Inter-Rater Reliability Study of New York State Grades 3–8 English Language Arts Test 2009 Test Administration

Technical Report

Prepared for the New York State Education Department by Pearson

September 21, 2009

Table of Contents

List of Tables

List of Tables (continued)

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 Regional Information Centers across the state and the constructed-response items are scored by teachers at regional scoring centers, in their districts, or in their 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 booklets 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 booklets 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 booklets for each score point using scoring rubrics for the constructed-response questions.

Schools identified for the 2009 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 2009 throughout the state. The operational data for these assessments were collected by NYSED and include both MC and CR scores. The Regional Information Centers scored the MC items and New York State teachers scored the CR items. In April 2009, Pearson conducted the audit study by rescoring the CR items from approximately five percent of all test booklets. Pearson identified a stratified sample of schools from across the state for each of the grade levels that contained approximately 15,000 student test booklets. 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 85,760 ELA test booklets were collected from sample schools and rescored in April 2009.

Audit notification letters were sent to the sample schools in February 2009 and the selected schools sent their student test booklets to Pearson for audit. Pearson rescored the CR 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 State 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 Booklets

Audit Samples

To achieve the target audit sample of 12,500 test booklets per grade level, approximately 15,000 test booklets were sampled. Six stratified random samples of schools were selected, one for each grade, from all New York State schools with Grades 3–8 enrollment to yield the target number of test booklets. Each school was selected for audit at only one grade level. All selected schools were requested to send Pearson their ELA test booklets 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 Need/Resource Capacity Index Districts:	**New York City**	New York City
	Large Cities	Buffalo, Rochester, Syracuse, Yonkers
	Urban-Suburban	Districts at or above 70th percentile on the index with at least 100 students per square mile or enrollment greater than 2500
	Rural	All districts at or above the 70th percentile with fewer than 50 students per square mile or enrollment of less than 2500
Average Need/Resource Capacity Index Districts		All districts between the 20th and 70th percentiles on the index
Low Need/Resource Capacity Index Districts		All districts below the 20th 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/RC 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 Capacity 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	190486	190994	190950	194628	202659	203995	1173712
New York City	63402	63314	62415	62783	65573	65539	383026
Large Cities	8249	7803	7709	7489	8249	8425	47924
High Need Urban/Suburban	16089	15895	15358	15627	16239	16389	95597
High Need Rural	11524	11432	11368	11893	12639	13278	72134
Average Need	58717	59728	60099	62181	65556	66811	373092
Low Need	29532	30436	30819	31771	32078	32129	186765
Charter	2973	2386	3182	2884	2325	1424	15174

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							
Need/Resource Capacity Index Category	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**	**Total**
New York City	0.33	0.33	0.33	0.32	0.32	0.32	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	0.06	0.07	0.06
Average Need	0.31	0.31	0.31	0.32	0.32	0.33	0.32
Low Need	0.16	0.16	0.16	0.16	0.16	0.16	0.16
Charter	0.02	0.01	0.02	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	4993	5059	4903	4839	4853	4819	29466
Large Cities	650	613	606	577	611	620	3675
High Need Urban/Suburban	1267	1248	1206	1204	1202	1205	7333
High Need Rural	907	898	893	917	935	976	5527
Average Need	4624	4691	4721	4792	4852	4913	28593
Low Need	2326	2390	2421	2449	2374	2362	14322
Charter	234	187	250	222	172	105	1171
Totals	15000	15087	15000	15000	15000	15000	90087

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	5002	5069	4902	4846	4858	4816	29493
Large Cities	652	630	621	585	630	626	3744
High Need Urban/Suburban	1263	1245	1201	1209	1199	1204	7321
High Need Rural	906	890	895	924	934	988	5537
Average Need	4625	4688	4727	4800	4857	4918	28615
Low Need	2328	2387	2423	2450	2379	2376	14343
Charter	237	192	263	227	167	118	1204
Totals	15013	15101	15032	15041	15024	15046	90257

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 State'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	0.33	0.34	0.33	0.32	0.32	0.32	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	0.06	0.07	0.06
Average Need	0.31	0.31	0.31	0.32	0.32	0.33	0.32
Low Need	0.16	0.16	0.16	0.16	0.16	0.16	0.16
Charter	0.02	0.01	0.02	0.02	0.01	0.01	0.01

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

III. Data Collection and School Participation

Pearson notified 795 schools and of those, 761 schools returned materials. This represents a participation rate of 96%.

After the test booklets 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 ELA

Obtained N-counts							
	Grade 3	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**	**Totals**
New York City	4909	5131	4882	5329	3888	4817	28956
Large Cities	621	524	559	680	494	626	3504
High Need Urban/Suburban	1139	1236	1227	1160	1151	1142	7055
High Need Rural	878	874	930	984	826	891	5383
Average Need	4501	4528	4174	4878	4082	4423	26586
Low Need	1978	2276	2175	2419	2351	1573	12772
Charter	190	210	293	263	131	125	1212
Totals	14216	14779	14240	15713	12923	13597	85468

Table 8. Obtained Proportions for ELA

Obtained Proportions						
	Grade 3	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**
New York City	0.35	0.35	0.34	0.34	0.30	0.35
Large Cities	0.04	0.04	0.04	0.04	0.04	0.05
High Need Urban/Suburban	0.08	0.08	0.09	0.07	0.09	0.08
High Need Rural	0.06	0.06	0.07	0.06	0.06	0.07
Average Need	0.32	0.31	0.29	0.31	0.32	0.33
Low Need	0.14	0.15	0.15	0.15	0.18	0.12
Charter	0.01	0.01	0.02	0.02	0.01	0.01

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. The largest difference in percents was for the Grade 8 Low Need group, given the actual samples differed by 4%. All other samples differed less than 4% 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 website 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.

Auditors possessed, at a minimum, a four-year college degree. They were selected to work on ELA based on their educational qualifications and their work or scoring experience.

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

Supervisor training took place in Auburn, Washington, from April 1–April 3, 2009. Supervisors trained on all booklets and all grades for which they would score. One hundred forty-five auditors began training on April 6, 2009. Auditors trained on items in a single booklet, completed scoring all booklets, and then trained on a new booklet 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 booklet. 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 booklets. Subsequent practice sets for a booklet 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 booklets and any necessary review, auditors completed the consistency assurance set (CAS) for that booklet. A review and discussion of the scores occurred after auditors had assigned scores to all booklets 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 booklets. Trainees who were unable to demonstrate accuracy and consistency in scoring were not allowed to participate in 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 included:

- **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** began 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

In Auburn, 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, twenty-one 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 statistic 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 was computed, providing 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 how scoring reliability impacted total score correlations. The amount of shared variance for the total scores when the constructed-response items were scored using the local and audit scoring methods was obtained by squaring the correlation.

VII. Results

Item Means

The average score and standard deviation for each constructed-response question was computed based on the local scoring and the audit scoring. The results from this analysis, presented in Table 9, show a very close agreement between the local scoring and the audit scoring on the ELA constructed-response questions. Specifically, 38% of the items have exactly the same mean raw scores and 29% of the items have an absolute mean difference of 0.1.

Percent of Agreement

Table 9 contains the percent of agreement and the 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 3 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 percent of exact agreement ranged from 42.4 - 98.9%; the total percent of agreement ranges from 90.3% to a high of 99.9%. 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 strong because they explain more than one-third of the variance. Table 9 shows that all of the items had correlations greater or equal to 0.69. Furthermore, 48% of the items had correlations equal to or greater than 0.80. The intra-class correlations ranged from 0.69 to 0.94, showing a high degree of consistency between the local and audit scores.

