

**Inter-Rater Reliability Study of New York State
Grades 3-8 English Language Arts Test
2008 Test Administration**



Technical Report

**Prepared by:
Pearson
February 19, 2009**

Table of Contents

I.	Purpose and Scope of Audit	1
	Purpose.....	1
	Scope.....	1
II.	Selection of School Sample and Test Papers.....	3
	Audit Samples	3
	Stratified Sampling Design at the School Level.....	5
III.	Data Collection and School Participation	7
IV.	Selection and Training of Auditors	8
	Description of How the Auditors Were Selected.....	8
	Training of Auditors	8
	Steps Used to Support That Trainees Were Adequately Trained.....	8
	Quality Control Procedures	9
V.	Audit Procedures	10
	Description of the Audit Procedures.....	10
VI.	Data Analysis.....	11
	Item Means	11
	Intra-class Correlation	11
	Weighted Kappa.....	11
	Inter-rater Agreement.....	12
	Total Score Correlation	12
VII.	Results	13
	Item Means	13
	Percent of Agreement	13
	Intra-Class Correlations.....	13
	Weighted Kappa.....	15
	Inter-rater Agreement.....	15
	Total Score Correlation	17
	Additional Analyses.....	18
VIII.	Summary	19
	References	19
	Appendix A	20
	Appendix B	22
	Appendix C	24

Table of Contents Continued

Appendix D	31
Appendix E	39
Appendix F.....	41
Appendix G	48
Appendix H	55

List of Tables

Table 1:	Need/Resource Capacity Category Definitions	3
Table 2:	State N-counts.....	4
Table 3:	Target Proportions.....	4
Table 4:	Target N-counts.....	5
Table 5:	Selected N-counts	5
Table 6:	Sample Proportions	6
Table 7:	Obtained N-counts for English	7
Table 8:	Obtained Proportions for English	7
Table 9:	New York State Public Schools ELA Operational Test 2008: Inter-rater Agreement.....	14
Table 10:	Percentage of Raw Score Differences for English (Local Scoring Minus Audit Scoring).....	16
Table 11:	Correlations Between Local and Audit Scores	17
Table C-1:	New York State Public Schools Grade 3 ELA Operational Test 2008: Inter-rater Agreement.....	25
Table C-2:	New York State Public Schools Grade 4 ELA Operational Test 2008: Inter-rater Agreement	26
Table C-3:	New York State Public Schools Grade 5 ELA Operational Test 2008: Inter-rater Agreement.....	27
Table C-4:	New York State Public Schools Grade 6 ELA Operational Test 2008: Inter-rater Agreement.....	28
Table C-5:	New York State Public Schools Grade 7 ELA Operational Test 2008: Inter-rater Agreement.....	29
Table C-6:	New York State Public Schools Grade 8 ELA Operational Test 2008: Inter-rater Agreement.....	30
Table D-1:	New York State Public Schools (Without NYC) Grade 3 ELA Operational Test 2008: Inter-rater Agreement	32

List of Tables Continued

Table D-2:	New York State Public Schools (Without NYC) Grade 4 ELA Operational Test 2008: Inter-rater Agreement	33
Table D-3:	New York State Public Schools (Without NYC) Grade 5 ELA Operational Test 2008: Inter-rater Agreement	34
Table D-4:	New York State Public Schools (Without NYC) Grade 6 ELA Operational Test 2008: Inter-rater Agreement	35
Table D-5:	New York State Public Schools (Without NYC) Grade 7 ELA Operational Test 2008: Inter-rater Agreement	36
Table D-6:	New York State Public Schools (Without NYC) Grade 8 ELA Operational Test 2008: Inter-rater Agreement	38
Table E-1:	NYC Public Schools Grades 3 - 8 ELA Operational Test 2008: Inter-rater Agreement.....	40
Table F-1:	New York State Public Schools Grade 3 ELA Operational Test 2005: Proportions of Score Differences [Local Scoring minus Audit Scoring]	42
Table F-2:	New York State Public Schools Grade 4 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]	43
Table F-3:	New York State Public Schools Grade 5 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]	44
Table F-4:	New York State Public Schools Grade 6 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]	45
Table F-5:	New York State Public Schools Grade 7 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]	46
Table F-6:	New York State Public Schools Grade 8 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]	47
Table G-1:	New York State Public Schools Grade 3 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]	49

List of Tables Continued

Table G-2: New York State Public Schools Grade 4 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]	50
Table G-3: New York State Public Schools Grade 5 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]	51
Table G-4: New York State Public Schools Grade 6 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]	52
Table G-5: New York State Public Schools Grade 7 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]	53
Table G-6: New York State Public Schools Grade 8 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]	54
Table H-1: New York State Public Schools Grades 3 – 8 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring].....	56

I. Purpose and Scope of Audit

Purpose

The New York State Grades 3-8 English Language Arts (ELA) assessments consist of both multiple-choice (MC) and constructed-response (CR) items. The multiple-choice items are scored at the Regional Information Centers across the state and the constructed-response items are scored by teachers at the regional scoring centers or in their districts or schools. To ensure that teachers apply the same rigorous scoring standards as intended by the New York State Education Department (NYSED) and to provide evidence of inter-rater reliability, the Department conducts annual scoring audits that involve independent rescoring of five percent of all test papers after each test administration. This audit is conducted on a stratified random sample of schools, selected from each of the grade levels.

To help teachers in the scoring process, NYSED distributes training materials, sample student papers for various score points, and scoring rubrics. School districts provide in-service training to teachers through the use of scoring DVDs and scoring guides provided by NYSED. Combined with this training, teachers score student papers for each score point using scoring rubrics for the constructed-response questions; teachers have consistently done a very good job scoring the state assessment papers.

Schools identified for the 2008 audit were instructed to send their student assessments to Pearson for rescoring. Pearson is a professional scoring company known throughout the country for their quality scoring in large-scale state assessment programs. After Pearson completed the scoring, various statistical comparisons were made to evaluate the effectiveness and accuracy of the teacher scoring process. This report contains the results from those analyses.

Scope

The Grades 3-8 ELA assessments were administered in January 2008 throughout the state. The operational data for these assessments were collected by NYSED, including both MC and CR scores. The Regional Information Centers scored the MC items and New York teachers scored the CR items. In March 2008, Pearson conducted the audit study by rescoring the CR items from approximately five percent of all test papers. Pearson identified a stratified sample of schools (about 180 schools per grade) from across the state for each of the grade levels that contained approximately 15,000 student test papers. The 15,000 student assessments represented a 20% over-sampling, with the intention of attaining a minimum of 12,500 student assessments in each sample

for rescoring and data analyses. A total of 81,442 ELA test papers were collected from sample schools and rescored in March and April 2008.

Audit notification letters were sent to the sample schools in February 2008 and the selected schools sent their student test papers to Pearson for audit. Pearson re-scored the constructed-response questions and matched the audit scores with the local scores collected by NYSED. This process produced two sets of test scores for each student assessment. One set came from the local scoring performed by the New York teachers, and the second set came from the audit scoring performed by Pearson. The data analysis performed in this study consisted of various comparisons between the local scores and the audit scores.

II. Selection of School Sample and Test Papers

Audit Samples

To achieve the target audit sample of 12,500 test papers per grade level, approximately 15,000 test papers were sampled. Six stratified random samples of schools, with approximately 180 schools per grade, were selected, one for each grade, from all New York schools with Grades 3-8 enrollment to yield the target number of test papers. Each school was selected for audit at only one grade level. All selected schools were requested to send Pearson their ELA test papers for the grade level selected for audit.

Each audit sample was stratified by need/resource capacity category that consists of 7 categories. The need/resource capacity index, a measure of a district's ability to meet the needs of its students with local resources, is the ratio of the estimated poverty percentage to the combined wealth ratio. The need/resource capacity (N/RC) index divides districts into four categories: those with the highest need relative to resource capacity (High N/RC), those with average need relative to resource capacity (Average N/RC), those with less than average need relative to resource capacity (Low N/RC), and charter schools. The High N/RC districts are further subdivided into four groups (see Table 1 for definition).

Table 1. Need/Resource Capacity Category Definitions

Need/Resource Capacity Category		Definition
High N/RC Districts:	New York City	New York City
	Large Cities	Buffalo, Rochester, Syracuse, Yonkers
	Urban-Suburban	Districts at or above the 70 th percentile on the index with at least 100 students per square mile or enrollment greater than 2500
	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 N/RC Districts		All districts between the 20 th and 70 th percentiles on the index
Low N/RC Districts		All districts below the 20 th percentile on the index
Charter Schools		Each charter school is a district

The first step in the sampling procedure was to calculate the state n-counts within the seven N/R groups used for sampling. Based on school enrollment data provided by NYSED, the total number of students by grade was calculated for each need/resource category. Table 2 identifies the n-counts for each N/RC group by grade.

Table 2. State N-counts

State N-counts							
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Total
Total	193870	191849	196740	197195	204999	207143	1191796
New York City	66213	64402	65832	64118	67921	69082	397568
Large Cities	7893	7723	7664	7794	8486	8263	47823
High Need Urban/Suburban	16146	15689	15612	15474	16219	16083	95223
High Need Rural	11671	11490	11849	12278	13534	13500	74322
Average Need	59417	60176	61602	64005	66413	68036	379649
Low Need	30021	30283	31350	30831	30909	30994	184388
Charter	2509	2086	2831	2695	1517	1185	12823

Once the total n-counts were calculated by code for each grade level, the proportions represented by these n-counts were calculated within each cell. The following table contains those proportions.

Table 3. Target Proportions

Target Proportions							
N/RC Category	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Total
New York City	0.34	0.34	0.33	0.33	0.33	0.33	0.33
Large Cities	0.04	0.04	0.04	0.04	0.04	0.04	0.04
High Need Urban/Suburban	0.08	0.08	0.08	0.08	0.08	0.08	0.08
High Need Rural	0.06	0.06	0.06	0.06	0.07	0.07	0.06
Average Need	0.31	0.31	0.31	0.32	0.32	0.33	0.32
Low Need	0.15	0.16	0.16	0.16	0.15	0.15	0.15
Charter	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Finally, the number of students in each cell as determined by the target proportions was computed. These numbers are the product of the proportions in Table 3 and 15,000 which was the target sample size. This target sample size includes a 20% over-sampling to ensure a minimum sample of 12,500. The following table summarizes these n-counts.

Table 4. Target N-counts

Target N-counts per Sample							
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Total
New York City	5123	5059	5019	4877	4970	5002	30050
Large Cities	611	604	584	593	621	598	3611
High Need Urban/Suburban	1249	1227	1190	1177	1187	1165	7195
High Need Rural	903	898	903	934	990	978	5606
Average Need	4597	4705	4697	4869	4860	4927	28655
Low Need	2323	2368	2390	2345	2262	2244	13932
Charter	194	163	216	205	111	86	975
Totals	15000	15024	14999	15000	15001	15000	90024

Stratified Sampling Design at the School Level

Based on the target n-counts in Table 4, schools were randomly selected by grade within each N/RC group until the desired n-count was reached. Once a school was selected for a grade level, it was removed from the selection process. This process helped ensure that a school would not be audited at more than one grade level. Some school replacements were necessary so that target n-counts were met. Table 5 lists the resulting n-counts from the school sampling.

Table 5. Selected N-counts

Selected N-counts per Sample							
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Total
New York City	5,154	5,078	5,065	4,915	4,992	5,005	30209
Large Cities	620	631	573	610	620	710	3764
High Need Urban/Suburban	1259	1249	1224	1186	1205	1211	7334
High Need Rural	921	915	908	956	1002	976	5678
Average Need	4600	4731	4709	4949	4858	4965	28812
Low Need	2333	2373	2428	2345	2280	2292	14051
Charter	207	165	231	210	120	83	1016
Totals	15094	15142	15138	15171	15077	15242	90864

Table 6 shows the proportions within each cell based on the selected schools. A comparison between the proportions in Table 6 with the state proportions presented in Table 3 shows a very close match, thus demonstrating that the samples at each grade level are representative of New York's student population.

