference and variability between the two groups of scorers on a given item.
4. **Total test mean difference and correlation** between school and audit scores were computed for each of the four parts of the examination mentioned above; all multiple-choice items, all constructed response items, and total scores show degree of agreement and consistency.
5. **Internal consistency reliability** (Cronbach's Alpha) was calculated as a measure of the reliability of the constructed response portion of the test.

IV. Results:

1. Item Raw Score Agreement

Item raw score agreement measures the absolute agreement/differences deviations between school scores and audit scores. As shown in Table 2, the exact raw score agreement (i.e. school score and audit score were exactly the same) between school scores and SED audit scores for the nine CR items ranged from 77% to 95%, with a mean exact rate of 85.3%. More specifically, seven out of the nine CR items had exact agreement of 80% or higher when rounded to the nearest whole percentage. The two items that exhibited relatively lower exact agreement rate (78% and 77%, respectively) were Item 34 and Item 38. It is expected that exact agreement rate is usually higher with 2-point items than with 3- or 4-point items. The results suggest a relatively high degree of inter-rater agreement between the school scorers and the audit scorers.

Table 2: Item Raw-Score Agreement between School Score and Audit Score

Item #	Max Points	N Count	Raw Score Agreement			Intraclass Correlation
			Exact	Adjacent*	Nonadjacent**	
31	2	584	79.8%	19.3%	0.9%	0.798
32	2	584	95.4%	4.5%	0.2%	0.972
33	2	583	92.3%	7.7%	0.0%	0.919
34	3	584	78.1%	17.8%	4.1%	0.837
35	3	584	79.8%	18.5%	1.7%	0.909
36	3	584	91.6%	7.7%	0.7%	0.937
37	4	584	80.1%	15.4%	4.5%	0.880
38	4	584	76.7%	15.6%	7.7%	0.838
39	4	584	94.2%	5.1%	0.7%	0.972

*Adjacent agreement: School score and audit score differ by +/-1 raw-score point.

**Nonadjacent agreement: School score and audit score differ by +/-2 or more raw-score points.

2. Intraclass correlation

As shown in Table 2, the intraclass correlation coefficients for all CR items were high, ranging from 0.80 to 0.97. More specifically, five of the nine CR items had intraclass correlation above 0.90. The results again suggest a high degree of consistency between the school scorers and SED audit scorers.

The distribution of raw-score agreement/differences is further detailed in Table 3. The positive raw-score differences (+1, +2, and +3) indicate that school scores were higher than the audit scores by one, two, or three raw score points, while the negative score discrepancies (-1, -2, and -3) indicate that school scores were lower than the audit scores by one, two or three score points. Again, the results suggest a high degree of agreement of 80% or higher for most CR items. It should be noted that, on some of the nine CR items, the distribution of score difference indicates that the positive raw-score differences have a higher percentage than that of negative score discrepancies.

Table 3: Percentage of Raw Score Difference between School Scoring and Audit Scoring (School Score minus Audit Score)

Item #	Max Points	N Count	School Score Lower				School Score Higher		
			(-3)	(-2)	(-1)	0	(+1)	(+2)	(+3)
31	2	584	0.0%	0.2%	2.9%	79.8%	16.4%	0.7%	0.0%
32	2	584	0.0%	0.2%	2.4%	95.4%	2.1%	0.0%	0.0%
33	2	583	0.0%	0.0%	0.7%	92.3%	7.0%	0.0%	0.0%
34	3	584	0.0%	0.2%	2.9%	78.1%	14.9%	3.4%	0.5%
35	3	584	0.0%	0.3%	3.3%	79.8%	15.2%	1.2%	0.2%
36	3	584	0.0%	0.5%	2.7%	91.6%	5.0%	0.2%	0.0%
37	4	584	0.2%	0.2%	2.4%	80.1%	13.0%	3.9%	0.2%
38	4	584	0.2%	2.2%	6.5%	76.7%	9.1%	3.9%	1.4%
39	4	584	0.0%	0.2%	1.5%	94.2%	3.6%	0.5%	0.0%

3. Item Mean Score and Standard Deviation

Table 4 presents the item raw score mean and standard deviation for all CR items from both school scoring and audit scoring. The mean-score difference was also computed and tested for statistical difference using pair-t test. The mean-score comparison indicated that the school mean scores on four out of nine CR items were exactly the same or comparable between school scores and audit scores. On five items, the school mean scores were slightly higher than the audit mean scores. The standard deviations of the school and audit score were generally similar. The average mean school score was 0.80 as compared to the average audit mean score of 0.70.