Table 9. NYS Public Schools ELA Operational Test 2009: Inter-Rater Agreement

Grade	Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
					Exact	Approx.	Total	Local	Audit	Local	Audit		
3	21	Overall	2	13825	78.5	21.0	99.5	1.3	1.3	0.72	0.73	0.88	0.71
	26	Overall	2	13825	98.9	1.0	99.9	2.0	2.0	0.22	0.21	0.92	0.84
	27	Overall	2	13825	80.4	19.1	99.5	1.6	1.6	0.56	0.58	0.81	0.62
	28	Overall	3	13825	86.6	12.6	99.2	2.6	2.5	0.8	0.83	0.94	0.79
4	29–31	Overall	4	14444	53.5	44.1	97.6	2.6	2.4	0.92	0.71	0.75	0.45
	32–35	Overall	4	14444	50.7	45.0	95.8	2.5	2.3	0.98	0.80	0.76	0.46
	31&35	Overall	3	14444	59.8	38.5	98.3	2.1	2.0	0.77	0.74	0.75	0.48
5	21	Overall	2	14046	88.2	11.4	99.6	1.6	1.6	0.61	0.62	0.91	0.80
	26	Overall	2	14046	85.6	13.9	99.5	1.8	1.7	0.49	0.53	0.82	0.65
	27	Overall	3	14046	73.1	25.4	98.5	1.8	1.8	1.03	1.03	0.92	0.75
6	27–30	Overall	5	15263	42.4	49.5	91.8	3.7	3.3	1.01	0.91	0.73	0.40
	31–34	Overall	5	15263	42.8	47.5	90.3	3.4	3.2	1.11	1.02	0.76	0.43
	30&34	Overall	3	15263	60.9	37.7	98.6	2.3	2.4	0.69	0.67	0.69	0.42
7	27	Overall	2	12675	61.4	37.1	98.5	1.4	1.2	0.68	0.67	0.70	0.44
	28	Overall	2	12675	69.0	29.9	98.9	1.4	1.3	0.69	0.70	0.79	0.56
	33	Overall	2	12675	83.5	16.4	99.9	1.6	1.6	0.52	0.54	0.83	0.68
	34	Overall	2	12675	77.0	22.2	99.2	1.6	1.6	0.57	0.60	0.78	0.56
	35	Overall	3	12675	78.8	20.4	99.3	1.6	1.6	1.05	1.04	0.94	0.81
8	27–30	Overall	5	12676	43.1	47.7	90.8	3.5	3.1	1.18	1.03	0.80	0.48
	31–34	Overall	5	12676	45.2	46.0	91.2	3.6	3.5	1.11	1.03	0.77	0.45
	30&34	Overall	3	12676	57.8	40.1	97.8	2.3	2.1	0.73	0.72	0.70	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.

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:

- $k > 0.75$ denotes excellent reproducibility
- $0.4 \leq k \leq 0.75$ 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. In particular, 19% of the items had a weighted Kappa statistic which denoted excellent reproducibility. The remaining 17 items produced a weighted Kappa statistic denoting good 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 totals 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 ELA (Local Scoring Minus Audit Scoring)

			Difference				
Grade	Item	MAX Points	0	1	2	3	4 or more
3	21	2	79%	21%	0%	0%	0%
N=13825	26	2	99%	1%	0%	0%	0%
	27	2	80%	19%	0%	0%	0%
	28	3	87%	13%	0%	0%	0%
4	29–31	4	53%	44%	2%	0%	0%
N=14444	32–35	4	51%	45%	4%	0%	0%
	31&35	3	60%	39%	2%	0%	0%
5	21	2	88%	11%	0%	0%	0%
N=14046	26	2	86%	14%	0%	0%	0%
	27	3	73%	25%	1%	0%	0%
6	27–30	5	42%	49%	8%	0%	0%
N=15263	31–34	5	43%	47%	9%	1%	0%
	30&34	3	61%	38%	1%	0%	0%
7	27	2	61%	37%	1%	0%	0%
N=12675	28	2	69%	30%	1%	0%	0%
	33	2	84%	16%	0%	0%	0%
	34	2	77%	22%	1%	0%	0%
	35	3	79%	20%	1%	0%	0%
8	27–30	5	43%	48%	9%	0%	0%
N=12676	31–34	5	45%	46%	8%	1%	0%
	30&34	3	58%	40%	2%	0%	0%

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 of perfect agreement was lower, though most agreement was within one score point. A possible explanation for this 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 three items above 70% and two 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, two sets of total scores were computed. One total uses both the MC and CR items, and the second total uses only the CR items. 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 Using MC and CR Items		Total Score Using CR Items Open-ended Only	
	Correlation	Common Variance	Correlation	Common Variance
3	0.99	0.98	0.87	0.76
4	0.98	0.96	0.80	0.64
5	0.99	0.98	0.88	0.77
6	0.96	0.92	0.78	0.61
7	0.98	0.96	0.86	0.74
8	0.97	0.94	0.80	0.64

The correlations show a very high degree of consistency between the local and audit scoring results with correlations ranging from 0.96 to 0.99. Based on these correlations, the amount of common variance between local and audit scoring ranges from 0.92 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 total score using CR items only range from a low of 0.78 to a high of 0.88, and the common variance ranges from 0.61 to 0.77. 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 ELA 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 ELA audit scores are in Appendix B.

Appendices C, D, and E contain summary item-level information for the ELA assessments. Analyses are computed for all schools and then by scoring model, where 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
- Weighted 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 ELA. 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 seven 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's 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 0.96 to a high of 0.99. The correlations based on the constructed-response items only range from 0.78 to 0.88. 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 ELA, the largest mean difference between local and audit scoring was 0.4, which occurred in Grade 6, item 27 and Grade 8, item 27. Considering these are 5-point items, that difference only represents 8% of the maximum points. All other items had absolute mean differences of 0.2 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 0.7 in magnitude. This difference is based on a very small number of papers (N=41). The remaining differences were all less than or equal to 0.6, 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

ELA Item Analysis for Local Scoring

Local Scoring ELA Grade 3 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	Point-Biserial
21	O	0.00	0.15	0.38	0.48	0.00	0.00	1.33	0.62
26	O	0.00	0.00	0.03	0.97	0.00	0.00	1.96	0.24
27	O	0.00	0.04	0.29	0.67	0.00	0.00	1.63	0.46
28	O	0.00	0.05	0.06	0.14	0.75	0.00	2.60	0.49

Local Scoring ELA Grade 4 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	Point-Biserial
29–31	O	0.00	0.01	0.10	0.31	0.42	0.16	2.60	0.74
32–35	O	0.00	0.02	0.13	0.31	0.37	0.17	2.54	0.77
31&35	O	0.00	0.02	0.20	0.46	0.32	0.00	2.08	0.67

Local Scoring ELA Grade 5 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	Point-Biserial
21	O	0.00	0.06	0.31	0.62	0.00	0.00	1.56	0.61
26	O	0.00	0.03	0.19	0.78	0.00	0.00	1.75	0.42
27	O	0.00	0.15	0.21	0.34	0.30	0.00	1.80	0.67

Local Scoring ELA Grade 6 Item Statistics

Item	Key	B	0	1	2	3	4	5	Mean	Point-Biserial
27–30	O	0.00	0.00	0.02	0.09	0.27	0.37	0.24	3.72	0.75
31–34	O	0.00	0.00	0.05	0.14	0.31	0.32	0.18	3.42	0.76
30&34	O	0.00	0.01	0.12	0.47	0.40	0.00	0.00	2.27	0.64

Local Scoring ELA Grade 7 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	Point-Biserial
27	O	0.00	0.11	0.41	0.47	0.00	0.00	1.35	0.58
28	O	0.02	0.10	0.38	0.51	0.00	0.00	1.39	0.60
33	O	0.00	0.02	0.33	0.65	0.00	0.00	1.63	0.45
34	O	0.00	0.05	0.27	0.68	0.00	0.00	1.63	0.48
35	O	0.00	0.20	0.23	0.35	0.23	0.00	1.60	0.72

Local Scoring ELA Grade 8 Item Statistics

Item	Key	B	0	1	2	3	4	5	Mean	Point-Biserial
27–30	O	0	0.01	0.05	0.13	0.26	0.31	0.23	3.51	0.78
31–34	O	0	0.01	0.04	0.12	0.27	0.34	0.22	3.55	0.76
30&34	O	0	0.01	0.12	0.41	0.47	0.00	0.00	2.32	0.66

Appendix B

ELA Item Analysis for Audit Scoring

Audit Scoring ELA Grade 3 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	Point-Biserial
21	O	0.00	0.17	0.38	0.46	0.00	0.00	1.29	0.61
26	O	0.00	0.00	0.02	0.97	0.00	0.00	1.96	0.23
27	O	0.00	0.05	0.30	0.66	0.00	0.00	1.61	0.46
28	O	0.00	0.05	0.06	0.18	0.70	0.00	2.53	0.46