Table 6. Sample Proportions

Selected Sample Proportions							
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Total
New York City	0.34	0.34	0.33	0.32	0.33	0.33	0.33
Large Cities	0.04	0.04	0.04	0.04	0.04	0.05	0.04
High Need Urban/Suburban	0.08	0.08	0.08	0.08	0.08	0.08	0.08
High Need Rural	0.06	0.06	0.06	0.06	0.07	0.06	0.06
Average Need	0.30	0.31	0.31	0.33	0.32	0.33	0.32
Low Need	0.15	0.16	0.16	0.15	0.15	0.15	0.15
Charter	0.01	0.01	0.02	0.01	0.01	0.01	0.01

The schools identified in the above sampling scheme were contacted by Pearson and their test papers were used in the audit study.

III. Data Collection and School Participation

Pearson notified 807 schools and of those 732 schools returned materials. This represents a participation rate of 91%.

After the test papers were scored by Pearson the audit score file was combined with the local score file. Table 7 shows the actual n-counts in the final data files after all scoring and matching of data. Table 8 shows the actual proportions in the final data files after all scoring and matching of data.

Table 7. Obtained N-counts for English

Obtained N-counts							
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Totals
New York City	4624	5151	5060	4062	4019	4667	27583
Large Cities	557	597	530	527	517	721	3449
High Need Urban/Suburban	1154	1011	1099	1062	1117	1200	6643
High Need Rural	886	1005	794	1156	878	924	5643
Average Need	4148	4323	4375	3949	4553	4467	25815
Low Need	1852	2063	1888	1914	1791	2071	11579
Charter	94	165	217	197	0	57	730
Totals	13315	14315	13963	12867	12875	14107	81442

Table 8. Obtained Proportions for English

Obtained Proportions						
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
New York City	35%	36%	36%	32%	31%	33%
Large Cities	4%	4%	4%	4%	4%	5%
High Need Urban/Suburban	9%	7%	8%	8%	9%	9%
High Need Rural	7%	7%	6%	9%	7%	7%
Average Need	31%	30%	31%	31%	35%	32%
Low Need	14%	14%	14%	15%	14%	15%
Charter	1%	1%	2%	2%	0%	0%

A comparison between these proportions and the desired proportions in Table 3 shows that the data files used in each grade level closely match the intended demographics and were representative of the state. All samples differed less than 5% from the targets.

IV. Selection and Training of Auditors

Description of How the Auditors Were Selected

Scoring directors who led the audit were content experts with degrees in the subject area or a related area. Scoring directors were also chosen based on their experience in scoring the subject area. Prior to auditor training, scoring directors reviewed the training materials provided by NYSED. Scoring directors also reviewed the FAQs listed on the NYSED Web site and viewed NYSED-provided DVDs containing original training presentations.

Scoring Supervisors for the audit also 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. Further, ELA supervisors on Grades 4, 6, and 8 were required to demonstrate strong grammar skills.

Auditors possessed, at a minimum, a four-year college degree. They were placed on the most appropriate subject area based on their educational qualifications and their work or scoring experience. Auditors who demonstrated strong grammar skills in a grammar placement test qualified to assign mechanics scores to linked items.

The high quality of the auditors and high rate of return for auditors was due in part to the scoring sites' proximity to major universities and scoring sites' access to a large pool of college graduates.

Training of Auditors

Steps Used to Support That Trainees Were Adequately Trained

Supervisor training took place in Atlanta from March 24 - 26, 2008. 17 supervisors trained on all books and all grades for which they would score. 128 auditors began training on March 27, 2008. Auditors trained on items in a single book, completed scoring all books, and then trained on a new book for the next grade level.

Pearson staff used only those training materials supplied by the NYSED and used in the original scorer training. Scoring directors began training by reviewing and discussing the scoring guides for items in a book. Scoring directors then gave auditors the practice set(s) and auditors assigned scores to these sample responses. After auditors completed the set, scoring directors reviewed and explained expert scores for the practice papers. Subsequent practice sets for a book were trained in the same manner. If auditor performance or discussion of

the practice sets indicated a need for reviewing or retraining, it occurred at that time.

After discussion of the practice papers and any necessary review, auditors completed the consistency assurance set (CAS) for that book. A review and discussion of the scores occurred after auditors had assigned scores to all papers in the set. The scores achieved on the CAS determined if a trainee understood and could apply the scoring criteria. To qualify to remain on the project, a trainee had to demonstrate accuracy and consistency in scoring the CAS papers. Trainees who were unable to demonstrate accuracy and consistency in scoring were not allowed to score the project.

Quality Control Procedures

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 and began immediately. Backreading is a process in which supervisors check the scores of auditors immediately after they score a booklet. 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 scoring of the project. Supervisors increased backreading focus on auditors whose scoring accuracy, based on statistical reports or backreading records, was falling below expectations.
- **Second Scoring** begins immediately with 10% of responses in the audit receiving an independent score by a different auditor than the original. Second score papers are randomly generated by the system. By having a different auditor score the paper a second time without knowledge of the score given by the original auditor, it generates the inter-rater reliability statistics to verify the accuracy of the score.
- **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 auditors and for teams. Auditors whose statistics were not meeting quality expectations received retraining and had to demonstrate the ability to meet expectations in order to remain on the project.

V. Audit Procedures

Description of the Audit Procedures

Auditors were divided into two groups per grade. Each group scored either Book 1 or Book 2 for Grades 3, 5 and 7. One group scored Book 2 only for Grades 4, 6, and 8. The second group of auditors scored all of Book 3 and assigned a mechanics score to the linked items in Books 2 and 3 for Grades 4, 6, and 8.

Auditors recorded their scores onto scoring monitors. Scoring monitors are scannable tracking sheets that auditors grid the appropriate score for the booklet onto. Completed scoring monitors are then scanned at regular intervals throughout the day. After monitors were scanned, reports were generated for scoring directors to review and take appropriate action based on the reports (e.g., identifying auditors with low-quality statistics, identifying retraining needs).

In total, 21 ELA constructed-response items were rescored by the Pearson auditors.

VI. Data Analysis

For every test booklet used in the data analysis, there were two sets of scores. The first set of scores consisted of the multiple-choice and the constructed-response scores provided by the local scoring. The second set of scores consisted of the same multiple-choice scores and the audit scores for the constructed response items. All data analysis and comparisons were based on these two sets of scores for each test booklet.

Inter-rater reliability requires various statistics to evaluate. A single number never provides a complete picture of the reliability. Instead, one needs to examine inter-rater reliability from different aspects. To achieve that goal several analyses were performed. Item means were calculated to provide a measure of the average agreement between the local and audit scoring. An intra-class correlation was computed between the local and audit scoring which provides an estimate of the reliability of the scoring. A weighted Kappa statistics was computed to quantify the level of agreement between the categorical data provided by the local and audit scoring. Inter-rater agreement was evaluated by examining the consensus between the local and audit scoring using percent of agreement. Finally, the correlation between the total scores resulting from the local and audit scoring is computed. This provides an overall evaluation of the scoring reliability.

Item Means

The average score for each constructed-response question was computed based on the local scoring and the audit scoring. Differences between the two scores were also computed. Item means for the multiple-choice items were not examined because the same item responses were used for both the local scoring and the audit scoring.

Intra-class Correlation

The mean intra-class correlation was computed for each item. This correlation estimates the reliability of the scoring based on an average of the local and audit scores.

Weighted Kappa

The weighted Kappa (Cohen, 1968) was calculated for each item based on the local and audit scoring. This statistic produces an estimate of the reliability of the

score classifications. Weighted Kappa is a measure of quantifying levels of agreement for categorical data, item scores in the case of this study. When raters tend to assign some scores more frequently than others, the agreement rates are affected. By using the weighted Kappa, larger differences between raters are given smaller weights, therefore this statistic can differ from the inter-rater agreement measure for certain items. In this study, lower scores were more frequently assigned than the higher scores; therefore, this statistic was evaluated only as one of the many pieces of evidence supporting the reliability of the state and school scores.

Inter-rater Agreement

For each constructed-response question, the difference between the local score and the audit score was computed and tallied. The total of the constructed-response items was also computed and the difference between the local scoring and audit scoring results were computed. The number of times the various differences occurred was counted and the proportions were calculated.

Two total scores were computed for each test booklet using the local scoring and audit scoring results. The correlation between these scores was also computed.

Total Score Correlation

For both the local and audit scoring results, a total score on the complete assessment was computed. Then the correlation between these total scores was computed. This statistic provides an overall measure of the scoring reliability. The amount of variance of the total scores that is shared by the local and audit scoring is obtained by squaring the correlation.

VII. Results

Item Means

The average score for each constructed-response question was computed based on the local scoring and the audit scoring as well as the standard deviation. The results from this analysis are presented in Table 9. They show a very close agreement between the local scoring and the audit scoring on the English constructed-response questions. 9 out of 21 items have exactly the same mean raw scores and an additional 10 items have a mean difference of 0.1.

Percent of Agreement

Table 9 contains the percent of agreement and percent of approximate agreement. Percent of approximate agreement pertains to scores where the local and audit scoring differed by only one score point.

When interpreting these statistics it is important to note the impact of the maximum points possible for a given item. That is, it is more likely that the two sets of scores will have exact agreement if there are only 2 maximum points versus an item with 5 maximum points. The total percent of agreement is the sum of the exact agreement and the approximate agreement i.e., ratings that differ by one point. This statistic is greatly influenced by the maximum points possible. Taken collectively, the total percent of agreement ranges from a low of 89.2% to a high of 99.8%. Consistent with the information in the item means, the percent of agreement shows a high level of agreement between the local and audit scoring.

Intra-Class Correlations

The Intra-Class Correlation (ICC) assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects. The mean intra-class correlation estimates the reliability of the scoring based on an average of the local and audit scores.

Generally, correlations greater than 0.60 are considered very strong because they explain more than one-third of the variance. Table 9 shows that all of the items had correlations greater than 0.60. One-third of the items had correlations greater than 0.80. The intra-class correlations range from 0.64 to 0.96. The intra-class correlations show a high degree of consistency between the local and audit scores.

Table 9. New York State Public Schools ELA Operational Test 2008: Inter-rater Agreement

Grade	Item #	Scoring Model	Score Points	Total N	Agreement (%)			RS Mean		RS SD		Intra-Class Correlation	Weighted Kappa
					Exact	Approx.	Total	Local	Audit	Local	Audit		
3	21	Overall	2	12985	96.5	3.2	99.8	1.9	1.9	0.43	0.44	0.94	0.85
	26	Overall	2	12985	81.2	18.1	99.4	1.6	1.6	0.61	0.64	0.85	0.67
	27	Overall	2	12985	84.6	14.1	98.7	1.1	1.0	0.87	0.88	0.93	0.82
	28	Overall	3	12985	94.7	4.4	99.2	2.7	2.7	0.77	0.75	0.96	0.89
4	29-31	Overall	4	14139	56.1	41.4	97.5	2.8	2.8	0.84	0.77	0.75	0.45
	32-35	Overall	4	14139	51.4	44.7	96.1	2.7	2.8	0.90	0.94	0.78	0.47
	31&35	Overall	3	14139	57.8	40.3	98.1	2.1	2.2	0.75	0.70	0.70	0.42
5	21	Overall	2	13698	74.8	23.5	98.3	1.2	1.1	0.64	0.68	0.79	0.61
	26	Overall	2	13698	83.1	16.4	99.5	1.6	1.6	0.55	0.60	0.84	0.68
	27	Overall	3	13698	69.6	27.7	97.4	1.3	1.3	1.07	1.06	0.91	0.72
6	27-30	Overall	5	12767	43.4	47.8	91.2	3.2	2.9	1.11	0.94	0.75	0.42
	31-34	Overall	5	12767	43.8	47.1	90.9	3.2	3.1	1.11	1.09	0.78	0.45
	30&34	Overall	3	12767	58.7	39.8	98.5	2.2	2.1	0.72	0.72	0.72	0.43
7	27	Overall	2	12429	64.5	34.5	99.0	1.4	1.4	0.67	0.65	0.71	0.46
	28	Overall	2	12429	64.4	34.0	98.4	1.3	1.2	0.73	0.71	0.76	0.51
	33	Overall	2	12429	86.5	13.0	99.5	1.8	1.8	0.46	0.44	0.77	0.56
	34	Overall	2	12429	68.4	30.2	98.6	1.6	1.6	0.55	0.61	0.64	0.40
	35	Overall	3	12429	70.8	26.8	97.6	1.0	0.9	0.94	0.93	0.88	0.69
8	27-30	Overall	5	13875	40.6	48.5	89.2	3.5	3.1	1.15	1.06	0.76	0.43
	31-34	Overall	5	13875	44.4	47.6	91.9	3.8	3.7	1.07	0.98	0.76	0.42
	30&34	Overall	3	13875	60.2	38.7	98.8	2.3	2.2	0.70	0.68	0.71	0.43

Approximate agreement (%) is the percent of pairs of readers that differ by one score point.

