
**New York State Regents Examination
Department Review of January 2008
Regents Comprehensive Examination in English**

Technical Report

May 2009



**Office of State Assessment, P-16
New York State Education Department**

Copyright

Developed and published by the Office of State Assessment of New York State Education Department. Copyright © 2009 by the New York State Education Department. Any part of the publication may be reproduced or distributed with this notice in any form or by any means.

Table of Contents

SECTION I: INTRODUCTION.....	1
SECTION II: REVIEW PROCEDURE	2
Sample Collection and School Participation.....	2
Rescoring Procedure	3
SECTION III: DATA ANALYSIS.....	5
Data Preparation.....	5
Methods Used	5
SECTION IV: RESULTS.....	7
Item Mean and Standard Deviation	7
Inter-ratter Agreement	8
Intra-class Correlation.....	8
Total Score Correlation.....	9
Internal Consistency.....	9
SECTION V: SUMMARY	10

List of Tables

TABLE 1. NEED/RESOURCE CATEGORY (NRC) DEFINITIONS.....	2
TABLE 2. DISTRIBUTION OF EXAMINATION PAPERS	3
TABLE 3. NUMBER OF RECORDS RECEIVED	5
TABLE 4. CONSTRUCTED-RESPONSE ITEM ANALYSIS	7
TABLE 5. COMPARISON OF ITEM MEAN AND STANDARD DEVIATION.....	7
TABLE 6. INTER-RATER AGREEMENT BETWEEN STATE AND LOCAL	8
TABLE 7. PERCENTAGE OF SCORE DIFFERENCES.....	8
TABLE 8. INTRA-CLASS CORRELATION.....	9

Section I: Introduction

This report summarizes the results of a department review of the Regents Comprehensive Examination in English administered in January 2008. Department review is an internal audit process conducted by the New York State Education Department to ensure the reliability of the Regents assessment program. Each year, to ensure the reliability of local scoring of Regents examinations, the department conducts audits of New York State teachers' local scoring of a selected number of Regents examinations. In 2008, the Regents Comprehensive Examination in English administered in January 2008 was chosen for department review. Due to limited resources, the 2008 department review was limited to the rescoring of the constructed-response (CR) items only. Student test papers from a sample of schools from across the state were collected and responses were rescored by the State's independent scorers.

The purpose of the rescoring is to provide the necessary test reliability and inter-rater reliability evidence for the Regents Examinations. The audit process also 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 of their students' tests. Department review also acts as a deterrent to schools and teachers ensuring that they score tests properly in accordance with overall state 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.

Section II: Review Procedure

Sample Collection and School Participation

As soon as the January 2008 Regents Examinations were administered and scored by local schools, a stratified random sample of 119 high schools was selected for the department review. The school sample was stratified by Need/Resource Capacity Category to represent New York State school population (see Table 1).

Table 1: Need/Resource Category (NRC) Definitions

Need/Resource Category	Definition
New York City	New York City
Big 4 Cities	Buffalo, Rochester, Syracuse, Yonkers
High Need Urban/Suburban	Districts at or above 70 th percentile on the index with at least 100 students per square mile or enrollment greater than 2500
High Need Rural	All districts at or above the 70 th percentile with fewer than 50 students per square mile or enrollment of less than 2500
Average Need	All districts between the 20 th and 70 th percentiles on the index
Low Need	All districts below the 20 th percentile on the index
Charter Schools	Each charter school is a district

Of the 119 selected, 106 schools submitted original papers of student Regents Comprehensive Examination in English to the department. Upon receipt of the examination papers from the sample schools, a random sample of approximately 10% of the obtained examination papers from each school was selected for rescoring by an independent group of raters. The maximum number of student papers selected from an individual school was 30. For schools with ten or fewer student papers, all papers were selected for rescoring. A total of 908 student papers were rescored by state's raters. The distribution of the submitted student papers is presented in Table 2.

Table 2: Distribution of Examination Papers by Need/Resource Category

Need/Resource Category	N-Count	Percent
New York City	450	49.6
Big 4 Cities	10	1.1
High Need Urban/Suburban	40	4.4
High Need Rural	60	6.6
Average Need	190	20.9
Low Need	158	17.4
Total	908	100

Rescoring Procedure

The state rescoring was carried out by a group of current and recently retired New York State certified high school English teachers who are all highly experienced in scoring this examination. A total of 13 high school English teachers were recruited from across the State to conduct the rescoring. Efforts were made to recruit raters who represent teachers from across the State.