Audit Scoring ELA Grade 4 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	Point-Biserial
29–31	O	0.00	0.01	0.08	0.46	0.42	0.03	2.38	0.68
32–35	O	0.00	0.02	0.12	0.46	0.36	0.05	2.30	0.73
31&35	O	0.00	0.02	0.22	0.49	0.27	0.00	2.02	0.67

Audit Scoring ELA Grade 5 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	Point-Biserial
21	O	0.00	0.07	0.32	0.62	0.00	0.00	1.55	0.61
26	O	0.00	0.03	0.22	0.74	0.00	0.00	1.71	0.44
27	O	0.00	0.15	0.21	0.33	0.30	0.00	1.79	0.68

Audit Scoring ELA Grade 6 Item Statistics

Item	Key	B	0	1	2	3	4	5	Mean	Point-Biserial
27–30	O	0	0.00	0.00	0.03	0.14	0.44	0.31	3.27	0.73
31–34	O	0	0.00	0.01	0.04	0.18	0.39	0.28	3.19	0.74
30&34	O	0	0.01	0.09	0.43	0.48	0.00	0.00	2.38	0.62

Audit Scoring ELA Grade 7 Item Statistics

Item	Key	B	0	1	2	3	4	Mean	Point-Biserial
27	O	0.00	0.15	0.51	0.34	0.00	0.00	1.19	0.57
28	O	0.02	0.12	0.42	0.43	0.00	0.00	1.29	0.59
33	O	0.00	0.02	0.40	0.58	0.00	0.00	1.55	0.46
34	O	0.00	0.06	0.27	0.67	0.00	0.00	1.61	0.45
35	O	0.00	0.19	0.23	0.35	0.23	0.00	1.62	0.71

Audit Scoring ELA Grade 8 Item Statistics

Item	Key	B	0	1	2	3	4	5	Mean	Point-Biserial
27–30	O	0.00	0.01	0.05	0.19	0.39	0.27	0.09	3.13	0.75
31–34	O	0.00	0.00	0.03	0.13	0.34	0.33	0.17	3.46	0.72
30&34	O	0.00	0.01	0.17	0.51	0.30	0.00	0.00	2.11	0.66

Appendix C

Item Level Statistics for ELA Including All Schools in State

Table C–1. NYS Public Schools Grade 3 ELA Operational Test 2009: Inter-Rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
21	Overall	2	13825	78.5	21.0	99.5	1.3	1.3	0.72	0.73	0.88	0.71
	1	2	6228	78.0	21.7	99.6	1.3	1.2	0.73	0.75	0.88	0.71
	2	2	215	87.4	12.1	99.5	1.6	1.5	0.57	0.65	0.90	0.78
	3	2	4334	77.5	21.8	99.3	1.4	1.3	0.68	0.72	0.86	0.68
	4	2	593	79.9	19.9	99.8	1.5	1.4	0.68	0.69	0.88	0.71
	5	2	2455	80.6	18.9	99.5	1.3	1.3	0.73	0.74	0.89	0.74
26	Overall	2	13825	98.9	1.0	99.9	2.0	2.0	0.22	0.21	0.92	0.84
	1	2	6228	98.9	1.0	99.9	2.0	2.0	0.24	0.23	0.93	0.86
	2	2	215	99.1	0.9	100.0	2.0	2.0	0.20	0.18	0.93	0.83
	3	2	4334	98.9	1.0	99.9	2.0	2.0	0.20	0.19	0.91	0.81
	4	2	593	98.8	0.8	99.7	2.0	2.0	0.21	0.25	0.89	0.80
	5	2	2455	99.0	0.9	99.9	2.0	2.0	0.22	0.22	0.93	0.86
27	Overall	2	13825	80.4	19.1	99.5	1.6	1.6	0.56	0.58	0.81	0.62
	1	2	6228	79.2	20.4	99.6	1.6	1.6	0.58	0.59	0.81	0.62
	2	2	215	80.9	19.1	100.0	1.7	1.8	0.50	0.45	0.74	0.54
	3	2	4334	81.4	18.2	99.5	1.7	1.6	0.54	0.57	0.80	0.62
	4	2	593	83.1	16.9	100.0	1.7	1.7	0.50	0.52	0.81	0.61
	5	2	2455	81.1	18.2	99.2	1.6	1.6	0.57	0.58	0.81	0.63
28	Overall	3	13825	86.6	12.6	99.2	2.6	2.5	0.80	0.83	0.94	0.79
	1	3	6228	87.2	11.8	99.1	2.6	2.5	0.86	0.88	0.94	0.82
	2	3	215	93.5	6.0	99.5	2.8	2.8	0.43	0.49	0.90	0.78
	3	3	4334	87.1	12.1	99.2	2.7	2.6	0.72	0.76	0.93	0.77
	4	3	593	87.5	12.1	99.7	2.7	2.7	0.68	0.70	0.92	0.76
	5	3	2455	83.2	15.9	99.1	2.6	2.5	0.83	0.86	0.93	0.76

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. NYS Public Schools Grade 4 ELA Operational Test 2009: Inter-Rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
29–31	Overall	4	14444	53.5	44.1	97.6	2.6	2.4	0.92	0.71	0.75	0.45
	1	4	7216	54.0	43.5	97.5	2.4	2.3	0.93	0.70	0.75	0.45
	2	4	601	56.4	41.6	98.0	2.6	2.3	0.81	0.63	0.70	0.42
	3	4	4971	53.8	44.2	98.0	2.8	2.5	0.86	0.70	0.74	0.44
	4	4	729	53.2	43.5	96.7	3.0	2.8	0.85	0.60	0.66	0.37
	5	4	927	45.7	50.6	96.3	2.8	2.4	0.89	0.70	0.68	0.37
32–35	Overall	4	14444	50.7	45.0	95.8	2.5	2.3	0.98	0.80	0.76	0.46
	1	4	7216	52.8	43.7	96.5	2.4	2.2	0.99	0.81	0.79	0.49
	2	4	601	48.9	45.9	94.8	2.6	2.2	0.90	0.75	0.70	0.40
	3	4	4971	48.8	46.4	95.3	2.7	2.4	0.93	0.79	0.73	0.42
	4	4	729	49.5	45.0	94.5	2.9	2.6	0.87	0.74	0.68	0.39
	5	4	927	47.0	47.5	94.5	2.8	2.4	0.93	0.76	0.71	0.40
31&35	Overall	3	14444	59.8	38.5	98.3	2.1	2.0	0.77	0.74	0.75	0.48
	1	3	7216	59.7	38.5	98.2	2.0	1.9	0.77	0.75	0.76	0.48
	2	3	601	61.9	35.3	97.2	2.1	2.0	0.75	0.76	0.74	0.48
	3	3	4971	59.4	39.0	98.4	2.2	2.1	0.75	0.73	0.74	0.46
	4	3	729	57.5	41.0	98.5	2.3	2.2	0.73	0.68	0.69	0.40
	5	3	927	62.8	35.9	98.7	2.1	2.1	0.76	0.73	0.77	0.51

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. NYS Public Schools Grade 5 ELA Operational Test 2009: Inter-Rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
21	Overall	2	14046	88.2	11.4	99.6	1.6	1.6	0.61	0.62	0.91	0.80
	1	2	6397	87.0	12.6	99.6	1.5	1.5	0.64	0.65	0.91	0.79
	2	2	227	90.7	9.3	100.0	1.6	1.6	0.56	0.57	0.92	0.82
	3	2	3786	88.6	11.0	99.6	1.6	1.6	0.58	0.59	0.90	0.78
	4	2	1177	90.0	9.6	99.6	1.7	1.6	0.56	0.58	0.91	0.80
	5	2	2459	89.8	10.0	99.8	1.6	1.6	0.59	0.58	0.91	0.81
26	Overall	2	14046	85.6	13.9	99.5	1.8	1.7	0.49	0.53	0.82	0.65
	1	2	6397	84.3	14.9	99.2	1.7	1.7	0.53	0.56	0.82	0.65
	2	2	227	85.0	13.7	98.7	1.8	1.7	0.45	0.53	0.76	0.58
	3	2	3786	86.9	12.9	99.8	1.8	1.7	0.44	0.49	0.82	0.64
	4	2	1177	88.4	11.1	99.6	1.8	1.7	0.44	0.49	0.83	0.68
	5	2	2459	85.6	14.1	99.7	1.7	1.7	0.47	0.48	0.80	0.63
27	Overall	3	14046	73.1	25.4	98.5	1.8	1.8	1.03	1.03	0.92	0.75
	1	3	6397	72.1	26.2	98.4	1.7	1.7	1.07	1.07	0.92	0.75
	2	3	227	75.3	23.8	99.1	1.8	1.8	1.00	0.98	0.93	0.76
	3	3	3786	74.0	24.7	98.7	2.0	1.9	0.97	0.98	0.91	0.74
	4	3	1177	76.0	23.2	99.2	2.0	2.0	0.95	0.96	0.92	0.76
	5	3	2459	72.6	25.7	98.3	1.7	1.8	1.01	0.99	0.91	0.73