Table 4: Item Mean and Standard Deviation

Item #	Max Points	N Count	Raw-Score Mean			Standard Deviation	
			School	Audit	Mean Difference	School	Audit
31	2	584	0.61	0.47	0.14	0.788	0.697
32	2	584	1.01	1.02	-0.01	0.955	0.953
33	2	583	0.39	0.32	0.07	0.701	0.678
34	3	584	0.88	0.67	0.21	1.085	1.021
35	3	584	1.32	1.18	0.14	1.213	1.182
36	3	584	0.42	0.40	0.02	0.911	0.912
37	4	584	0.84	0.66	0.18	1.237	1.161
38	4	584	1.17	1.08	0.09	1.288	1.295
39	4	584	0.54	0.51	0.03	1.182	1.192

4. Total Test Mean Scores and Correlation

Inter-rater reliability was also examined at the total test level and the subtest level. Mean score and standard deviation were computed for the following:

- Raw score for each part of the exam
  - MC items
  - CR items
- Total raw score
- Scale score

Mean-score differences and correlation between school scores and audit scores were computed.

As shown in Table 5, mean-score difference is minimal between school and audit mean scores for Part I containing MC items. Small mean-score differences were found between school and audit mean scores for Part II, Part III, and Part IV, as well as total CR items. School mean scores of total test raw and scale scores were slightly higher than the audit mean scores. However, it should be noted that the overall impact of the CR mean-score differences on the overall raw score and scale score was less than one score point out of a total of 87 raw score points and 100 scale score points.

Despite the mean difference found in the CR section of the test, the school scores correlated highly with the audit scores for each part of the examination as well as the total score, with correlation coefficients ranging from 0.934 to 0.998. The high correlations indicate a very high degree of consistency between school and audit scoring results.

Table 5: Total Test Mean Score and Correlation

Item #	Max Points	N Count	Raw-Score Mean			Raw-Score SD		Corr. Between School and Audit Scores
			School	Audit	Diff.	School	Audit	
Part I (MC)	60	584	25.53	25.57	-0.04	10.175	10.156	0.998
Part II (CR)	6	584	2.02	1.81	0.21	1.800	1.730	0.941
Part III (CR)	9	584	2.62	2.26	0.36	2.439	2.407	0.934
Part IV (CR)	12	584	2.55	2.24	0.31	2.967	2.931	0.938
Total CR Items	27	584	7.19	6.32	0.87	6.315	6.220	0.958
Total Raw Score	87	584	32.72	31.88	0.84	15.503	15.534	0.992
Scale Score	100	584	63.72	62.82	0.90*	12.986	13.118	0.983

5. Internal Consistency

Internal consistency is another measure of test reliability. Cronbach's Alpha was computed to measure the internal consistency of CR items in Part II, Part III, and Part IV of the examination, respectively, as well as all CR items for both the school score and audit score. The high Cronbach's Alpha coefficients suggest that the CR scores from both the school scoring and audit scoring were highly consistent and reliable. The Cronbach Alpha reliability coefficients for CR items in Part II, Part III, and Part IV ranged from 0.56 to 0.72 for the school scores and from 0.57 to 0.72 for audit scores. These coefficients appear lower than might be expected, but keeping in mind the small number of available score points (as few as six), these values are acceptable for consistency evidence. The reliability for all CR items was 0.84 for the school scores and 0.85 for the audit scores. The results suggest that the CR scores from both school and audit scoring were highly consistent.