Four of these highly experienced subject specialists, one for each CR item, were appointed as table leaders to organize and supervise the scoring activities. They provided training to the raters, including a review of the scoring rubrics and the rating guide. The same hand scoring training materials developed for local scoring of the January 2008 Regents Comprehensive Examination in English were used to train the raters.

All four CR items in the examination were rescored during department review. The raters were divided into four groups. Each group was assigned one of the four CR items. Rater 1 in each group scored all the papers selected from a school for rescoring. This was a blind scoring. The table leader determined which papers required a second state score. Papers required a second score if Rater 1’s score was not within ½ point of the school’s score. Rater 2 then scored any papers that needed a second scoring. This was also a blind rating. If a third state score was required, the table leader served as the third rater. This third rating was required if Rater 1 and Rater 2 scores were more than 1 point apart and the school score was not between them. In that case, the department rescore was determined by the table leader.

At the conclusion of the department review, schools received one department Rescoring Record Sheet for each CR item rescored. The Record Sheet listed all the papers rescored for that CR item. This feedback to the schools helps schools implement appropriate changes in school procedures for rating future examinations if there were significant discrepancies between the schools' scores and the department's ratings.

Section III: Data Analysis

Data Preparation

An ACCESS database was designed for data entry. A four-digit school code and a two-digit student ID were built into a data entry form to record student scores. Both the final local school scores and the state rescores were entered and saved for data analysis. A total of 908 records were received. One record was deleted since it contained missing data on all of the constructed items.

Response data were obtained from two sources. Each student had one score from local scoring and one score from state scoring for each of the four constructed responses. Student local scores and state audit scores were matched by student ID number for data analysis. The matching local score and state rescore ranged from 533 to 892 across the four CR items (See Table 3). Only records with matching data for both local and state scoring were used in data analysis.

Table 3: Number of Records Received

	Number of Records	
	Local	State
Session I – Part A	903	533
Session I – Part B	885	678
Session II – Part A	900	855
Session II – Part B	895	892

Methods Used

Multiple methods were employed to assess the scoring reliability of the four CR items in the January 2008 Regents Comprehensive Examination in English. The following methods address the degree of agreement between local school scores and state rescores and the internal consistency of the CR component of the examination.

1. **Item Means and Standard Deviations:** Item raw score mean difference and standard deviation between the local school scores and state rescores were calculated as measures of average agreement/difference and variability between the two groups of scorers on a given item.
2. **Inter-rater Agreement:** Raw score agreement, as a measure of consensus between local school scorers and state rescorsers, was calculated for each item. In this method, the percentage of exact agreement (i.e. local scores match state rescores) and the percentage of adjacent and nonadjacent agreement (local scores and state scores differ in their score assignment by 1, 2, 3, or more score points) were calculated.
3. **Intra-class Correlation:** Intra-class 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. It was used to evaluate the inter-rater agreement between local school scores and state rescores.
4. **Total Score Correlation:** A local total score and state total score based on all four CR items were calculated. Correlation between the two total scores as calculated to provide an overall measure of the scoring reliability.
5. **Internal Consistency:** Internal consistency reliability (Cronbach Alpha) was calculated to provide another measure of the reliability of the CR component of the examination.

Section IV: Results

Item Mean and Standard Deviation

Item analysis was performed on all CR items based on both local school scores on the January 2008 Regents Comprehensive Examination in English. Table 4 presents the item analysis results for the January 2008 operational test. The item analysis results include maximum score points, total number of student counts, mean scores, and percent of students scoring at each score point.

Table 4: Constructed-response Item Analysis

Constructed-Response Item	Max Point	N-Count	P-Value	Percent of Students at Each Score Point							
				B	0	1	2	3	4	5	6
Session I – Part A	6	903	0.66	0.00	0.00	0.01	0.08	0.23	0.38	0.24	0.06
Session I – Part B	6	885	0.63	0.02	0.01	0.01	0.07	0.28	0.37	0.18	0.07
Session II – Part A	6	900	0.60	0.00	0.01	0.03	0.08	0.32	0.35	0.16	0.04
Session II – Part B	6	895	0.60	0.00	0.01	0.04	0.11	0.30	0.32	0.17	0.05

Table 5 presents the comparison of local and state raw score mean and standard deviation. Item mean and standard deviation are measures of average agreement/difference and variability between the two groups of scorers. The results show very close agreement between local and state item mean and standard deviation. Specifically, Session I – Part A has a mean difference of 0.2. Session I – Part B and Session II – Part B have a difference of 0.1. Session II – Part A has exactly the same mean raw scores. The differences in standard deviation between local and state scoring were minimal for all four items.