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. NYS Public Schools Grade 6 ELA Operational Test 2009: Inter-Rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
27–30	Overall	5	15263	42.4	49.5	91.8	3.7	3.3	1.01	0.91	0.73	0.40
	1	5	8250	42.2	49.6	91.7	3.6	3.2	1.03	0.91	0.73	0.41
	2	5	41	24.4	65.9	90.2	3.4	2.7	0.96	0.74	0.56	0.22
	3	5	1512	43.5	48.9	92.3	3.7	3.4	1.03	0.92	0.74	0.42
	4	5	1191	38.8	51.9	90.7	3.9	3.4	0.93	0.90	0.66	0.34
	5	5	4269	43.5	48.6	92.2	3.9	3.4	0.94	0.89	0.71	0.39
31–34	Overall	5	15263	42.8	47.5	90.3	3.4	3.2	1.11	1.02	0.76	0.43
	1	5	8250	42.6	48.0	90.6	3.3	3.1	1.11	1.00	0.76	0.42
	2	5	41	56.1	39.0	95.1	2.5	2.8	0.89	0.97	0.81	0.54
	3	5	1512	45.4	46.3	91.7	3.4	3.4	1.12	1.06	0.79	0.47
	4	5	1191	40.8	46.9	87.7	3.7	3.4	1.06	1.08	0.72	0.39
	5	5	4269	42.8	47.0	89.9	3.7	3.3	1.04	0.99	0.73	0.41
30&34	Overall	3	15263	60.9	37.7	98.6	2.3	2.4	0.69	0.67	0.69	0.42
	1	3	8250	60.0	38.4	98.4	2.2	2.3	0.70	0.67	0.69	0.42
	2	3	41	56.1	41.5	97.6	2.1	1.9	0.62	0.60	0.50	0.25
	3	3	1512	59.8	38.8	98.5	2.3	2.4	0.68	0.67	0.68	0.40
	4	3	1191	62.6	35.8	98.4	2.4	2.4	0.66	0.69	0.70	0.43
	5	3	4269	62.8	36.4	99.1	2.3	2.5	0.65	0.64	0.69	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. NYS Public Schools Grade 7 ELA Operational Test 2009: Inter-Rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
27	Overall	2	12675	61.4	37.1	98.5	1.4	1.2	0.68	0.67	0.70	0.44
	1	2	6197	62.7	35.6	98.3	1.3	1.1	0.70	0.69	0.72	0.47
	2	2	86	64.0	34.9	98.8	0.9	1.0	0.75	0.67	0.76	0.51
	3	2	933	63.8	35.7	99.5	1.2	1.1	0.70	0.69	0.76	0.50
	4	2	1785	61.5	37.0	98.5	1.6	1.4	0.58	0.63	0.61	0.38
	5	2	3674	58.5	40.1	98.6	1.4	1.2	0.65	0.63	0.63	0.37
28	Overall	2	12675	69.0	29.9	98.9	1.4	1.3	0.69	0.70	0.79	0.56
	1	2	6197	68.1	30.7	98.8	1.3	1.2	0.72	0.73	0.80	0.57
	2	2	86	75.6	24.4	100.0	1.1	1.1	0.67	0.63	0.83	0.63
	3	2	933	70.0	29.0	99.0	1.2	1.2	0.70	0.71	0.80	0.59
	4	2	1785	69.9	28.9	98.8	1.7	1.5	0.56	0.63	0.69	0.47
	5	2	3674	69.8	29.3	99.1	1.4	1.4	0.65	0.66	0.77	0.54
33	Overall	2	12675	83.5	16.4	99.9	1.6	1.6	0.52	0.54	0.83	0.68
	1	2	6197	84.4	15.5	99.9	1.6	1.5	0.54	0.55	0.85	0.71
	2	2	86	82.6	17.4	100.0	1.4	1.4	0.56	0.53	0.83	0.68
	3	2	933	80.5	19.4	99.9	1.5	1.5	0.54	0.56	0.80	0.64
	4	2	1785	84.1	15.8	99.9	1.7	1.6	0.46	0.51	0.80	0.65
	5	2	3674	82.4	17.5	99.9	1.7	1.6	0.48	0.53	0.79	0.64
34	Overall	2	12675	77.0	22.2	99.2	1.6	1.6	0.57	0.60	0.78	0.56
	1	2	6197	76.5	22.7	99.1	1.6	1.6	0.61	0.61	0.79	0.57
	2	2	86	70.9	26.7	97.7	1.4	1.4	0.64	0.72	0.76	0.55
	3	2	933	68.8	30.0	98.8	1.5	1.4	0.62	0.66	0.73	0.50
	4	2	1785	81.6	17.7	99.3	1.8	1.7	0.47	0.50	0.72	0.51
	5	2	3674	78.1	21.4	99.5	1.7	1.6	0.54	0.58	0.77	0.56
35	Overall	3	12675	78.8	20.4	99.3	1.6	1.6	1.05	1.04	0.94	0.81
	1	3	6197	78.6	20.7	99.3	1.5	1.5	1.06	1.07	0.94	0.81
	2	3	86	88.4	11.6	100.0	1.4	1.4	0.97	0.99	0.97	0.89
	3	3	933	79.4	19.7	99.1	1.3	1.4	1.06	1.04	0.94	0.82
	4	3	1785	78.0	21.2	99.2	1.9	1.9	0.97	0.98	0.93	0.78
	5	3	3674	79.3	20.0	99.3	1.7	1.8	1.00	0.98	0.94	0.80

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. NYS Public Schools Grade 8 ELA Operational Test 2009: Inter-Rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
27–30	Overall	5	12676	43.1	47.7	90.8	3.5	3.1	1.18	1.03	0.80	0.48
	1	5	6251	45.9	46.7	92.7	3.3	3.0	1.23	1.07	0.83	0.52
	2	5	156	48.1	44.2	92.3	3.6	3.1	1.04	0.95	0.77	0.48
	3	5	1586	42.9	47.7	90.5	3.5	3.0	1.29	1.09	0.83	0.52
	4	5	1599	41.5	50.2	91.7	3.8	3.3	0.94	0.91	0.70	0.38
	5	5	3084	38.2	48.4	86.6	3.9	3.3	0.99	0.96	0.66	0.36
31–34	Overall	5	12676	45.2	46.0	91.2	3.6	3.5	1.11	1.03	0.77	0.45
	1	5	6251	44.4	46.1	90.5	3.4	3.4	1.15	1.04	0.78	0.45
	2	5	156	41.0	52.6	93.6	3.8	3.5	0.93	0.90	0.69	0.34
	3	5	1586	43.9	44.8	88.7	3.6	3.3	1.24	1.09	0.79	0.48
	4	5	1599	48.8	44.8	93.7	3.8	3.6	0.89	0.96	0.74	0.43
	5	5	3084	45.9	46.4	92.3	3.8	3.6	0.98	1.00	0.75	0.43
30&34	Overall	3	12676	57.8	40.1	97.8	2.3	2.1	0.73	0.72	0.70	0.43
	1	3	6251	59.6	38.4	98.0	2.2	2.0	0.75	0.72	0.73	0.46
	2	3	156	65.4	34.0	99.4	2.3	2.2	0.63	0.67	0.73	0.48
	3	3	1586	57.9	40.1	98.0	2.2	2.0	0.81	0.73	0.75	0.47
	4	3	1599	49.0	49.1	98.1	2.5	2.1	0.59	0.65	0.51	0.27
	5	3	3084	58.1	39.1	97.2	2.5	2.3	0.61	0.71	0.62	0.37