Weighted Kappa

The weighted Kappa is an estimate of the reliability of the score classifications. That is, the Kappa statistic is a measure of reproducibility for categorical data. A common stumbling block in evaluating scoring reliability or consistency is the basic concept of agreement beyond chance and, in turn, the importance of correcting for chance agreement. The Kappa statistic corrects for this chance agreement and tells us how much of the possible agreement over and above chance the scorers have achieved.

Guidelines for the evaluation of Kappa are:

- $k > .75$ denotes excellent reproducibility
- $0.4 \leq k \leq .75$ denotes good reproducibility
- $0 \leq k \leq 0.4$ denotes marginal reproducibility

The results found in Table 9 show a high degree of consistency between the local and audit scoring. Item 34 in Grade 7 produced a weighted Kappa statistic of 0.40 which is right at the margin for the category of “good reproducibility.” All other items had weighted Kappa statistics denoted good or excellent reproducibility.

Inter-rater Agreement

For each constructed-response question, the difference between the local score and the audit score was computed and tallied. The total of the constructed-response items was also computed and the difference between the local scoring and audit scoring results were computed. The absolute value of the differences between the local scores and the audit scores were then tallied and the proportions computed. Those proportions are presented in Table 10. Appendices F through H contain the proportion of actual differences instead of the absolute values.

Table 10. Percentage of Raw Score Differences for English (Local Scoring Minus Audit Scoring)

Grade	Item	MAX Points	Difference				
			0	1	2	3	4 or more
3 N=12985	21	2	97%	3%	0%		
	26	2	81%	18%	0%		
	27	2	85%	14%	1%		
	28	3	95%	4%	1%	0%	
4 N=14139	29-31	4	56%	41%	2%	0%	0%
	32-35	4	51%	45%	4%	0%	0%
	31&35	3	58%	40%	2%	0%	
5 N=13698	21	2	75%	23%	2%		
	26	2	83%	16%	0%		
	27	3	70%	28%	3%	0%	
6 N=12767	27-30	5	43%	48%	8%	0%	0%
	31-34	5	44%	47%	9%	0%	0%
	30&34	3	59%	40%	2%	0%	
7 N=12429	27	2	65%	34%	1%		
	28	2	64%	34%	2%		
	33	2	87%	13%	0%		
	34	2	68%	30%	1%		
	35	3	71%	27%	2%	0%	
8 N=13875	27-30	5	41%	49%	10%	1%	0%
	31-34	5	44%	48%	8%	0%	0%
	30&34	3	60%	39%	1%		

The information provided in Table 10 shows a high degree of consistency between the local and audit scoring. Specifically, the percentage of ratings that were exactly the same across local and audit scoring met or exceeded 70% for all items in Grades 3 and 5. For Grades 4, 6, and 8, the percent perfect agreement was lower, though most agreement was within one score point. A possible explanation for such observation might be because the maximum score points for items in Grades 4, 6, and 8 were relatively higher than the maximum score points for items in other grades, under which case agreement is relatively harder to achieve. Grade 7 had two items above 70% and three below with very few differences greater than one. The percent of scores that differed by two or more points fell below 5% for all items, except for the items with maximum score points of 5.

Total Score Correlation

For both the local and audit scoring results, total scores on the entire assessment and the open-ended questions only were computed. Then the correlation between the local and audit total scores was computed. This statistic provides an overall measure of the scoring reliability. The amount of variance of the total scores that is shared by the local and audit scoring is obtained by squaring the correlation. This statistic is an indication of the consistency between the local scoring and audit scoring on the total test score level.

Table 11. Correlations Between Local and Audit Scores

Grade	Total Score		Total of Open-ended Questions Only	
	Correlation	Common Variance	Correlation	Common Variance
3	0.99	0.98	0.91	0.83
4	0.98	0.96	0.76	0.58
5	0.98	0.96	0.84	0.71
6	0.97	0.94	0.78	0.61
7	0.98	0.96	0.79	0.62
8	0.97	0.94	0.78	0.61

The correlations show a very high degree of consistency between the local and audit scoring results with correlations ranging from 0.97 to 0.99. Based on these correlations, the amount of common variance between local and audit scoring ranges from 0.94 to 0.98, which means that differences in CR scores between the local and audit scoring results did not impact the total score level much. Given that most decisions using test results are based on the total score, these statistics provide valuable evidence of the reliability and consistency in students' total scores across local and audit scoring methods. The correlations based on the open-ended questions range from a low of 0.76 to 0.91, and the common variance ranges from 0.58 to 0.83. This again shows a high degree of agreement between local and audit scoring.

Additional Analyses

The results from additional analyses are presented in the appendices. Appendix A contains a detailed item analysis for the English constructed-response items resulting from the local scoring. These tables show the proportion of students obtaining each of the possible score points for each item. The tables also provide the item mean and point-biserial (PBS).

The same item analysis for the English audit scores are in Appendix B.

Appendices C, D, and E contain summary item-level information for the English assessments. Analyses are computed for all schools and then by scoring model. The scoring models are:

1. Regional scoring
2. Schools from two districts
3. Three or more schools within a district
4. Two schools within a district
5. Only one school

The appendices are for:

1. All schools in the state,
2. All schools without the New York City schools, and
3. New York City schools only.

These tables summarize the following item-level information:

- Maximum score points
- Exact agreement
- Approximate agreement
- Item mean and standard deviation from Audit and Local Scoring
- Intra-class correlation
- Kappa statistic

Appendices F, G, and H contain the distribution of differences at the item level between the Audit scoring and the Local scoring for English. This information was computed for the various scoring models. The appendices are for:

1. All schools in the state,
2. All schools without the New York City schools, and
3. New York City schools only.

VIII. Summary

The sample acquisition was very successful. A comparison between the obtained proportions with the state proportions found in Tables 3 and 8 show that the samples mirrored the State in these categories. For all grades the obtained proportions in each of the 7 Need/Resource Capacity categories were virtually identical to the state proportions. As a result, the analysis performed in the study is based on data which is representative of the State demographics.

A summary of the analyses performed in this study indicates that the local scoring results were very close to the audit scoring results. Correlations between the total scores resulting from the audit scoring and the local scoring range from a low of 0.97 to a high of 0.99. The correlations based on the open-ended only range from 0.76 to 0.91. These correlations indicate a high degree of agreement between local and audit scoring results.

Examination of the differences between local scoring and audit scoring at the item level also shows a high degree of consistency. In English the largest mean difference between local and audit scoring was 0.4, which occurred in Grade 8, item 27. Considering this is a 5-point item, that difference only represents 8% of the maximum points. All other items had mean differences of 0.3 or less.

Appendix C contains the scoring results for each of the scoring models. By inspection it appears that there is little difference between the local and audit scoring results by scoring model. The largest differences occurred in grade 6, item 27, scoring model 2, where differences reached 0.8 in magnitude. Also in grade 6, item 30, again scoring model 2, the difference reached 0.7. However, these differences are based on a very small number of papers (N=31). The remaining differences were all less than or equal to 0.5, with the vast majority at 0.1 or less. This shows a high degree of consistency not only between the local and audit scoring, but also across scoring models.

In conclusion, the local scoring results are very consistent with the audit scoring. No major discrepancies were found in these analyses.

References

Cohen J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*. 70:213-20, 1968.

Appendix A

English Item Analysis for Local Scoring

Local Scoring ELA Grade 3 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	PBS
21	O	0	0.04	0.06	0.91	0	0	1.87	0.47
26	O	0	0.06	0.27	0.66	0	0	1.6	0.47
27	O	0	0.34	0.24	0.41	0	0	1.07	0.52
28	O	0	0.05	0.04	0.11	0.81	0	2.67	0.5

Local Scoring ELA Grade 4 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	PBS
29-31	O	0	0	0.05	0.28	0.46	0.21	2.81	0.65
32-35	O	0	0.01	0.08	0.3	0.44	0.17	2.68	0.7
31&35	O	0	0.02	0.19	0.48	0.31	0	2.09	0.62

Local Scoring ELA Grade 5 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	PBS
21	O	0	0.13	0.55	0.31	0	0	1.18	0.55
26	O	0	0.04	0.3	0.66	0	0	1.63	0.44
27	O	0	0.31	0.26	0.27	0.16	0	1.28	0.71

Local Scoring ELA Grade 6 Item Statistics

Item	Key	B	0	1	2	3	4	5	Mean	PBS
27-30	O	0	0.01	0.07	0.19	0.34	0.28	0.12	3.17	0.74
31-34	O	0	0.01	0.06	0.18	0.33	0.3	0.12	3.2	0.76
30&34	O	0	0.01	0.15	0.49	0.35	0	0	2.17	0.64

Local Scoring ELA Grade 7 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	PBS
27	O	0	0.09	0.36	0.54	0	0	1.44	0.56
28	O	0.02	0.15	0.39	0.44	0	0	1.27	0.56
33	O	0	0.02	0.16	0.82	0	0	1.79	0.41
34	O	0	0.03	0.31	0.66	0	0	1.62	0.41
35	O	0	0.4	0.29	0.25	0.06	0	0.96	0.64

Local Scoring ELA Grade 8 Item Statistics

Item	Key	B	0	1	2	3	4	5	Mean	PBS
27-30	O	0	0.01	0.05	0.14	0.28	0.31	0.22	3.5	0.77
31-34	O	0	0	0.03	0.09	0.25	0.33	0.29	3.76	0.75
30&34	O	0	0.01	0.11	0.44	0.44	0	0	2.32	0.64

Appendix B

English Item Analysis for Audit Scoring

Audit Scoring ELA Grade 3 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	PBS
21	O	0	0.04	0.06	0.9	0	0	1.86	0.48
26	O	0	0.08	0.29	0.63	0	0	1.55	0.46
27	O	0	0.36	0.23	0.41	0	0	1.04	0.52
28	O	0	0.05	0.04	0.1	0.82	0	2.69	0.49

Audit Scoring ELA Grade 4 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	PBS
29-31	O	0	0	0.03	0.28	0.5	0.19	2.84	0.58
32-35	O	0	0.01	0.08	0.24	0.44	0.23	2.8	0.67
31&35	O	0	0.02	0.13	0.53	0.33	0	2.17	0.59

Audit Scoring ELA Grade 5 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	PBS
21	O	0	0.19	0.53	0.28	0	0	1.09	0.54
26	O	0	0.06	0.3	0.64	0	0	1.58	0.45
27	O	0	0.29	0.27	0.27	0.17	0	1.32	0.7

Audit Scoring ELA Grade 6 Item Statistics

Item	Key	B	0	1	2	3	4	5	Mean	PBS
27-30	O	0	0.01	0.06	0.21	0.48	0.2	0.04	2.94	0.67
31-34	O	0	0.01	0.06	0.2	0.36	0.28	0.09	3.1	0.74
30&34	O	0	0.01	0.17	0.49	0.33	0	0	2.14	0.64