Table 5: Comparison of Item Mean and Standard Deviation

	N-Count	Mean			Standard Deviation		
		Local	State	Difference	Local	State	Difference
Session I – Part A	533	4.1	3.9	0.2	1.0	1.1	-0.1
Session I – Part B	678	4.0	3.9	0.1	1.0	1.0	0.0
Session II – Part A	855	3.8	3.8	0.0	1.1	1.0	0.1
Session II – Part B	892	3.8	3.7	0.1	1.1	1.1	0.0

Inter-rater Agreement

Inter-rater agreement was conducted to measure the difference between local scoring and state rescoring. The percentage of times local scores and state rescoring agreed and differed was calculated. Table 6 shows the exact agreement between local and state scores ranged from 76% to 82% and the adjacent agreement ranged from 17% to 21%. The total agreement for the four items were 97% or higher. Table 7 presents the percentage of score differences.

Table 6: Inter-rater Agreement between State and Local Scores

Item	Max Points	N-Count	Agreement (%)		
			Exact Agreement	Adjacent Agreement (+/- 1 Point)	Total Agreement
Session I – Part A	6	533	76	21	97
Session I – Part B	6	678	80	17	97
Session II – Part A	6	855	82	17	99
Session II – Part B	6	892	78	20	98

Table 7: Percentage of Score Differences

Item	Max Points	N-Count	Percentage of Score Difference (State Rescore minus Local Scoring)						
			- 3	- 2	- 1	0	1	2	3
Session I – Part A	6	533	0	3	18	76	3	0	0
Session I – Part B	6	678	1	1	12	80	5	1	0
Session II – Part A	6	855	0	1	11	82	6	0	0
Session II – Part B	6	892	0	2	10	78	10	1	0

Intra-class Correlation

The intra-class correlation was computed for each CR item. This correlation is an estimate of the reliability of scoring based on an average of the local and state scores. The intra-class correlation coefficients were 0.87 or higher (See Table 8). Consistent with other measures of inter-rater reliability provided in this study, these values indicate a very high level of scoring reliability.

Table 8: Intra-class Correlation

	N-Count		Intra-class Correlation Coefficient
	Local	State	
Session I – Part A	533	533	0.89
Session I – Part B	677	677	0.89
Session II – Part A	853	853	0.91
Session II – Part B	889	889	0.90

Total Score Correlation

As an overall measure of scoring reliability, the Pearson Correlation Coefficient between the local and state total CR scores was computed. This statistic is often used as an overall indicator of scoring reliability and generally ranges from 0.00 to near 1.00. The correlation coefficient between the local and state total CR scores was 0.89, which indicates a high degree of scoring reliability.

Internal Consistency

The Reliability Alpha, as a measure of internal consistency based on the average inter-item correlation, provides another score of reliability evidence. The internal consistency reliability of all CR items on the January 2008 Regents Comprehensive Examination in English was 0.95, indicating a very high degree of internal scoring consistency.

Section V: Summary

The department review is an internal audit process to ensure the scoring reliability of the New York State Regents examinations. In January 2008, the Regents Comprehensive Examination in English administered was chosen for department review. A sample of over one hundred schools submitted their January 2008 operational test papers to the department for rescoring by an independent group of New York State certified English teachers. This group of experienced high school English teachers was overseen by four highly experienced subject specialists. Due to limited resources, only the constructed-response (CR) items were rescored.

A total of 908 examination papers from 106 schools across New York State were rescored. Multiple statistical methods were employed to assess the scoring reliability of the four CR items on the January 2008 Regents Comprehensive Examination in English. A comparison based on average means and standard deviations showed a very close agreement between local scoring and state rescoring. The inter-rater agreement between local scoring and state rescoring also indicated a high degree of agreement. The exact agreement between local and state scoring ranges from 76% to 82% and total agreement was 97% to 99%. The intra-class correlation coefficients between the two sets of scores ranged from 0.89 to 0.91. For test items with a maximum score of 6, these values indicate a very high level of scoring reliability.

As an overall measure of scoring reliability, the Pearson Correlation Coefficient between the total local scores and total state CR scoring was .89, which also indicated a high degree of scoring reliability. The internal consistency reliability of the four CR items on January 2008 Regents Comprehensive Examination in English showed a Cronbach Alpha of 0.95, indicating a very high level of scoring consistency.

In general, the department review has found a high degree of agreement between local scoring and state rescoring. Analysis using multiple statistical methods indicates a high scoring reliability.