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 ELA Including All Schools in State

Without New York City Schools

Table D–1. NYS Public Schools (Without NYC) Grade 3 ELA Operational Test 2009: Inter-Rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
21	Overall	2	8906	79.5	20.0	99.5	1.4	1.3	0.70	0.72	0.88	0.72
	1	2	1309	82.7	17.1	99.8	1.3	1.3	0.72	0.72	0.91	0.77
	2	2	215	87.4	12.1	99.5	1.6	1.5	0.57	0.65	0.90	0.78
	3	2	4334	77.5	21.8	99.3	1.4	1.3	0.68	0.72	0.86	0.68
	4	2	593	79.9	19.9	99.8	1.5	1.4	0.68	0.69	0.88	0.71
	5	2	2455	80.6	18.9	99.5	1.3	1.3	0.73	0.74	0.89	0.74
26	Overall	2	8906	98.9	1.0	99.9	2.0	2.0	0.21	0.21	0.92	0.84
	1	2	1309	98.5	1.5	100.0	1.9	2.0	0.26	0.25	0.94	0.85
	2	2	215	99.1	0.9	100.0	2.0	2.0	0.20	0.18	0.93	0.83
	3	2	4334	98.9	1.0	99.9	2.0	2.0	0.20	0.19	0.91	0.81
	4	2	593	98.8	0.8	99.7	2.0	2.0	0.21	0.25	0.89	0.80
	5	2	2455	99.0	0.9	99.9	2.0	2.0	0.22	0.22	0.93	0.86
27	Overall	2	8906	81.9	17.7	99.5	1.7	1.6	0.55	0.57	0.81	0.64
	1	2	1309	84.6	15.2	99.8	1.6	1.6	0.57	0.59	0.86	0.71
	2	2	215	80.9	19.1	100.0	1.7	1.8	0.50	0.45	0.74	0.54
	3	2	4334	81.4	18.2	99.5	1.7	1.6	0.54	0.57	0.80	0.62
	4	2	593	83.1	16.9	100.0	1.7	1.7	0.50	0.52	0.81	0.61
	5	2	2455	81.1	18.2	99.2	1.6	1.6	0.57	0.58	0.81	0.63
28	Overall	3	8906	86.8	12.4	99.2	2.6	2.6	0.76	0.80	0.93	0.78
	1	3	1309	91.3	8.0	99.3	2.6	2.6	0.82	0.83	0.96	0.86
	2	3	215	93.5	6.0	99.5	2.8	2.8	0.43	0.49	0.90	0.78
	3	3	4334	87.1	12.1	99.2	2.7	2.6	0.72	0.76	0.93	0.77
	4	3	593	87.5	12.1	99.7	2.7	2.7	0.68	0.70	0.92	0.76
	5	3	2455	83.2	15.9	99.1	2.6	2.5	0.83	0.86	0.93	0.76

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. NYS Public Schools (Without NYC) Grade 4 ELA Operational Test 2009: Inter-Rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
29–31	Overall	4	9211	53.3	44.4	97.7	2.7	2.5	0.88	0.69	0.74	0.43
	1	4	1983	54.7	43.0	97.6	2.5	2.4	0.90	0.69	0.74	0.43
	2	4	601	56.4	41.6	98.0	2.6	2.3	0.81	0.63	0.70	0.42
	3	4	4971	53.8	44.2	98.0	2.8	2.5	0.86	0.70	0.74	0.44
	4	4	729	53.2	43.5	96.7	3.0	2.8	0.85	0.60	0.66	0.37
	5	4	927	45.7	50.6	96.3	2.8	2.4	0.89	0.70	0.68	0.37
32–35	Overall	4	9211	49.5	45.8	95.4	2.7	2.4	0.93	0.79	0.74	0.43
	1	4	1983	52.6	43.7	96.4	2.5	2.3	0.94	0.80	0.77	0.47
	2	4	601	48.9	45.9	94.8	2.6	2.2	0.90	0.75	0.70	0.40
	3	4	4971	48.8	46.4	95.3	2.7	2.4	0.93	0.79	0.73	0.42
	4	4	729	49.5	45.0	94.5	2.9	2.6	0.87	0.74	0.68	0.39
	5	4	927	47.0	47.5	94.5	2.8	2.4	0.93	0.76	0.71	0.40
31&35	Overall	3	9211	60.0	38.3	98.4	2.1	2.1	0.76	0.73	0.75	0.47
	1	3	1983	60.8	37.7	98.4	2.0	2.0	0.76	0.74	0.76	0.49
	2	3	601	61.9	35.3	97.2	2.1	2.0	0.75	0.76	0.74	0.48
	3	3	4971	59.4	39.0	98.4	2.2	2.1	0.75	0.73	0.74	0.46
	4	3	729	57.5	41.0	98.5	2.3	2.2	0.73	0.68	0.69	0.40
	5	3	927	62.8	35.9	98.7	2.1	2.1	0.76	0.73	0.77	0.51

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. NYS Public Schools (Without NYC) Grade 5 ELA Operational Test 2009: Inter-Rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
21	Overall	2	9094	89.5	10.2	99.7	1.6	1.6	0.58	0.59	0.91	0.80
	1	2	1445	90.8	8.9	99.7	1.6	1.6	0.61	0.60	0.93	0.83
	2	2	227	90.7	9.3	100.0	1.6	1.6	0.56	0.57	0.92	0.82
	3	2	3786	88.6	11.0	99.6	1.6	1.6	0.58	0.59	0.90	0.78
	4	2	1177	90.0	9.6	99.6	1.7	1.6	0.56	0.58	0.91	0.80
	5	2	2459	89.8	10.0	99.8	1.6	1.6	0.59	0.58	0.91	0.81
26	Overall	2	9094	86.5	13.1	99.6	1.8	1.7	0.46	0.50	0.81	0.65
	1	2	1445	85.7	13.6	99.2	1.7	1.7	0.49	0.52	0.80	0.65
	2	2	227	85.0	13.7	98.7	1.8	1.7	0.45	0.53	0.76	0.58
	3	2	3786	86.9	12.9	99.8	1.8	1.7	0.44	0.49	0.82	0.64
	4	2	1177	88.4	11.1	99.6	1.8	1.7	0.44	0.49	0.83	0.68
	5	2	2459	85.6	14.1	99.7	1.7	1.7	0.47	0.48	0.80	0.63
27	Overall	3	9094	74.1	24.6	98.7	1.9	1.9	0.99	0.99	0.92	0.75
	1	3	1445	75.2	23.7	98.9	1.9	1.9	0.98	1.01	0.92	0.76
	2	3	227	75.3	23.8	99.1	1.8	1.8	1.00	0.98	0.93	0.76
	3	3	3786	74.0	24.7	98.7	2.0	1.9	0.97	0.98	0.91	0.74
	4	3	1177	76.0	23.2	99.2	2.0	2.0	0.95	0.96	0.92	0.76
	5	3	2459	72.6	25.7	98.3	1.7	1.8	1.01	0.99	0.91	0.73