Audit Scoring ELA Grade 7 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	PBS
27	O	0	0.09	0.42	0.49	0	0	1.4	0.54
28	O	0.02	0.15	0.44	0.39	0	0	1.22	0.53
33	O	0	0.03	0.1	0.87	0	0	1.84	0.38
34	O	0	0.07	0.3	0.64	0	0	1.57	0.39
35	O	0	0.4	0.29	0.25	0.05	0	0.95	0.64

Audit Scoring ELA Grade 8 Item Statistics

Item	Key	B	0	1	2	3	4	5	Mean	PBS
27-30	O	0	0.01	0.06	0.17	0.39	0.28	0.08	3.1	0.74
31-34	O	0	0	0.03	0.08	0.25	0.43	0.21	3.7	0.68
30&34	O	0	0.01	0.12	0.53	0.34	0	0	2.2	0.64

Appendix C

Item Level Statistics for English Including All Schools in State

Table C- 1. New York State Public Schools Grade 3 ELA Operational Test 2008: Inter-rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			RS Mean		RS SD		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
21	Overall	2	12985	96.5	3.2	99.8	1.9	1.9	0.43	0.44	0.94	0.85
	1	2	6657	96.2	3.5	99.7	1.8	1.8	0.47	0.49	0.95	0.86
	2	2	221	98.2	1.8	100.0	2.0	1.9	0.27	0.29	0.94	0.83
	3	2	3689	97.0	2.7	99.7	1.9	1.9	0.39	0.40	0.93	0.84
	4	2	832	97.1	2.9	100.0	1.9	1.9	0.38	0.40	0.95	0.85
	5	2	1586	96.3	3.5	99.7	1.9	1.9	0.39	0.39	0.92	0.81
26	Overall	2	12985	81.2	18.1	99.4	1.6	1.6	0.61	0.64	0.85	0.67
	1	2	6657	80.1	19.1	99.2	1.5	1.5	0.64	0.66	0.85	0.67
	2	2	221	79.6	20.4	100.0	1.7	1.7	0.52	0.52	0.77	0.56
	3	2	3689	82.6	16.9	99.5	1.7	1.6	0.56	0.61	0.84	0.67
	4	2	832	82.8	16.9	99.8	1.6	1.6	0.58	0.62	0.86	0.69
	5	2	1586	82.2	17.3	99.6	1.7	1.6	0.58	0.60	0.84	0.66
27	Overall	2	12985	84.6	14.1	98.7	1.1	1.0	0.87	0.88	0.93	0.82
	1	2	6657	81.9	16.2	98.1	1.0	1.0	0.86	0.88	0.92	0.79
	2	2	221	82.8	16.3	99.1	1.3	1.2	0.81	0.87	0.92	0.80
	3	2	3689	86.8	12.5	99.3	1.1	1.1	0.88	0.88	0.95	0.85
	4	2	832	89.1	10.7	99.8	1.2	1.2	0.86	0.87	0.96	0.88
	5	2	1586	88.7	10.7	99.4	1.1	1.1	0.88	0.87	0.95	0.87
28	Overall	3	12985	94.7	4.4	99.2	2.7	2.7	0.77	0.75	0.96	0.89
	1	3	6657	93.9	5.1	99.1	2.6	2.7	0.82	0.80	0.96	0.88
	2	3	221	98.2	1.8	100.0	2.8	2.8	0.52	0.50	0.98	0.94
	3	3	3689	95.9	3.4	99.3	2.7	2.7	0.72	0.71	0.97	0.90
	4	3	832	96.9	2.8	99.6	2.7	2.7	0.71	0.69	0.98	0.93
	5	3	1586	93.6	5.2	98.8	2.7	2.7	0.69	0.69	0.94	0.84

Approximate agreement (%) is the percent of pairs of readers that differ by one score point.

Total agreement (%) is the sum of exact and approximate percents.

The scoring models are: 1) Regional scoring; 2) Schools from two districts; 3) Three or more schools within a district; 4) Two schools within a district; and 5) Only one school.

Table C- 2. New York State Public Schools Grade 4 ELA Operational Test 2008: Inter-rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			RS Mean		RS SD		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
29-31	Overall	4	14139	56.1	41.4	97.5	2.8	2.8	0.84	0.77	0.75	0.45
	1	4	8064	54.6	42.3	97.0	2.7	2.8	0.86	0.78	0.75	0.45
	2	4	55	54.5	41.8	96.4	2.5	2.8	0.66	0.73	0.63	0.35
	3	4	4256	57.4	40.5	97.9	3.0	2.9	0.78	0.76	0.74	0.45
	4	4	465	57.4	41.7	99.1	3.0	2.8	0.71	0.68	0.70	0.40
	5	4	1299	61.0	37.9	98.8	2.9	2.8	0.79	0.72	0.77	0.50
32-35	Overall	4	14139	51.4	44.7	96.1	2.7	2.8	0.90	0.94	0.78	0.47
	1	4	8064	49.3	46.4	95.6	2.6	2.7	0.92	0.96	0.78	0.46
	2	4	55	49.1	45.5	94.5	2.5	2.8	0.78	1.07	0.78	0.45
	3	4	4256	54.1	42.4	96.5	2.8	2.9	0.83	0.90	0.76	0.46
	4	4	465	51.0	45.8	96.8	2.9	2.8	0.81	0.84	0.73	0.41
	5	4	1299	56.1	41.7	97.8	2.8	2.8	0.88	0.92	0.82	0.52
31&35	Overall	3	14139	57.8	40.3	98.1	2.1	2.2	0.75	0.70	0.70	0.42
	1	3	8064	56.7	41.2	97.9	2.1	2.1	0.77	0.70	0.70	0.41
	2	3	55	47.3	47.3	94.5	1.8	2.1	0.74	1.00	0.73	0.42
	3	3	4256	58.7	39.5	98.2	2.1	2.2	0.73	0.70	0.70	0.42
	4	3	465	61.7	37.8	99.6	2.1	2.2	0.70	0.61	0.70	0.42
	5	3	1299	60.7	37.5	98.2	2.1	2.2	0.72	0.71	0.72	0.45

Approximate agreement (%) is the percent of pairs of readers that differ by one score point.

Total agreement (%) is the sum of exact and approximate percents.

The scoring models are: 1) Regional scoring; 2) Schools from two districts; 3) Three or more schools within a district; 4) Two schools within a district; and 5) Only one school.

Table C- 3. New York State Public Schools Grade 5 ELA Operational Test 2008: Inter-rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			RS Mean		RS SD		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
21	Overall	2	13698	74.8	23.5	98.3	1.2	1.1	0.64	0.68	0.79	0.61
	1	2	6763	73.2	25.2	98.4	1.1	1.1	0.67	0.71	0.80	0.61
	2	2	353	70.0	27.2	97.2	1.1	1.0	0.67	0.66	0.72	0.53
	3	2	4090	76.5	21.6	98.1	1.2	1.1	0.61	0.65	0.78	0.61
	4	2	1309	80.7	18.4	99.2	1.3	1.2	0.58	0.63	0.83	0.67
	5	2	1183	72.7	25.4	98.1	1.2	1.1	0.63	0.64	0.75	0.56
26	Overall	2	13698	83.1	16.4	99.5	1.6	1.6	0.55	0.60	0.84	0.68
	1	2	6763	81.5	17.9	99.4	1.5	1.5	0.59	0.64	0.85	0.68
	2	2	353	82.4	17.3	99.7	1.7	1.7	0.49	0.53	0.78	0.60
	3	2	4090	84.8	14.7	99.5	1.7	1.6	0.50	0.56	0.83	0.67
	4	2	1309	84.6	15.3	99.9	1.7	1.7	0.49	0.51	0.82	0.65
	5	2	1183	85.3	14.2	99.5	1.7	1.6	0.51	0.58	0.85	0.70
27	Overall	3	13698	69.6	27.7	97.4	1.3	1.3	1.07	1.06	0.91	0.72
	1	3	6763	69.6	27.8	97.4	1.1	1.1	1.03	1.03	0.90	0.71
	2	3	353	75.6	23.8	99.4	1.3	1.3	1.05	1.06	0.94	0.79
	3	3	4090	69.7	27.5	97.1	1.5	1.5	1.08	1.06	0.91	0.72
	4	3	1309	68.8	28.3	97.2	1.6	1.6	1.03	1.04	0.89	0.70
	5	3	1183	68.6	28.8	97.5	1.4	1.5	1.05	1.02	0.90	0.70

Approximate agreement (%) is the percent of pairs of readers that differ by one score point.

Total agreement (%) is the sum of exact and approximate percents.

The scoring models are: 1) Regional scoring; 2) Schools from two districts; 3) Three or more schools within a district; 4) Two schools within a district; and 5) Only one school.

Table C- 4. New York State Public Schools Grade 6 ELA Operational Test 2008: Inter-rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			RS Mean		RS SD		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
27-30	Overall	5	12767	43.4	47.8	91.2	3.2	2.9	1.11	0.94	0.75	0.42
	1	5	6413	45.3	47.1	92.4	2.9	2.8	1.13	0.96	0.78	0.45
	2	5	31	29.0	41.9	71.0	3.5	2.7	1.16	0.64	0.43	0.23
	3	5	1673	38.1	50.0	88.1	3.2	2.9	1.08	0.88	0.66	0.32
	4	5	1375	41.9	49.4	91.3	3.5	3.1	1.12	0.91	0.75	0.42
	5	5	3275	43.1	47.5	90.7	3.5	3.1	0.97	0.90	0.68	0.36
31-34	Overall	5	12767	43.8	47.1	90.9	3.2	3.1	1.11	1.09	0.78	0.45
	1	5	6413	44.5	46.4	90.9	3.0	3.0	1.12	1.08	0.79	0.46
	2	5	31	45.2	45.2	90.3	3.2	2.9	1.15	1.00	0.74	0.44
	3	5	1673	39.9	50.3	90.2	3.3	3.2	1.10	1.06	0.75	0.40
	4	5	1375	44.9	46.7	91.6	3.5	3.3	1.08	1.12	0.79	0.47
	5	5	3275	44.0	46.9	90.9	3.4	3.2	1.04	1.06	0.76	0.43
30&34	Overall	3	12767	58.7	39.8	98.5	2.2	2.1	0.72	0.72	0.72	0.43
	1	3	6413	56.9	41.1	98.1	2.1	2.1	0.74	0.73	0.71	0.41
	2	3	31	38.7	51.6	90.3	2.4	1.7	0.65	0.52	0.03	0.10
	3	3	1673	59.1	39.6	98.7	2.2	2.2	0.67	0.70	0.69	0.41
	4	3	1375	63.6	35.7	99.3	2.4	2.3	0.68	0.74	0.76	0.50
	5	3	3275	60.1	38.7	98.8	2.3	2.2	0.69	0.70	0.71	0.43

Approximate agreement (%) is the percent of pairs of readers that differ by one score point.

Total agreement (%) is the sum of exact and approximate percents.

The scoring models are: 1) Regional scoring; 2) Schools from two districts; 3) Three or more schools within a district; 4) Two schools within a district; and 5) Only one school.