Table 6: Internal Consistency of Constructed Response Items

	Max Points	N Count	(Cronbach's Alpha)	
			School Score	Audit Score
Part II (CR)	6	584	0.56	0.57
Part III (CR)	9	584	0.62	0.65
Part IV (CR)	12	584	0.72	0.72
All CR Items	27	584	0.84	0.85

Mechanical Review Results

The mechanical review was conducted to check whether school scorers added up the total score for each section of the test and the total score of the test correctly. The following scores from the school scorers and SED mechanic reviewer's scores were compared for the following:

- Raw score for each of the four parts of the exam (MC and CR)
- Total CR score
- Total raw score
- Total scale score

As shown in Table 7, there was a high degree of exact agreement between the school scores and the mechanical review scores for multiple-choice items (95%). The exact agreement for each section of CR scores range from 60% to 73%. The percent of the CR scores that were within +/-1 for Part II, III, and IV are 96.9, 90.0, and 85.9, respectively.

At the total raw-score level, 34% of the school scores and the mechanical-review scores were exactly the same, 69% of the scores were within +/-1 raw-score point, and 86% of the scores were within +/-2 score points.

At the total scale-score level, 40% of the school scores and mechanical-review scores were exactly the same, 69% of the scores were within +/-1 raw-score point, and 84% of the scores were within +/-2 score points.

Table 7: Percentage of Score Difference Between School and Mechanical-Review Scores (School score minus mechanical-review score)

	Max Points	N Count	School Score Lower			Exact	School Score Higher		
			(≤-3)	(-2)	(-1)	0	(+1)	(+2)	(≥+3)
Part I Raw Score (MC)	60	584	0.5%	2.6%	0.0%	94.9%	0.0%	1.9%	0.2%
Part II Raw Score (CR)	6	583	0.2%	0.2%	4.3%	73.4%	19.2%	2.1%	0.7%
Part III Raw Score (CR)	9	584	0.0%	0.9%	7.0%	60.4%	22.6%	6.3%	2.7%
Part IV Raw Score (CR)	12	584	0.2%	2.6%	8.2%	61.8%	15.9%	7.5%	3.8%
Total CR Items	27	584	0.7%	3.6%	9.2%	36.0%	26.5%	11.3%	12.7%
Total Raw Score	87	584	1.2%	4.5%	9.9%	34.2%	25.3%	12.0%	12.8%
Total Scale Score	100	584	2.2%	3.3%	8.0%	39.7%	21.1%	12.2%	13.5%

Additional Analysis

Item analysis was performed on the nine CR items for the school scores and SED audit scores. The results are presented in Appendix B and Appendix C. School-level audit reports were generated and provided to the participating schools (see report template in Appendix D).

V. Summary

A total of 584 test papers from the August 2010 administration of the Regents Examination in Integrated Algebra from 67 schools were rescored by the Department's audit scorers during October of 2010. The audit sample was representative of the student population that took the August examination. Multiple methods were used to assess the reliability of the test and the inter-rater reliability of the constructed response items, including item raw score agreement, intraclass correlation, item mean score and standard deviation, total test mean score difference and correlation, and finally, internal consistency reliability.

A summary of the item level analysis indicated a relatively high level of agreement between the school scores and the SED audit scores, with a mean exact agreement rate of 85% for the nine CR items. The intraclass correlation between the school scores and the SED audit scores was at 0.80 or higher for all items. The school and audit item mean scores were comparable for a large majority of the items. The results suggested a high degree of inter-rated reliability and scoring consistency for the CR portion of the August examination.

At the total mean score level, the school scores correlated highly with the SED audit scores, with correlation coefficients ranging from 0.934 to 0.998. The internal consistency analysis of CR portion of the examination indicated a high degree of consistency for both the school scores and SED audit scores with the Cronbach's Alpha of 0.84 and 0.85, respectively.

The mechanical review of the total test score and subscores for the four parts of the examination showed a high level of accuracy, despite a few counting errors. The exact agreement for Part I (MC) was 94.9%. The exact agreement for each section of CR scores ranged from 60% to 73%. The percent of the CR scores that were within +/-1 for Part II, III, and IV were 96.9, 90.0, and 85.9, respectively.

In conclusion, the school scores of the CR items were very consistent with the SED audit scores. The total test score for the CR portion of the August examination was highly reliable for both the school scores and the SED audit scores.