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. NYS Public Schools (Without NYC) Grade 6 ELA Operational Test 2009: Inter-Rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
27–30	Overall	5	9814	42.6	49.5	92.1	3.8	3.3	0.98	0.91	0.72	0.40
	1	5	2801	42.8	49.8	92.5	3.6	3.2	0.99	0.92	0.74	0.41
	2	5	41	24.4	65.9	90.2	3.4	2.7	0.96	0.74	0.56	0.22
	3	5	1512	43.5	48.9	92.3	3.7	3.4	1.03	0.92	0.74	0.42
	4	5	1191	38.8	51.9	90.7	3.9	3.4	0.93	0.90	0.66	0.34
	5	5	4269	43.5	48.6	92.2	3.9	3.4	0.94	0.89	0.71	0.39
31–34	Overall	5	9814	43.5	47.0	90.5	3.5	3.2	1.09	1.03	0.76	0.43
	1	5	2801	44.3	47.4	91.8	3.2	3.0	1.08	1.01	0.77	0.45
	2	5	41	56.1	39.0	95.1	2.5	2.8	0.89	0.97	0.81	0.54
	3	5	1512	45.4	46.3	91.7	3.4	3.4	1.12	1.06	0.79	0.47
	4	5	1191	40.8	46.9	87.7	3.7	3.4	1.06	1.08	0.72	0.39
	5	5	4269	42.8	47.0	89.9	3.7	3.3	1.04	0.99	0.73	0.41
30&34	Overall	3	9814	62.0	36.8	98.8	2.3	2.4	0.67	0.68	0.71	0.44
	1	3	2801	62.0	36.6	98.6	2.2	2.3	0.69	0.71	0.73	0.47
	2	3	41	56.1	41.5	97.6	2.1	1.9	0.62	0.60	0.50	0.25
	3	3	1512	59.8	38.8	98.5	2.3	2.4	0.68	0.67	0.68	0.40
	4	3	1191	62.6	35.8	98.4	2.4	2.4	0.66	0.69	0.70	0.43
	5	3	4269	62.8	36.4	99.1	2.3	2.5	0.65	0.64	0.69	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. NYS Public Schools (Without NYC) Grade 7 ELA Operational Test 2009: Inter-Rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
27	Overall	2	8764	60.7	37.8	98.5	1.4	1.2	0.67	0.66	0.68	0.42
	1	2	2286	62.3	35.7	98.1	1.3	1.2	0.70	0.68	0.71	0.46
	2	2	86	64.0	34.9	98.8	0.9	1.0	0.75	0.67	0.76	0.51
	3	2	933	63.8	35.7	99.5	1.2	1.1	0.70	0.69	0.76	0.50
	4	2	1785	61.5	37.0	98.5	1.6	1.4	0.58	0.63	0.61	0.38
	5	2	3674	58.5	40.1	98.6	1.4	1.2	0.65	0.63	0.63	0.37
28	Overall	2	8764	69.8	29.2	99.0	1.4	1.3	0.67	0.68	0.78	0.55
	1	2	2286	69.4	29.5	98.9	1.3	1.3	0.70	0.71	0.79	0.57
	2	2	86	75.6	24.4	100.0	1.1	1.1	0.67	0.63	0.83	0.63
	3	2	933	70.0	29.0	99.0	1.2	1.2	0.70	0.71	0.80	0.59
	4	2	1785	69.9	28.9	98.8	1.7	1.5	0.56	0.63	0.69	0.47
	5	2	3674	69.8	29.3	99.1	1.4	1.4	0.65	0.66	0.77	0.54
33	Overall	2	8764	82.9	17.0	99.9	1.7	1.6	0.51	0.53	0.81	0.66
	1	2	2286	83.9	16.1	99.9	1.6	1.5	0.54	0.54	0.84	0.70
	2	2	86	82.6	17.4	100.0	1.4	1.4	0.56	0.53	0.83	0.68
	3	2	933	80.5	19.4	99.9	1.5	1.5	0.54	0.56	0.80	0.64
	4	2	1785	84.1	15.8	99.9	1.7	1.6	0.46	0.51	0.80	0.65
	5	2	3674	82.4	17.5	99.9	1.7	1.6	0.48	0.53	0.79	0.64
34	Overall	2	8764	76.9	22.3	99.2	1.6	1.6	0.57	0.60	0.77	0.56
	1	2	2286	74.9	23.8	98.7	1.6	1.6	0.63	0.63	0.78	0.56
	2	2	86	70.9	26.7	97.7	1.4	1.4	0.64	0.72	0.76	0.55
	3	2	933	68.8	30.0	98.8	1.5	1.4	0.62	0.66	0.73	0.50
	4	2	1785	81.6	17.7	99.3	1.8	1.7	0.47	0.50	0.72	0.51
	5	2	3674	78.1	21.4	99.5	1.7	1.6	0.54	0.58	0.77	0.56

Table D–5 NYS Public Schools (Without NYC) Grade 7 ELA Operational Test 2009: Inter-Rater Agreement (continued)

Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
35	Overall	3	8764	79.1	20.2	99.3	1.7	1.7	1.03	1.01	0.94	0.81
	1	3	2286	79.3	20.3	99.6	1.5	1.6	1.04	1.04	0.95	0.82
	2	3	86	88.4	11.6	100.0	1.4	1.4	0.97	0.99	0.97	0.89
	3	3	933	79.4	19.7	99.1	1.3	1.4	1.06	1.04	0.94	0.82
	4	3	1785	78.0	21.2	99.2	1.9	1.9	0.97	0.98	0.93	0.78
	5	3	3674	79.3	20.0	99.3	1.7	1.8	1.00	0.98	0.94	0.80

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. NYS Public Schools (Without NYC) Grade 8 ELA Operational Test 2009: Inter-Rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
27–30	Overall	5	7836	41.4	48.4	89.8	3.7	3.3	1.07	0.99	0.74	0.42
	1	5	1411	45.9	47.6	93.4	3.6	3.4	0.99	0.96	0.76	0.43
	2	5	156	48.1	44.2	92.3	3.6	3.1	1.04	0.95	0.77	0.48
	3	5	1586	42.9	47.7	90.5	3.5	3.0	1.29	1.09	0.83	0.52
	4	5	1599	41.5	50.2	91.7	3.8	3.3	0.94	0.91	0.70	0.38
	5	5	3084	38.2	48.4	86.6	3.9	3.3	0.99	0.96	0.66	0.36
31–34	Overall	5	7836	46.4	45.7	92.0	3.7	3.5	1.03	1.01	0.77	0.45
	1	5	1411	47.8	45.2	93.1	3.6	3.5	1.00	0.99	0.77	0.45
	2	5	156	41.0	52.6	93.6	3.8	3.5	0.93	0.90	0.69	0.34
	3	5	1586	43.9	44.8	88.7	3.6	3.3	1.24	1.09	0.79	0.48
	4	5	1599	48.8	44.8	93.7	3.8	3.6	0.89	0.96	0.74	0.43
	5	5	3084	45.9	46.4	92.3	3.8	3.6	0.98	1.00	0.75	0.43
30&34	Overall	3	7836	56.9	41.0	97.8	2.4	2.2	0.67	0.70	0.66	0.39
	1	3	1411	61.0	37.5	98.5	2.4	2.2	0.63	0.67	0.66	0.41
	2	3	156	65.4	34.0	99.4	2.3	2.2	0.63	0.67	0.73	0.48
	3	3	1586	57.9	40.1	98.0	2.2	2.0	0.81	0.73	0.75	0.47
	4	3	1599	49.0	49.1	98.1	2.5	2.1	0.59	0.65	0.51	0.27
	5	3	3084	58.1	39.1	97.2	2.5	2.3	0.61	0.71	0.62	0.37

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 ELA Including New York City Schools Only

Table E–1. NYC Public Schools Grades 3–8 ELA Operational Test 2009: Inter-Rater Agreement