Table C- 5. New York State Public Schools Grade 7 ELA Operational Test 2008: Inter-rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			RS Mean		RS SD		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
27	Overall	2	12429	64.5	34.5	99.0	1.4	1.4	0.67	0.65	0.71	0.46
	1	2	6235	61.6	37.0	98.6	1.4	1.3	0.68	0.67	0.70	0.44
	2	2	321	68.8	30.8	99.7	1.5	1.7	0.59	0.52	0.66	0.42
	3	2	192	59.9	38.5	98.4	1.0	1.2	0.77	0.66	0.73	0.46
	4	2	737	67.6	31.9	99.5	1.5	1.4	0.61	0.64	0.72	0.48
	5	2	4944	67.7	31.7	99.3	1.5	1.5	0.64	0.61	0.72	0.47
28	Overall	2	12429	64.4	34.0	98.4	1.3	1.2	0.73	0.71	0.76	0.51
	1	2	6235	63.9	34.3	98.3	1.2	1.2	0.74	0.72	0.76	0.51
	2	2	321	67.6	31.8	99.4	1.3	1.4	0.71	0.67	0.78	0.54
	3	2	192	59.4	39.6	99.0	0.7	0.9	0.71	0.69	0.72	0.45
	4	2	737	68.8	30.0	98.8	1.4	1.2	0.67	0.71	0.78	0.56
	5	2	4944	64.2	34.2	98.4	1.3	1.3	0.72	0.70	0.75	0.50
33	Overall	2	12429	86.5	13.0	99.5	1.8	1.8	0.46	0.44	0.77	0.56
	1	2	6235	85.3	14.2	99.5	1.8	1.8	0.49	0.47	0.79	0.57
	2	2	321	88.8	10.3	99.1	1.8	1.9	0.42	0.34	0.69	0.48
	3	2	192	80.7	18.2	99.0	1.5	1.5	0.68	0.67	0.86	0.69
	4	2	737	89.7	10.3	100.0	1.8	1.9	0.41	0.37	0.80	0.58
	5	2	4944	87.7	11.7	99.4	1.8	1.9	0.42	0.39	0.72	0.51
34	Overall	2	12429	68.4	30.2	98.6	1.6	1.6	0.55	0.61	0.64	0.40
	1	2	6235	67.9	30.8	98.7	1.6	1.6	0.56	0.62	0.65	0.41
	2	2	321	76.0	23.7	99.7	1.7	1.7	0.46	0.48	0.62	0.40
	3	2	192	59.4	37.5	96.9	1.2	1.1	0.66	0.74	0.67	0.42
	4	2	737	68.4	30.1	98.5	1.6	1.5	0.53	0.66	0.67	0.43
	5	2	4944	68.9	29.6	98.5	1.7	1.6	0.52	0.59	0.60	0.37
35	Overall	3	12429	70.8	26.8	97.6	1.0	0.9	0.94	0.93	0.88	0.69
	1	3	6235	72.5	25.2	97.6	0.9	0.8	0.92	0.91	0.88	0.69
	2	3	321	67.0	31.5	98.4	1.2	1.2	0.94	1.01	0.89	0.68
	3	3	192	78.6	21.4	100.0	0.4	0.4	0.68	0.68	0.87	0.65
	4	3	737	76.1	22.4	98.5	1.2	1.1	0.95	0.94	0.91	0.75
	5	3	4944	67.8	29.5	97.3	1.0	1.1	0.95	0.92	0.87	0.66

Approximate agreement (%) is the percent of pairs of readers that differ by one score point.

Total agreement (%) is the sum of exact and approximate percents. The scoring models are: 1) Regional scoring; 2) Schools from two districts; 3) Three or more schools within a district; 4) Two schools within a district; and 5) Only one school.

Table C- 6. New York State Public Schools Grade 8 ELA Operational Test 2008: Inter-rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			RS Mean		RS SD		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
27-30	Overall	5	13875	40.6	48.5	89.2	3.5	3.1	1.15	1.06	0.76	0.43
	1	5	6946	41.3	48.0	89.3	3.4	3.0	1.18	1.11	0.79	0.46
	2	5	92	46.7	48.9	95.7	3.2	3.3	1.04	0.86	0.78	0.45
	3	5	1524	45.3	46.1	91.3	3.1	2.9	1.14	1.02	0.78	0.47
	4	5	1542	39.8	48.8	88.7	3.9	3.4	1.01	0.96	0.69	0.38
	5	5	3771	37.8	50.4	88.2	3.7	3.2	1.05	0.97	0.69	0.36
31-34	Overall	5	13875	44.4	47.6	91.9	3.8	3.7	1.07	0.98	0.76	0.42
	1	5	6946	44.0	47.8	91.7	3.6	3.7	1.11	0.99	0.77	0.43
	2	5	92	43.5	54.3	97.8	3.9	3.7	0.97	0.72	0.73	0.38
	3	5	1524	43.6	47.9	91.5	3.5	3.4	1.13	1.10	0.80	0.47
	4	5	1542	46.6	45.8	92.4	4.1	4.0	0.93	0.88	0.68	0.36
	5	5	3771	44.5	47.7	92.1	4.0	3.8	0.97	0.92	0.71	0.38
30&34	Overall	3	13875	60.2	38.7	98.8	2.3	2.2	0.70	0.68	0.71	0.43
	1	3	6946	59.1	39.9	98.9	2.3	2.2	0.73	0.69	0.72	0.43
	2	3	92	65.2	33.7	98.9	2.3	2.3	0.70	0.56	0.69	0.46
	3	3	1524	58.1	39.6	97.7	2.2	2.0	0.74	0.71	0.70	0.43
	4	3	1542	62.5	37.2	99.6	2.4	2.4	0.61	0.61	0.65	0.39
	5	3	3771	61.9	36.9	98.8	2.4	2.3	0.65	0.64	0.67	0.42

Approximate agreement (%) is the percent of pairs of readers that differ by one score point.

Total agreement (%) is the sum of exact and approximate percents.

The scoring models are: 1) Regional scoring; 2) Schools from two districts; 3) Three or more schools within a district; 4) Two schools within a district; and 5) Only one school.

Appendix D

Item Level Statistics for English Including All Schools in State Without New York City Schools

Table D- 1. New York State Public Schools (Without NYC) Grade 3 ELA Operational Test 2008: Inter-rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			RS Mean		RS SD		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
21	Overall	2	8423	96.5	3.3	99.8	1.9	1.9	0.40	0.42	0.93	0.83
	1	2	2095	95.2	4.6	99.8	1.9	1.8	0.45	0.47	0.93	0.82
	2	2	221	98.2	1.8	100.0	2.0	1.9	0.27	0.29	0.94	0.83
	3	2	3689	97.0	2.7	99.7	1.9	1.9	0.39	0.40	0.93	0.84
	4	2	832	97.1	2.9	100.0	1.9	1.9	0.38	0.40	0.95	0.85
	5	2	1586	96.3	3.5	99.7	1.9	1.9	0.39	0.39	0.92	0.81
26	Overall	2	8423	81.8	17.6	99.5	1.6	1.6	0.58	0.62	0.84	0.67
	1	2	2095	80.1	19.1	99.2	1.6	1.5	0.61	0.64	0.84	0.66
	2	2	221	79.6	20.4	100.0	1.7	1.7	0.52	0.52	0.77	0.56
	3	2	3689	82.6	16.9	99.5	1.7	1.6	0.56	0.61	0.84	0.67
	4	2	832	82.8	16.9	99.8	1.6	1.6	0.58	0.62	0.86	0.69
	5	2	1586	82.2	17.3	99.6	1.7	1.6	0.58	0.60	0.84	0.66
27	Overall	2	8423	87.3	12.0	99.3	1.1	1.1	0.87	0.87	0.95	0.86
	1	2	2095	86.9	12.1	99.0	1.0	1.0	0.87	0.87	0.94	0.85
	2	2	221	82.8	16.3	99.1	1.3	1.2	0.81	0.87	0.92	0.80
	3	2	3689	86.8	12.5	99.3	1.1	1.1	0.88	0.88	0.95	0.85
	4	2	832	89.1	10.7	99.8	1.2	1.2	0.86	0.87	0.96	0.88
	5	2	1586	88.7	10.7	99.4	1.1	1.1	0.88	0.87	0.95	0.87
28	Overall	3	8423	95.1	4.1	99.1	2.7	2.7	0.73	0.72	0.96	0.89
	1	3	2095	93.7	5.2	98.8	2.6	2.7	0.80	0.79	0.96	0.87
	2	3	221	98.2	1.8	100.0	2.8	2.8	0.52	0.50	0.98	0.94
	3	3	3689	95.9	3.4	99.3	2.7	2.7	0.72	0.71	0.97	0.90
	4	3	832	96.9	2.8	99.6	2.7	2.7	0.71	0.69	0.98	0.93
	5	3	1586	93.6	5.2	98.8	2.7	2.7	0.69	0.69	0.94	0.84

Approximate agreement (%) is the percent of pairs of readers that differ by one score point.

Total agreement (%) is the sum of exact and approximate percents.

The scoring models are: 1) Regional scoring; 2) Schools from two districts; 3) Three or more schools within a district; 4) Two schools within a district; and 5) Only one school.

Table D- 2. New York State Public Schools (Without NYC) Grade 4 ELA Operational Test 2008: Inter-rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			RS Mean		RS SD		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
29-31	Overall	4	9003	57.1	40.9	98.0	2.9	2.8	0.80	0.76	0.75	0.45
	1	4	2928	55.0	42.7	97.7	2.8	2.8	0.82	0.78	0.74	0.44
	2	4	55	54.5	41.8	96.4	2.5	2.8	0.66	0.73	0.63	0.35
	3	4	4256	57.4	40.5	97.9	3.0	2.9	0.78	0.76	0.74	0.45
	4	4	465	57.4	41.7	99.1	3.0	2.8	0.71	0.68	0.70	0.40
	5	4	1299	61.0	37.9	98.8	2.9	2.8	0.79	0.72	0.77	0.50
32-35	Overall	4	9003	53.7	43.2	97.0	2.8	2.9	0.85	0.91	0.78	0.47
	1	4	2928	52.7	44.6	97.3	2.7	2.8	0.87	0.93	0.79	0.48
	2	4	55	49.1	45.5	94.5	2.5	2.8	0.78	1.07	0.78	0.45
	3	4	4256	54.1	42.4	96.5	2.8	2.9	0.83	0.90	0.76	0.46
	4	4	465	51.0	45.8	96.8	2.9	2.8	0.81	0.84	0.73	0.41
	5	4	1299	56.1	41.7	97.8	2.8	2.8	0.88	0.92	0.82	0.52
31&35	Overall	3	9003	59.1	39.1	98.2	2.1	2.2	0.73	0.70	0.70	0.42
	1	3	2928	58.8	39.3	98.1	2.1	2.2	0.73	0.70	0.70	0.42
	2	3	55	47.3	47.3	94.5	1.8	2.1	0.74	1.00	0.73	0.42
	3	3	4256	58.7	39.5	98.2	2.1	2.2	0.73	0.70	0.70	0.42
	4	3	465	61.7	37.8	99.6	2.1	2.2	0.70	0.61	0.70	0.42
	5	3	1299	60.7	37.5	98.2	2.1	2.2	0.72	0.71	0.72	0.45

Approximate agreement (%) is the percent of pairs of readers that differ by one score point.

Total agreement (%) is the sum of exact and approximate percents.

The scoring models are: 1) Regional scoring; 2) Schools from two districts; 3) Three or more schools within a district; 4) Two schools within a district; and 5) Only one school.

Table D- 3. New York State Public Schools (Without NYC) Grade 5 ELA Operational Test 2008: Inter-rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			RS Mean		RS SD		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
21	Overall	2	8544	76.1	22.2	98.2	1.2	1.1	0.62	0.65	0.78	0.61
	1	2	1609	75.0	23.1	98.0	1.1	1.0	0.64	0.68	0.78	0.60
	2	2	353	70.0	27.2	97.2	1.1	1.0	0.67	0.66	0.72	0.53
	3	2	4090	76.5	21.6	98.1	1.2	1.1	0.61	0.65	0.78	0.61
	4	2	1309	80.7	18.4	99.2	1.3	1.2	0.58	0.63	0.83	0.67
	5	2	1183	72.7	25.4	98.1	1.2	1.1	0.63	0.64	0.75	0.56
26	Overall	2	8544	84.5	15.1	99.6	1.7	1.6	0.50	0.56	0.83	0.67
	1	2	1609	83.6	15.9	99.5	1.7	1.6	0.53	0.60	0.84	0.68
	2	2	353	82.4	17.3	99.7	1.7	1.7	0.49	0.53	0.78	0.60
	3	2	4090	84.8	14.7	99.5	1.7	1.6	0.50	0.56	0.83	0.67
	4	2	1309	84.6	15.3	99.9	1.7	1.7	0.49	0.51	0.82	0.65
	5	2	1183	85.3	14.2	99.5	1.7	1.6	0.51	0.58	0.85	0.70
27	Overall	3	8544	69.5	27.8	97.4	1.5	1.5	1.07	1.06	0.91	0.72
	1	3	1609	68.9	28.6	97.5	1.3	1.3	1.05	1.04	0.90	0.71
	2	3	353	75.6	23.8	99.4	1.3	1.3	1.05	1.06	0.94	0.79
	3	3	4090	69.7	27.5	97.1	1.5	1.5	1.08	1.06	0.91	0.72
	4	3	1309	68.8	28.3	97.2	1.6	1.6	1.03	1.04	0.89	0.70
	5	3	1183	68.6	28.8	97.5	1.4	1.5	1.05	1.02	0.90	0.70

Approximate agreement (%) is the percent of pairs of readers that differ by one score point.