Appendix A

2010 Department Review - Regents Examination in Integrated Algebra Organization Chart for Rescoring of Constructed Response Items								
			Student (within School)					
			1	2	3	…	…	10
Team	**Item #**	**Max. Credit**	Round 1 = Team 1, 2,…5 Round 2 (If necessary) = Alternate Team					
1	31 34 37 Total points	2 3 4 9	If necessary, papers exchanged with Group 4 for agreement					
2	32 35 38 Total points	2 3 4 9	If necessary, papers exchanged with Group 5 for agreement					
3	33 36 39 Total points	2 3 4 9	If necessary, papers exchanged with other groups for agreement					
4	31 34 37 Total points	2 3 4 9	If necessary, papers exchanged with Group 1 for agreement					
5	32 35 38 Total points	2 3 4 9	If necessary, papers exchanged with Group 2 for agreement					

Appendix B

Item Statistics Based on School Scores								
Item #	Max Points	Score Points (%)					Item Mean	Item-Total Correlation
		0	1	2	3	4		
31	2	57.7	23.1	19.2			0.61	0.553
32	2	44.9	8.9	46.2			1.01	0.579
33	2	73.9	13.4	12.7			0.39	0.682
34	3	52.9	19.5	14.7	12.8		0.88	0.656
35	3	37.5	17.6	19.9	25.0		1.32	0.674
36	3	78.8	9.1	3.6	8.6		0.42	0.675
37	4	59.8	16.1	11.5	6.0	6.7	0.84	0.760
38	4	45.7	14.6	23.1	9.9	6.7	1.17	0.666
39	4	78.8	6.0	5.5	2.2	7.5	0.54	0.764
Mean							**0.80**	**0.668**

Appendix C

Item Statistics Based on SED Audit Scores								
Item #	Max Points	Score Points (%)					Item Mean	Item-Total Correlation
		0	1	2	3	4		
31	2	64.9	23.3	11.8			0.47	0.523
32	2	44.3	9.2	46.4			1.02	0.595
33	2	79.6	8.4	12.0			0.32	0.710
34	3	64.0	14.0	12.3	9.6		0.67	0.657
35	3	43.0	15.2	22.4	19.3		1.18	0.679
36	3	80.8	6.2	4.8	8.2		0.40	0.705
37	4	68.7	13.4	7.2	5.3	5.5	0.66	0.769
38	4	53.1	7.0	25.2	8.6	6.2	1.08	0.681
39	4	81.0	5.3	3.4	2.2	8.0	0.51	0.768
Mean							**0.70**	**0.676**

The University of the State of New York
THE STATE EDUCATION DEPARTMENT
Albany, New York 12234

New York State Department Review Report
August 2010 Regents Examinations

Examination Title: Regents Examination in Integrated Algebra

	Part 1	Part 2			Part 3			Part 4				
Item Type	MC Total	Q31	Q32	Q33	Q34	Q35	Q36	Q37	Q38	Q39	CR Total	SS Total
		Constructed-Response (CR) Questions										
Number of Questions	30	1	1	1	1	1	1	1	1	1	9	NA
Maxmum Credits	60	2	2	2	3	3	3	4	4	4	27	100
Number of Papers Reviewed	10											
Number of Papers with School Score 3 or More Credits Lower than SED Score	0	NA	NA	NA	0	0	0	0	0	0	0	0
Number of Papers with School Score 2 Credits Lower than SED Score	0	0	0	0	0	0	0	0	0	0	0	0
Number of Papers with School Score 1 Credit Lower than SED Score	0	0	0	0	2	0	0	1	0	0	1	1
Number of Papers with Exact Agreement between School Score and SED Score	10	7	10	10	8	8	10	7	9	10	5	6
Number of Papers with School Score 1 Credit Higher than SED Score	0	3	0	0	0	2	0	2	0	0	2	3
Number of Papers with School Score 2 Credits Higher than SED Score	0	0	0	0	0	0	0	0	1	0	1	0
Number of Papers with School Score 3 or More Credits Higher than SED Score	0	NA	NA	NA	0	0	0	0	0	0	1	0
Total School Score Mean	27.4	0.8	1.2	0.2	0.8	1.7	0.3	0.6	2.0	0.4	8.0	68.3
Total SED Audit Score Mean	27.4	0.5	1.2	0.2	1.0	1.5	0.3	0.5	1.8	0.4	7.4	68.1

Abbreviation: MC = Multiple Choice, CR = Constructed Response, SS = Scale Score and SED = State Education Department