Item #	Scoring Model	Score Points	Total N	Agreement (%)			Raw Score Mean		Raw Score Standard Deviation		Intra-Class Correlation	Weighted Kappa
				Exact	Approx.	Total	Local	Audit	Local	Audit		
Grade 3												
21	NYC Overall	2	4919	76.7	22.9	99.6	1.2	1.2	0.74	0.75	0.88	0.70
26	NYC Overall	2	4919	99.0	0.9	99.9	2.0	2.0	0.24	0.22	0.93	0.86
27	NYC Overall	2	4919	77.7	21.8	99.5	1.6	1.6	0.58	0.59	0.79	0.59
28	NYC Overall	3	4919	86.1	12.9	99.0	2.5	2.5	0.87	0.89	0.94	0.81
Grade 4												
29–31	NYC Overall	4	5233	53.7	43.7	97.4	2.4	2.2	0.93	0.70	0.75	0.45
32–35	NYC Overall	4	5233	52.8	43.6	96.5	2.3	2.2	1.00	0.81	0.79	0.49
31&35	NYC Overall	3	5233	59.3	38.8	98.1	2.0	1.9	0.78	0.75	0.75	0.48
Grade 5												
21	NYC Overall	2	4952	85.9	13.7	99.6	1.5	1.4	0.65	0.66	0.90	0.78
26	NYC Overall	2	4952	83.9	15.3	99.2	1.7	1.7	0.54	0.58	0.83	0.65
27	NYC Overall	3	4952	71.2	27.0	98.2	1.6	1.6	1.08	1.08	0.92	0.75
Grade 6												
27–30	NYC Overall	5	5449	41.8	49.5	91.3	3.6	3.2	1.05	0.91	0.73	0.41
31–34	NYC Overall	5	5449	41.7	48.3	90.0	3.3	3.1	1.12	1.00	0.75	0.41
30&34	NYC Overall	3	5449	58.9	39.4	98.3	2.2	2.4	0.71	0.64	0.67	0.40
Grade 7												
27	NYC Overall	2	3911	62.9	35.5	98.4	1.3	1.1	0.70	0.70	0.73	0.48
28	NYC Overall	2	3911	67.3	31.4	98.7	1.3	1.2	0.73	0.73	0.80	0.57
33	NYC Overall	2	3911	84.8	15.2	99.9	1.6	1.5	0.55	0.55	0.85	0.72
34	NYC Overall	2	3911	77.4	22.0	99.4	1.6	1.6	0.59	0.59	0.79	0.57
35	NYC Overall	3	3911	78.2	20.9	99.1	1.4	1.4	1.07	1.08	0.94	0.81
Grade 8												
27–30	NYC Overall	5	4840	46.0	46.5	92.5	3.1	2.9	1.27	1.07	0.84	0.53
31–34	NYC Overall	5	4840	43.4	46.4	89.8	3.3	3.4	1.18	1.05	0.77	0.45
30&34	NYC Overall	3	4840	59.2	38.6	97.8	2.1	2.0	0.77	0.73	0.73	0.46

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 ELA Including All Schools in State

Table F–1. NYS Public Schools Grade 3 ELA Operational Test 2009: Percentages 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	0.79	0.13	0.00	0.00		
	1			0.00	0.00	0.10	0.78	0.12	0.00	0.00		
	2			0.00	0.00	0.02	0.87	0.10	0.00	0.00		
	3			0.00	0.00	0.08	0.78	0.14	0.00	0.00		
	4			0.00	0.00	0.07	0.80	0.13	0.00	0.00		
	5			0.00	0.00	0.07	0.81	0.12	0.00	0.00		
26	Overall			0.00	0.00	0.01	0.99	0.00	0.00	0.00		
	1			0.00	0.00	0.01	0.99	0.00	0.00	0.00		
	2			0.00	0.00	0.01	0.99	0.00	0.00	0.00		
	3			0.00	0.00	0.01	0.99	0.00	0.00	0.00		
	4			0.00	0.00	0.00	0.99	0.01	0.00	0.00		
	5			0.00	0.00	0.01	0.99	0.00	0.00	0.00		
27	Overall			0.00	0.00	0.09	0.80	0.10	0.00	0.00		
	1			0.00	0.00	0.10	0.79	0.11	0.00	0.00		
	2			0.00	0.00	0.13	0.81	0.06	0.00	0.00		
	3			0.00	0.00	0.07	0.81	0.11	0.00	0.00		
	4			0.00	0.00	0.07	0.83	0.10	0.00	0.00		
	5			0.00	0.00	0.09	0.81	0.09	0.00	0.00		
28	Overall			0.00	0.00	0.03	0.87	0.10	0.00	0.00		
	1			0.00	0.00	0.03	0.87	0.09	0.00	0.00		
	2			0.00	0.00	0.02	0.93	0.04	0.00	0.00		
	3			0.00	0.00	0.03	0.87	0.09	0.01	0.00		
	4			0.00	0.00	0.04	0.88	0.08	0.00	0.00		
	5			0.00	0.00	0.04	0.83	0.12	0.01	0.00		

Table F–2. NYS Public Schools Grade 4 ELA Operational Test 2009: Percentages 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	0.00	0.13	0.53	0.32	0.02	0.00	0.00	
	1		0.00	0.00	0.01	0.16	0.54	0.27	0.02	0.00	0.00	
	2		0.00	0.00	0.01	0.09	0.56	0.32	0.01	0.00	0.00	
	3		0.00	0.00	0.00	0.09	0.54	0.36	0.02	0.00	0.00	
	4		0.00	0.00	0.00	0.12	0.53	0.31	0.03	0.00	0.00	
	5		0.00	0.00	0.00	0.09	0.46	0.41	0.04	0.00	0.00	
32–35	Overall		0.00	0.00	0.01	0.13	0.51	0.32	0.04	0.00	0.00	
	1		0.00	0.00	0.01	0.17	0.53	0.27	0.03	0.00	0.00	
	2		0.00	0.00	0.00	0.10	0.49	0.36	0.05	0.00	0.00	
	3		0.00	0.00	0.00	0.10	0.49	0.36	0.04	0.00	0.00	
	4		0.00	0.00	0.00	0.08	0.50	0.37	0.05	0.00	0.00	
	5		0.00	0.00	0.00	0.09	0.47	0.39	0.05	0.00	0.00	
31&35	Overall		0.00	0.00	0.01	0.17	0.60	0.22	0.01	0.00	0.00	
	1		0.00	0.00	0.01	0.17	0.60	0.21	0.01	0.00	0.00	
	2		0.00	0.00	0.01	0.15	0.62	0.20	0.02	0.00	0.00	
	3		0.00	0.00	0.01	0.16	0.59	0.23	0.01	0.00	0.00	
	4		0.00	0.00	0.01	0.16	0.57	0.26	0.01	0.00	0.00	
	5		0.00	0.00	0.01	0.16	0.63	0.20	0.01	0.00	0.00	

Table F–3. NYS Public Schools Grade 5 ELA Operational Test 2009: Percentages 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	0.05	0.88	0.06	0.00	0.00		
	1			0.00	0.00	0.06	0.87	0.07	0.00	0.00		
	2			0.00	0.00	0.04	0.91	0.06	0.00	0.00		
	3			0.00	0.00	0.05	0.89	0.06	0.00	0.00		
	4			0.00	0.00	0.03	0.90	0.07	0.00	0.00		
	5			0.00	0.00	0.05	0.90	0.05	0.00	0.00		
26	Overall			0.00	0.00	0.05	0.86	0.09	0.00	0.00		
	1			0.00	0.00	0.05	0.84	0.10	0.00	0.00		
	2			0.00	0.00	0.04	0.85	0.10	0.01	0.00		
	3			0.00	0.00	0.04	0.87	0.09	0.00	0.00		
	4			0.00	0.00	0.03	0.88	0.08	0.00	0.00		
	5			0.00	0.00	0.07	0.86	0.07	0.00	0.00		
27	Overall			0.00	0.01	0.12	0.73	0.13	0.01	0.00		
	1			0.00	0.01	0.13	0.72	0.13	0.01	0.00		
	2			0.00	0.00	0.09	0.75	0.15	0.00	0.00		
	3			0.00	0.01	0.11	0.74	0.14	0.01	0.00		
	4			0.00	0.00	0.11	0.76	0.12	0.00	0.00		
	5			0.00	0.01	0.15	0.73	0.11	0.00	0.00		

Table F–4. NYS Public Schools Grade 6 ELA Operational Test 2009: Percentages 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	0.09	0.42	0.40	0.07	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.01	0.09	0.42	0.40	0.07	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.10	0.24	0.56	0.10	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.01	0.12	0.43	0.37	0.06	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.01	0.08	0.39	0.44	0.08	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.01	0.08	0.44	0.40	0.07	0.00	0.00	0.00
31–34	Overall	0.00	0.00	0.00	0.02	0.17	0.43	0.30	0.07	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.03	0.18	0.43	0.30	0.06	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.05	0.29	0.56	0.10	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.04	0.21	0.45	0.25	0.04	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.02	0.15	0.41	0.31	0.10	0.01	0.00	0.00
	5	0.00	0.00	0.00	0.02	0.14	0.43	0.33	0.08	0.00	0.00	0.00
30&34	Overall	0.00	0.00	0.00	0.01	0.24	0.61	0.14	0.00	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.01	0.25	0.60	0.14	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.02	0.10	0.56	0.32	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.01	0.22	0.60	0.16	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.01	0.17	0.63	0.19	0.01	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.01	0.24	0.63	0.12	0.00	0.00	0.00	0.00