Total agreement (%) is the sum of exact and approximate percents.

The scoring models are: 1) Regional scoring; 2) Schools from two districts; 3) Three or more schools within a district; 4) Two schools within a district; and 5) Only one school.

Table D- 4. New York State Public Schools (Without NYC) Grade 6 ELA Operational Test 2008: Inter-rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			RS Mean		RS SD		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
27-30	Overall	5	8610	42.1	48.7	90.8	3.3	3.0	1.06	0.90	0.71	0.38
	1	5	2256	44.1	48.9	93.0	3.2	3.0	1.09	0.90	0.75	0.42
	2	5	31	29.0	41.9	71.0	3.5	2.7	1.16	0.64	0.43	0.23
	3	5	1673	38.1	50.0	88.1	3.2	2.9	1.08	0.88	0.66	0.32
	4	5	1375	41.9	49.4	91.3	3.5	3.1	1.12	0.91	0.75	0.42
	5	5	3275	43.1	47.5	90.7	3.5	3.1	0.97	0.90	0.68	0.36
31-34	Overall	5	8610	43.7	47.4	91.1	3.3	3.2	1.08	1.09	0.78	0.45
	1	5	2256	45.2	46.4	91.6	3.2	3.1	1.09	1.13	0.80	0.47
	2	5	31	45.2	45.2	90.3	3.2	2.9	1.15	1.00	0.74	0.44
	3	5	1673	39.9	50.3	90.2	3.3	3.2	1.10	1.06	0.75	0.40
	4	5	1375	44.9	46.7	91.6	3.5	3.3	1.08	1.12	0.79	0.47
	5	5	3275	44.0	46.9	90.9	3.4	3.2	1.04	1.06	0.76	0.43
30&34	Overall	3	8610	60.1	38.7	98.8	2.3	2.2	0.69	0.72	0.72	0.44
	1	3	2256	59.1	39.5	98.6	2.2	2.1	0.69	0.73	0.71	0.43
	2	3	31	38.7	51.6	90.3	2.4	1.7	0.65	0.52	0.03	0.10
	3	3	1673	59.1	39.6	98.7	2.2	2.2	0.67	0.70	0.69	0.41
	4	3	1375	63.6	35.7	99.3	2.4	2.3	0.68	0.74	0.76	0.50
	5	3	3275	60.1	38.7	98.8	2.3	2.2	0.69	0.70	0.71	0.43

Approximate agreement (%) is the percent of pairs of readers that differ by one score point.

Total agreement (%) is the sum of exact and approximate percents.

The scoring models are: 1) Regional scoring; 2) Schools from two districts; 3) Three or more schools within a district; 4) Two schools within a district; and 5) Only one school.

Table D- 5. New York State Public Schools (Without NYC) Grade 7 ELA Operational Test 2008: Inter-rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			RS Mean		RS SD		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
27	Overall	2	8570	67.4	31.9	99.3	1.5	1.5	0.64	0.62	0.72	0.47
	1	2	2376	67.3	32.1	99.4	1.4	1.5	0.63	0.63	0.72	0.47
	2	2	321	68.8	30.8	99.7	1.5	1.7	0.59	0.52	0.66	0.42
	3	2	192	59.9	38.5	98.4	1.0	1.2	0.77	0.66	0.73	0.46
	4	2	737	67.6	31.9	99.5	1.5	1.4	0.61	0.64	0.72	0.48
	5	2	4944	67.7	31.7	99.3	1.5	1.5	0.64	0.61	0.72	0.47
28	Overall	2	8570	65.1	33.5	98.7	1.3	1.3	0.71	0.70	0.76	0.51
	1	2	2376	66.0	33.0	99.1	1.3	1.3	0.68	0.68	0.76	0.51
	2	2	321	67.6	31.8	99.4	1.3	1.4	0.71	0.67	0.78	0.54
	3	2	192	59.4	39.6	99.0	0.7	0.9	0.71	0.69	0.72	0.45
	4	2	737	68.8	30.0	98.8	1.4	1.2	0.67	0.71	0.78	0.56
	5	2	4944	64.2	34.2	98.4	1.3	1.3	0.72	0.70	0.75	0.50
33	Overall	2	8570	87.6	11.9	99.5	1.8	1.9	0.43	0.40	0.74	0.53
	1	2	2376	87.2	12.4	99.6	1.8	1.9	0.42	0.39	0.74	0.52
	2	2	321	88.8	10.3	99.1	1.8	1.9	0.42	0.34	0.69	0.48
	3	2	192	80.7	18.2	99.0	1.5	1.5	0.68	0.67	0.86	0.69
	4	2	737	89.7	10.3	100.0	1.8	1.9	0.41	0.37	0.80	0.58
	5	2	4944	87.7	11.7	99.4	1.8	1.9	0.42	0.39	0.72	0.51
34	Overall	2	8570	68.9	29.8	98.6	1.6	1.6	0.53	0.61	0.63	0.40
	1	2	2376	68.9	30.1	98.9	1.6	1.6	0.54	0.60	0.65	0.41
	2	2	321	76.0	23.7	99.7	1.7	1.7	0.46	0.48	0.62	0.40
	3	2	192	59.4	37.5	96.9	1.2	1.1	0.66	0.74	0.67	0.42
	4	2	737	68.4	30.1	98.5	1.6	1.5	0.53	0.66	0.67	0.43
	5	2	4944	68.9	29.6	98.5	1.7	1.6	0.52	0.59	0.60	0.37

Table D- 5 continued. New York State Public Schools (Without NYC) Grade 7 ELA Operational Test 2008: Inter-rater Agreement (continued)

35	Overall	3	8570	70.3	27.3	97.6	1.0	1.1	0.95	0.93	0.88	0.69
	1	3	2376	73.7	24.2	97.8	1.0	1.1	0.95	0.94	0.90	0.72
	2	3	321	67.0	31.5	98.4	1.2	1.2	0.94	1.01	0.89	0.68
	3	3	192	78.6	21.4	100.0	0.4	0.4	0.68	0.68	0.87	0.65
	4	3	737	76.1	22.4	98.5	1.2	1.1	0.95	0.94	0.91	0.75
	5	3	4944	67.8	29.5	97.3	1.0	1.1	0.95	0.92	0.87	0.66

Approximate agreement (%) is the percent of pairs of readers that differ by one score point.

Total agreement (%) is the sum of exact and approximate percents.

The scoring models are: 1) Regional scoring; 2) Schools from two districts; 3) Three or more schools within a district; 4) Two schools within a district; and 5) Only one school.

Table D- 6. New York State Public Schools (Without NYC) Grade 8 ELA Operational Test 2008: Inter-rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			RS Mean		RS SD		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
27-30	Overall	5	9191	40.7	48.9	89.7	3.6	3.2	1.09	1.00	0.74	0.41
	1	5	2262	43.0	48.5	91.5	3.6	3.3	1.07	1.02	0.76	0.43
	2	5	92	46.7	48.9	95.7	3.2	3.3	1.04	0.86	0.78	0.45
	3	5	1524	45.3	46.1	91.3	3.1	2.9	1.14	1.02	0.78	0.47
	4	5	1542	39.8	48.8	88.7	3.9	3.4	1.01	0.96	0.69	0.38
	5	5	3771	37.8	50.4	88.2	3.7	3.2	1.05	0.97	0.69	0.36
31-34	Overall	5	9191	44.7	47.5	92.1	3.8	3.7	1.03	0.96	0.75	0.41
	1	5	2262	44.5	47.7	92.2	3.7	3.8	1.03	0.94	0.74	0.40
	2	5	92	43.5	54.3	97.8	3.9	3.7	0.97	0.72	0.73	0.38
	3	5	1524	43.6	47.9	91.5	3.5	3.4	1.13	1.10	0.80	0.47
	4	5	1542	46.6	45.8	92.4	4.1	4.0	0.93	0.88	0.68	0.36
	5	5	3771	44.5	47.7	92.1	4.0	3.8	0.97	0.92	0.71	0.38
30&34	Overall	3	9191	61.4	37.5	98.9	2.4	2.2	0.67	0.66	0.69	0.42
	1	3	2262	61.7	37.6	99.3	2.3	2.3	0.66	0.66	0.69	0.42
	2	3	92	65.2	33.7	98.9	2.3	2.3	0.70	0.56	0.69	0.46
	3	3	1524	58.1	39.6	97.7	2.2	2.0	0.74	0.71	0.70	0.43
	4	3	1542	62.5	37.2	99.6	2.4	2.4	0.61	0.61	0.65	0.39
	5	3	3771	61.9	36.9	98.8	2.4	2.3	0.65	0.64	0.67	0.42

Approximate agreement (%) is the percent of pairs of readers that differ by one score point.

Total agreement (%) is the sum of exact and approximate percents.

The scoring models are: 1) Regional scoring; 2) Schools from two districts; 3) Three or more schools within a district; 4) Two schools within a district; and 5) Only one school.

Appendix E

Item Level Statistics for English Including New York City Schools Only

Table E- 1. NYC Public Schools Grades 3 - 8 ELA Operational Test 2008: Inter-rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			RS Mean		RS SD		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
Grade 3												
21	NYC	2	4562	96.6	3.1	99.7	1.8	1.8	0.49	0.49	0.95	0.88
26	NYC	2	4562	80.1	19.1	99.2	1.5	1.5	0.65	0.67	0.85	0.68
27	NYC	2	4562	79.7	18.1	97.7	1.0	1.0	0.86	0.88	0.90	0.76
28	NYC	3	4562	94.1	5.1	99.2	2.6	2.6	0.83	0.81	0.96	0.89
Grade 4												
29-31	NYC	4	5136	54.4	42.2	96.5	2.7	2.8	0.89	0.79	0.75	0.46
32-35	NYC	4	5136	47.3	47.4	94.7	2.5	2.7	0.94	0.97	0.77	0.45
31&35	NYC	3	5136	55.5	42.3	97.8	2.1	2.1	0.78	0.70	0.70	0.40
Grade 5												
21	NYC	2	5154	72.7	25.8	98.5	1.1	1.1	0.68	0.72	0.81	0.61
26	NYC	2	5154	80.8	18.6	99.4	1.5	1.5	0.61	0.65	0.85	0.68
27	NYC	3	5154	69.8	27.6	97.4	1.0	1.0	1.02	1.02	0.90	0.70
Grade 6												
27-30	NYC	5	4157	46.0	46.1	92.1	2.8	2.7	1.13	0.98	0.79	0.46
31-34	NYC	5	4157	44.0	46.5	90.5	2.9	2.9	1.13	1.05	0.78	0.45
30&34	NYC	3	4157	55.8	42.0	97.8	2.0	2.0	0.75	0.73	0.70	0.40
Grade 7												
27	NYC	2	3859	58.0	40.1	98.1	1.4	1.2	0.72	0.68	0.68	0.42
28	NYC	2	3859	62.6	35.1	97.7	1.2	1.1	0.77	0.73	0.76	0.51
33	NYC	2	3859	84.1	15.4	99.4	1.7	1.8	0.52	0.51	0.80	0.59
34	NYC	2	3859	67.3	31.3	98.6	1.6	1.5	0.57	0.63	0.66	0.41
35	NYC	3	3859	71.7	25.8	97.5	0.8	0.7	0.89	0.87	0.87	0.66
Grade 8												
27-30	NYC	5	4684	40.5	47.8	88.2	3.3	2.9	1.22	1.13	0.79	0.46
31-34	NYC	5	4684	43.7	47.8	91.5	3.6	3.6	1.14	1.01	0.78	0.44
30&34	NYC	3	4684	57.8	40.9	98.8	2.2	2.1	0.76	0.70	0.72	0.43

Approximate agreement (%) is the percent of pairs of readers that differ by one score point.
 Total agreement (%) is the sum of exact and approximate percents.

Appendix F

Item Level Differences for English Including All Schools in State

Table F- 1. New York State Public Schools Grade 3 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]

Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
21	Overall			0.00	0.00	0.01	0.97	0.02	0.00	0.00		
	1			0.00	0.00	0.01	0.96	0.02	0.00	0.00		
	2			0.00	0.00	0.00	0.98	0.02	0.00	0.00		
	3			0.00	0.00	0.01	0.97	0.02	0.00	0.00		
	4			0.00	0.00	0.01	0.97	0.02	0.00	0.00		
	5			0.00	0.00	0.02	0.96	0.02	0.00	0.00		
26	Overall			0.00	0.00	0.07	0.81	0.11	0.00	0.00		
	1			0.00	0.00	0.08	0.80	0.11	0.01	0.00		
	2			0.00	0.00	0.10	0.80	0.11	0.00	0.00		
	3			0.00	0.00	0.05	0.83	0.12	0.01	0.00		
	4			0.00	0.00	0.06	0.83	0.11	0.00	0.00		
	5			0.00	0.00	0.06	0.82	0.11	0.00	0.00		
27	Overall			0.00	0.00	0.06	0.85	0.08	0.01	0.00		
	1			0.00	0.00	0.07	0.82	0.09	0.02	0.00		
	2			0.00	0.00	0.05	0.83	0.12	0.01	0.00		
	3			0.00	0.00	0.06	0.87	0.07	0.00	0.00		
	4			0.00	0.00	0.06	0.89	0.05	0.00	0.00		
	5			0.00	0.00	0.05	0.89	0.06	0.00	0.00		
28	Overall			0.00	0.00	0.03	0.95	0.02	0.00	0.00		
	1			0.00	0.00	0.03	0.94	0.02	0.00	0.00		
	2			0.00	0.00	0.02	0.98	0.00	0.00	0.00		
	3			0.00	0.00	0.02	0.96	0.01	0.00	0.00		
	4			0.00	0.00	0.02	0.97	0.01	0.00	0.00		
	5			0.00	0.00	0.03	0.94	0.02	0.01	0.00		

Table F- 2. New York State Public Schools Grade 4 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]

Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
29-31	Overall		0.00	0.00	0.01	0.21	0.56	0.20	0.01	0.00	0.00	
	1		0.00	0.00	0.02	0.25	0.55	0.17	0.01	0.00	0.00	
	2		0.00	0.00	0.04	0.36	0.55	0.05	0.00	0.00	0.00	
	3		0.00	0.00	0.01	0.17	0.57	0.23	0.01	0.00	0.00	
	4		0.00	0.00	0.01	0.12	0.57	0.29	0.00	0.00	0.00	
	5		0.00	0.00	0.00	0.13	0.61	0.25	0.01	0.00	0.00	
32-35	Overall		0.00	0.00	0.03	0.27	0.51	0.18	0.01	0.00	0.00	
	1		0.00	0.00	0.03	0.30	0.49	0.17	0.01	0.00	0.00	
	2		0.00	0.00	0.05	0.33	0.49	0.13	0.00	0.00	0.00	
	3		0.00	0.00	0.02	0.24	0.54	0.18	0.01	0.00	0.00	
	4		0.00	0.00	0.01	0.20	0.51	0.26	0.02	0.00	0.00	
	5		0.00	0.00	0.01	0.21	0.56	0.21	0.01	0.00	0.00	
31&35	Overall		0.00	0.00	0.01	0.23	0.58	0.17	0.01	0.00	0.00	
	1		0.00	0.00	0.01	0.24	0.57	0.18	0.00	0.00	0.00	
	2		0.00	0.00	0.04	0.36	0.47	0.11	0.02	0.00	0.00	
	3		0.00	0.00	0.01	0.23	0.59	0.17	0.01	0.00	0.00	
	4		0.00	0.00	0.00	0.22	0.62	0.16	0.00	0.00	0.00	
	5		0.00	0.00	0.01	0.20	0.61	0.17	0.01	0.00	0.00	

Table F- 3. New York State Public Schools Grade 5 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]

Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
21	Overall			0.00	0.00	0.08	0.75	0.15	0.01	0.00		
	1			0.00	0.00	0.09	0.73	0.16	0.01	0.00		
	2			0.00	0.01	0.13	0.70	0.14	0.02	0.00		
	3			0.00	0.00	0.07	0.76	0.15	0.02	0.00		
	4			0.00	0.00	0.05	0.81	0.14	0.01	0.00		
	5			0.00	0.00	0.07	0.73	0.18	0.02	0.00		
26	Overall			0.00	0.00	0.06	0.83	0.10	0.00	0.00		
	1			0.00	0.00	0.08	0.82	0.10	0.00	0.00		
	2			0.00	0.00	0.07	0.82	0.10	0.00	0.00		
	3			0.00	0.00	0.04	0.85	0.11	0.00	0.00		
	4			0.00	0.00	0.07	0.85	0.09	0.00	0.00		
	5			0.00	0.00	0.04	0.85	0.10	0.00	0.00		
27	Overall			0.00	0.01	0.16	0.70	0.12	0.01	0.00		
	1			0.00	0.01	0.16	0.70	0.12	0.01	0.00		
	2			0.00	0.00	0.13	0.76	0.10	0.00	0.00		
	3			0.00	0.02	0.15	0.70	0.12	0.01	0.00		
	4			0.00	0.01	0.14	0.69	0.14	0.01	0.00		
	5			0.00	0.02	0.18	0.69	0.11	0.01	0.00		

Table F- 4. New York State Public Schools Grade 6 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]

Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
27-30	Overall	0.00	0.00	0.00	0.02	0.17	0.43	0.31	0.06	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.03	0.20	0.45	0.27	0.05	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.06	0.29	0.35	0.29	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.02	0.16	0.38	0.34	0.09	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.01	0.14	0.42	0.36	0.07	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.01	0.13	0.43	0.35	0.08	0.01	0.00	0.00
31-34	Overall	0.00	0.00	0.00	0.03	0.21	0.44	0.26	0.05	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.04	0.23	0.44	0.24	0.05	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.19	0.45	0.26	0.06	0.03	0.00	0.00
	3	0.00	0.00	0.00	0.04	0.21	0.40	0.29	0.05	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.02	0.18	0.45	0.29	0.06	0.01	0.00	0.00
	5	0.00	0.00	0.00	0.03	0.19	0.44	0.28	0.06	0.00	0.00	0.00
30&34	Overall	0.00	0.00	0.00	0.01	0.18	0.59	0.22	0.01	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.01	0.20	0.57	0.21	0.01	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.03	0.39	0.48	0.10	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.01	0.19	0.59	0.20	0.01	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00	0.13	0.64	0.22	0.01	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.01	0.16	0.60	0.23	0.01	0.00	0.00	0.00

Table F- 5. New York State Public Schools Grade 7 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]

Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
27	Overall			0.00	0.00	0.15	0.65	0.19	0.01	0.00		
	1			0.00	0.00	0.14	0.62	0.23	0.01	0.00		
	2			0.00	0.00	0.21	0.69	0.10	0.00	0.00		
	3			0.00	0.02	0.25	0.60	0.14	0.00	0.00		
	4			0.00	0.00	0.12	0.68	0.20	0.00	0.00		
	5			0.00	0.00	0.16	0.68	0.16	0.00	0.00		
28	Overall			0.00	0.01	0.15	0.64	0.19	0.01	0.00		
	1			0.00	0.01	0.15	0.64	0.20	0.01	0.00		
	2			0.00	0.01	0.19	0.68	0.13	0.00	0.00		
	3			0.00	0.01	0.27	0.59	0.13	0.00	0.00		
	4			0.00	0.00	0.08	0.69	0.22	0.01	0.00		
	5			0.00	0.01	0.15	0.64	0.19	0.01	0.00		
33	Overall			0.00	0.00	0.09	0.87	0.04	0.00	0.00		
	1			0.00	0.00	0.10	0.85	0.05	0.00	0.00		
	2			0.00	0.01	0.09	0.89	0.02	0.00	0.00		
	3			0.00	0.01	0.12	0.81	0.06	0.01	0.00		
	4			0.00	0.00	0.08	0.90	0.03	0.00	0.00		
	5			0.00	0.00	0.08	0.88	0.04	0.00	0.00		
34	Overall			0.00	0.00	0.13	0.68	0.17	0.01	0.00		
	1			0.00	0.00	0.14	0.68	0.17	0.01	0.00		
	2			0.00	0.00	0.13	0.76	0.11	0.00	0.00		
	3			0.00	0.01	0.14	0.59	0.23	0.03	0.00		
	4			0.00	0.00	0.13	0.68	0.17	0.01	0.00		
	5			0.00	0.00	0.12	0.69	0.17	0.01	0.00		
35	Overall			0.00	0.01	0.13	0.71	0.14	0.01	0.00		
	1			0.00	0.01	0.11	0.72	0.14	0.01	0.00		
	2			0.00	0.01	0.14	0.67	0.17	0.00	0.00		
	3			0.00	0.00	0.09	0.79	0.13	0.00	0.00		
	4			0.00	0.01	0.10	0.76	0.12	0.01	0.00		
	5			0.00	0.01	0.15	0.68	0.14	0.01	0.00		

Table F- 6. New York State Public Schools Grade 8 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]

Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
27-30	Overall	0.00	0.00	0.00	0.01	0.13	0.41	0.36	0.09	0.01	0.00	0.00
	1	0.00	0.00	0.00	0.01	0.14	0.41	0.34	0.09	0.01	0.00	0.00
	2	0.00	0.00	0.00	0.04	0.23	0.47	0.26	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.02	0.17	0.45	0.29	0.06	0.01	0.00	0.00
	4	0.00	0.00	0.00	0.01	0.08	0.40	0.41	0.10	0.01	0.00	0.00
	5	0.00	0.00	0.00	0.01	0.10	0.38	0.40	0.10	0.01	0.00	0.00
31-34	Overall	0.00	0.00	0.00	0.04	0.21	0.44	0.26	0.04	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.05	0.24	0.44	0.23	0.03	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.02	0.17	0.43	0.37	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.04	0.21	0.44	0.26	0.04	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.03	0.18	0.47	0.28	0.04	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.02	0.17	0.44	0.30	0.05	0.00	0.00	0.00
30&34	Overall	0.00	0.00	0.00	0.01	0.14	0.60	0.25	0.01	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.01	0.15	0.59	0.25	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.16	0.65	0.17	0.01	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.01	0.12	0.58	0.28	0.02	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00	0.17	0.62	0.20	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00	0.11	0.62	0.26	0.01	0.00	0.00	0.00

Appendix G

Item Level Differences for English All Schools in State Without New York City Schools

Table G- 1. New York State Public Schools Grade 3 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]

Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
21	Overall			0.00	0.00	0.01	0.96	0.02	0.00	0.00		
	1			0.00	0.00	0.02	0.95	0.03	0.00	0.00		
	2			0.00	0.00	0.00	0.98	0.02	0.00	0.00		
	3			0.00	0.00	0.01	0.97	0.02	0.00	0.00		
	4			0.00	0.00	0.01	0.97	0.02	0.00	0.00		
	5			0.00	0.00	0.02	0.96	0.02	0.00	0.00		
26	Overall			0.00	0.00	0.06	0.82	0.12	0.00	0.00		
	1			0.00	0.00	0.08	0.80	0.11	0.01	0.00		
	2			0.00	0.00	0.10	0.80	0.11	0.00	0.00		
	3			0.00	0.00	0.05	0.83	0.12	0.01	0.00		
	4			0.00	0.00	0.06	0.83	0.11	0.00	0.00		
	5			0.00	0.00	0.06	0.82	0.11	0.00	0.00		
27	Overall			0.00	0.00	0.06	0.87	0.06	0.00	0.00		
	1			0.00	0.00	0.06	0.87	0.06	0.01	0.00		
	2			0.00	0.00	0.05	0.83	0.12	0.01	0.00		
	3			0.00	0.00	0.06	0.87	0.07	0.00	0.00		
	4			0.00	0.00	0.06	0.89	0.05	0.00	0.00		
	5			0.00	0.00	0.05	0.89	0.06	0.00	0.00		
28	Overall			0.00	0.00	0.03	0.95	0.01	0.00	0.00		
	1			0.00	0.01	0.03	0.94	0.02	0.00	0.00		
	2			0.00	0.00	0.02	0.98	0.00	0.00	0.00		
	3			0.00	0.00	0.02	0.96	0.01	0.00	0.00		
	4			0.00	0.00	0.02	0.97	0.01	0.00	0.00		
	5			0.00	0.00	0.03	0.94	0.02	0.01	0.00		

Table G- 2. New York State Public Schools Grade 4 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]

Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
29-31	Overall		0.00	0.00	0.01	0.18	0.57	0.23	0.01	0.00	0.00	
	1		0.00	0.00	0.01	0.23	0.55	0.20	0.01	0.00	0.00	
	2		0.00	0.00	0.04	0.36	0.55	0.05	0.00	0.00	0.00	
	3		0.00	0.00	0.01	0.17	0.57	0.23	0.01	0.00	0.00	
	4		0.00	0.00	0.01	0.12	0.57	0.29	0.00	0.00	0.00	
	5		0.00	0.00	0.00	0.13	0.61	0.25	0.01	0.00	0.00	
32-35	Overall		0.00	0.00	0.02	0.25	0.54	0.18	0.01	0.00	0.00	
	1		0.00	0.00	0.02	0.28	0.53	0.16	0.01	0.00	0.00	
	2		0.00	0.00	0.05	0.33	0.49	0.13	0.00	0.00	0.00	
	3		0.00	0.00	0.02	0.24	0.54	0.18	0.01	0.00	0.00	
	4		0.00	0.00	0.01	0.20	0.51	0.26	0.02	0.00	0.00	
	5		0.00	0.00	0.01	0.21	0.56	0.21	0.01	0.00	0.00	
31&35	Overall		0.00	0.00	0.01	0.23	0.59	0.16	0.01	0.00	0.00	
	1		0.00	0.00	0.01	0.23	0.59	0.16	0.00	0.00	0.00	
	2		0.00	0.00	0.04	0.36	0.47	0.11	0.02	0.00	0.00	
	3		0.00	0.00	0.01	0.23	0.59	0.17	0.01	0.00	0.00	
	4		0.00	0.00	0.00	0.22	0.62	0.16	0.00	0.00	0.00	
	5		0.00	0.00	0.01	0.20	0.61	0.17	0.01	0.00	0.00	

Table G- 3. New York State Public Schools Grade 5 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]

Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
21	Overall			0.00	0.00	0.07	0.76	0.15	0.02	0.00		
	1			0.00	0.01	0.09	0.75	0.14	0.01	0.00		
	2			0.00	0.01	0.13	0.70	0.14	0.02	0.00		
	3			0.00	0.00	0.07	0.76	0.15	0.02	0.00		
	4			0.00	0.00	0.05	0.81	0.14	0.01	0.00		
	5			0.00	0.00	0.07	0.73	0.18	0.02	0.00		
26	Overall			0.00	0.00	0.05	0.85	0.10	0.00	0.00		
	1			0.00	0.00	0.05	0.84	0.11	0.00	0.00		
	2			0.00	0.00	0.07	0.82	0.10	0.00	0.00		
	3			0.00	0.00	0.04	0.85	0.11	0.00	0.00		
	4			0.00	0.00	0.07	0.85	0.09	0.00	0.00		
	5			0.00	0.00	0.04	0.85	0.10	0.00	0.00		
27	Overall			0.00	0.01	0.16	0.70	0.12	0.01	0.00		
	1			0.00	0.01	0.16	0.69	0.12	0.01	0.00		
	2			0.00	0.00	0.13	0.76	0.10	0.00	0.00		
	3			0.00	0.02	0.15	0.70	0.12	0.01	0.00		
	4			0.00	0.01	0.14	0.69	0.14	0.01	0.00		
	5			0.00	0.02	0.18	0.69	0.11	0.01	0.00		

Table G- 4. New York State Public Schools Grade 6 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]

Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
27-30	Overall	0.00	0.00	0.00	0.02	0.15	0.42	0.34	0.07	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.02	0.18	0.44	0.31	0.05	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.06	0.29	0.35	0.29	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.02	0.16	0.38	0.34	0.09	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.01	0.14	0.42	0.36	0.07	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.01	0.13	0.43	0.35	0.08	0.01	0.00	0.00
31-34	Overall	0.00	0.00	0.00	0.03	0.19	0.44	0.28	0.06	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.03	0.20	0.45	0.27	0.05	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.19	0.45	0.26	0.06	0.03	0.00	0.00
	3	0.00	0.00	0.00	0.04	0.21	0.40	0.29	0.05	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.02	0.18	0.45	0.29	0.06	0.01	0.00	0.00
	5	0.00	0.00	0.00	0.03	0.19	0.44	0.28	0.06	0.00	0.00	0.00
30&34	Overall	0.00	0.00	0.00	0.01	0.16	0.60	0.22	0.01	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.01	0.16	0.59	0.23	0.01	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.03	0.39	0.48	0.10	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.01	0.19	0.59	0.20	0.01	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00	0.13	0.64	0.22	0.01	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.01	0.16	0.60	0.23	0.01	0.00	0.00	0.00

Table G- 5. New York State Public Schools Grade 7 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]

Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
27	Overall			0.00	0.00	0.16	0.67	0.16	0.00	0.00		
	1			0.00	0.00	0.18	0.67	0.14	0.00	0.00		
	2			0.00	0.00	0.21	0.69	0.10	0.00	0.00		
	3			0.00	0.02	0.25	0.60	0.14	0.00	0.00		
	4			0.00	0.00	0.12	0.68	0.20	0.00	0.00		
	5			0.00	0.00	0.16	0.68	0.16	0.00	0.00		
28	Overall			0.00	0.01	0.15	0.65	0.18	0.01	0.00		
	1			0.00	0.01	0.17	0.66	0.16	0.00	0.00		
	2			0.00	0.01	0.19	0.68	0.13	0.00	0.00		
	3			0.00	0.01	0.27	0.59	0.13	0.00	0.00		
	4			0.00	0.00	0.08	0.69	0.22	0.01	0.00		
	5			0.00	0.01	0.15	0.64	0.19	0.01	0.00		
33	Overall			0.00	0.00	0.08	0.88	0.03	0.00	0.00		
	1			0.00	0.00	0.09	0.87	0.04	0.00	0.00		
	2			0.00	0.01	0.09	0.89	0.02	0.00	0.00		
	3			0.00	0.01	0.12	0.81	0.06	0.01	0.00		
	4			0.00	0.00	0.08	0.90	0.03	0.00	0.00		
	5			0.00	0.00	0.08	0.88	0.04	0.00	0.00		
34	Overall			0.00	0.00	0.13	0.69	0.17	0.01	0.00		
	1			0.00	0.00	0.14	0.69	0.16	0.01	0.00		
	2			0.00	0.00	0.13	0.76	0.11	0.00	0.00		
	3			0.00	0.01	0.14	0.59	0.23	0.03	0.00		
	4			0.00	0.00	0.13	0.68	0.17	0.01	0.00		
	5			0.00	0.00	0.12	0.69	0.17	0.01	0.00		
35	Overall			0.00	0.01	0.14	0.70	0.13	0.01	0.00		
	1			0.00	0.01	0.13	0.74	0.11	0.01	0.00		
	2			0.00	0.01	0.14	0.67	0.17	0.00	0.00		
	3			0.00	0.00	0.09	0.79	0.13	0.00	0.00		
	4			0.00	0.01	0.10	0.76	0.12	0.01	0.00		
	5			0.00	0.01	0.15	0.68	0.14	0.01	0.00		

Table G- 6. New York State Public Schools Grade 8 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]

Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
27-30	Overall	0.00	0.00	0.00	0.01	0.12	0.41	0.37	0.09	0.01	0.00	0.00
	1	0.00	0.00	0.00	0.01	0.14	0.43	0.35	0.07	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.04	0.23	0.47	0.26	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.02	0.17	0.45	0.29	0.06	0.01	0.00	0.00
	4	0.00	0.00	0.00	0.01	0.08	0.40	0.41	0.10	0.01	0.00	0.00
	5	0.00	0.00	0.00	0.01	0.10	0.38	0.40	0.10	0.01	0.00	0.00
31-34	Overall	0.00	0.00	0.00	0.03	0.20	0.45	0.27	0.04	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.05	0.26	0.44	0.21	0.03	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.02	0.17	0.43	0.37	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.04	0.21	0.44	0.26	0.04	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.03	0.18	0.47	0.28	0.04	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.02	0.17	0.44	0.30	0.05	0.00	0.00	0.00
30&34	Overall	0.00	0.00	0.00	0.00	0.14	0.61	0.24	0.01	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.00	0.17	0.62	0.21	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.16	0.65	0.17	0.01	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.01	0.12	0.58	0.28	0.02	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00	0.17	0.62	0.20	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00	0.11	0.62	0.26	0.01	0.00	0.00	0.00

Appendix H

Item Level Differences for English Including New York City Schools Only

Table H- 1. New York State Public Schools Grades 3 – 8 ELA Operational Test 2008: Proportions of Score Differences [Local Scoring minus Audit Scoring]

Grade 3												
Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
21	NYC			0.00	0.00	0.01	0.97	0.02	0.00	0.00		
26	NYC			0.00	0.00	0.09	0.80	0.10	0.01	0.00		
27	NYC			0.00	0.00	0.07	0.80	0.11	0.02	0.00		
28	NYC			0.00	0.00	0.03	0.94	0.02	0.00	0.00		
Grade 4												
Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
29-31	NYC		0.00	0.00	0.02	0.27	0.54	0.15	0.01	0.00	0.00	
32-35	NYC		0.00	0.00	0.04	0.30	0.47	0.17	0.01	0.00	0.00	
31&35	NYC		0.00	0.00	0.01	0.24	0.55	0.18	0.01	0.00	0.00	
Grade 5												
Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
21	NYC			0.00	0.00	0.10	0.73	0.16	0.01	0.00		
26	NYC			0.00	0.00	0.09	0.81	0.10	0.00	0.00		
27	NYC			0.00	0.01	0.16	0.70	0.12	0.01	0.00		
Grade 6												
Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
27-30	NYC	0.00	0.00	0.00	0.03	0.22	0.46	0.24	0.05	0.00	0.00	0.00
31-34	NYC	0.00	0.00	0.00	0.04	0.24	0.44	0.22	0.05	0.00	0.00	0.00
30&34	NYC	0.00	0.00	0.00	0.01	0.22	0.56	0.20	0.01	0.00	0.00	0.00
Grade 7												
Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
27	NYC			0.00	0.01	0.12	0.58	0.28	0.01	0.00		
28	NYC			0.00	0.01	0.13	0.63	0.22	0.01	0.00		
33	NYC			0.00	0.00	0.10	0.84	0.05	0.00	0.00		
34	NYC			0.00	0.00	0.14	0.67	0.18	0.01	0.00		
35	NYC			0.00	0.01	0.10	0.72	0.16	0.01	0.00		
Grade 8												
Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
27-30	NYC	0.00	0.00	0.00	0.01	0.14	0.40	0.34	0.10	0.01	0.00	0.00
31-34	NYC	0.00	0.00	0.00	0.05	0.23	0.44	0.24	0.03	0.00	0.00	0.00
30&34	NYC	0.00	0.00	0.00	0.01	0.14	0.58	0.27	0.00	0.00	0.00	0.00