Table F–5. NYS Public Schools Grade 7 ELA Operational Test 2009: Percentages 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	0.11	0.61	0.26	0.01	0.00		
	1			0.00	0.00	0.12	0.63	0.24	0.01	0.00		
	2			0.00	0.01	0.21	0.64	0.14	0.00	0.00		
	3			0.00	0.00	0.14	0.64	0.22	0.00	0.00		
	4			0.00	0.00	0.08	0.61	0.29	0.01	0.00		
	5			0.00	0.00	0.12	0.58	0.28	0.01	0.00		
28	Overall			0.00	0.00	0.11	0.69	0.19	0.01	0.00		
	1			0.00	0.00	0.11	0.68	0.19	0.01	0.00		
	2			0.00	0.00	0.10	0.76	0.14	0.00	0.00		
	3			0.00	0.01	0.12	0.70	0.17	0.00	0.00		
	4			0.00	0.00	0.06	0.70	0.23	0.01	0.00		
	5			0.00	0.00	0.12	0.70	0.18	0.01	0.00		
33	Overall			0.00	0.00	0.04	0.84	0.12	0.00	0.00		
	1			0.00	0.00	0.05	0.84	0.10	0.00	0.00		
	2			0.00	0.00	0.05	0.83	0.13	0.00	0.00		
	3			0.00	0.00	0.07	0.80	0.13	0.00	0.00		
	4			0.00	0.00	0.03	0.84	0.13	0.00	0.00		
	5			0.00	0.00	0.03	0.82	0.14	0.00	0.00		
34	Overall			0.00	0.00	0.10	0.77	0.12	0.01	0.00		
	1			0.00	0.00	0.12	0.76	0.11	0.01	0.00		
	2			0.00	0.00	0.14	0.71	0.13	0.02	0.00		
	3			0.00	0.00	0.10	0.69	0.20	0.01	0.00		
	4			0.00	0.00	0.07	0.82	0.11	0.00	0.00		
	5			0.00	0.00	0.09	0.78	0.12	0.00	0.00		
35	Overall			0.00	0.00	0.11	0.79	0.10	0.00	0.00		
	1			0.00	0.01	0.10	0.79	0.11	0.00	0.00		
	2			0.00	0.00	0.06	0.88	0.06	0.00	0.00		
	3			0.00	0.01	0.12	0.79	0.07	0.00	0.00		
	4			0.00	0.00	0.10	0.78	0.11	0.01	0.00		
	5			0.00	0.00	0.12	0.79	0.08	0.00	0.00		

Table F–6. NYS Public Schools Grade 8 ELA Operational Test 2009: Percentages 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	0.43	0.36	0.08	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.02	0.15	0.46	0.31	0.05	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.04	0.48	0.40	0.07	0.01	0.00	0.00
	3	0.00	0.00	0.00	0.01	0.10	0.43	0.38	0.08	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00	0.09	0.42	0.41	0.08	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00	0.07	0.38	0.41	0.12	0.01	0.00	0.00
31–34	Overall	0.00	0.00	0.00	0.03	0.19	0.45	0.26	0.05	0.00	0.00	0.00
	1	0.00	0.00	0.01	0.05	0.23	0.44	0.23	0.04	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.19	0.41	0.34	0.06	0.01	0.00	0.00
	3	0.00	0.00	0.00	0.03	0.17	0.44	0.28	0.07	0.01	0.00	0.00
	4	0.00	0.00	0.00	0.01	0.15	0.49	0.30	0.05	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.02	0.15	0.46	0.31	0.05	0.00	0.00	0.00
30&34	Overall	0.00	0.00	0.00	0.00	0.10	0.58	0.30	0.02	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.01	0.12	0.60	0.27	0.01	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.17	0.65	0.17	0.01	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.00	0.13	0.58	0.27	0.02	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00	0.05	0.49	0.44	0.02	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00	0.08	0.58	0.31	0.03	0.00	0.00	0.00

Appendix G

Item Level Differences for ELA Including All Schools in State

Without New York City Schools

Table G–1. NYS Public Schools (Without NYC) Grade 3 ELA Operational Test 2009: Percentages 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	0.80	0.13	0.00	0.00		
	1			0.00	0.00	0.07	0.83	0.10	0.00	0.00		
	2			0.00	0.00	0.02	0.87	0.10	0.00	0.00		
	3			0.00	0.00	0.08	0.78	0.14	0.00	0.00		
	4			0.00	0.00	0.07	0.80	0.13	0.00	0.00		
	5			0.00	0.00	0.07	0.81	0.12	0.00	0.00		
26	Overall			0.00	0.00	0.01	0.99	0.00	0.00	0.00		
	1			0.00	0.00	0.01	0.99	0.00	0.00	0.00		
	2			0.00	0.00	0.01	0.99	0.00	0.00	0.00		
	3			0.00	0.00	0.01	0.99	0.00	0.00	0.00		
	4			0.00	0.00	0.00	0.99	0.01	0.00	0.00		
	5			0.00	0.00	0.01	0.99	0.00	0.00	0.00		
27	Overall			0.00	0.00	0.08	0.82	0.10	0.00	0.00		
	1			0.00	0.00	0.06	0.85	0.09	0.00	0.00		
	2			0.00	0.00	0.13	0.81	0.06	0.00	0.00		
	3			0.00	0.00	0.07	0.81	0.11	0.00	0.00		
	4			0.00	0.00	0.07	0.83	0.10	0.00	0.00		
	5			0.00	0.00	0.09	0.81	0.09	0.00	0.00		
28	Overall			0.00	0.00	0.03	0.87	0.09	0.00	0.00		
	1			0.00	0.01	0.03	0.91	0.06	0.00	0.00		
	2			0.00	0.00	0.02	0.93	0.04	0.00	0.00		
	3			0.00	0.00	0.03	0.87	0.09	0.01	0.00		
	4			0.00	0.00	0.04	0.88	0.08	0.00	0.00		
	5			0.00	0.00	0.04	0.83	0.12	0.01	0.00		

Table G–2. NYS Public Schools (Without NYC) Grade 4 ELA Operational Test 2009: Percentages 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	0.00	0.11	0.53	0.34	0.02	0.00	0.00	
	1		0.00	0.00	0.01	0.17	0.55	0.26	0.02	0.00	0.00	
	2		0.00	0.00	0.01	0.09	0.56	0.32	0.01	0.00	0.00	
	3		0.00	0.00	0.00	0.09	0.54	0.36	0.02	0.00	0.00	
	4		0.00	0.00	0.00	0.12	0.53	0.31	0.03	0.00	0.00	
	5		0.00	0.00	0.00	0.09	0.46	0.41	0.04	0.00	0.00	
32–35	Overall		0.00	0.00	0.00	0.10	0.50	0.36	0.04	0.00	0.00	
	1		0.00	0.00	0.01	0.12	0.53	0.32	0.03	0.00	0.00	
	2		0.00	0.00	0.00	0.10	0.49	0.36	0.05	0.00	0.00	
	3		0.00	0.00	0.00	0.10	0.49	0.36	0.04	0.00	0.00	
	4		0.00	0.00	0.00	0.08	0.50	0.37	0.05	0.00	0.00	
	5		0.00	0.00	0.00	0.09	0.47	0.39	0.05	0.00	0.00	
31&35	Overall		0.00	0.00	0.01	0.17	0.60	0.21	0.01	0.00	0.00	
	1		0.00	0.00	0.01	0.21	0.61	0.17	0.01	0.00	0.00	
	2		0.00	0.00	0.01	0.15	0.62	0.20	0.02	0.00	0.00	
	3		0.00	0.00	0.01	0.16	0.59	0.23	0.01	0.00	0.00	
	4		0.00	0.00	0.01	0.16	0.57	0.26	0.01	0.00	0.00	
	5		0.00	0.00	0.01	0.16	0.63	0.20	0.01	0.00	0.00	

Table G–3. NYS Public Schools (Without NYC) Grade 5 ELA Operational Test 2009: Percentages 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	0.05	0.89	0.05	0.00	0.00		
	1			0.00	0.00	0.05	0.91	0.04	0.00	0.00		
	2			0.00	0.00	0.04	0.91	0.06	0.00	0.00		
	3			0.00	0.00	0.05	0.89	0.06	0.00	0.00		
	4			0.00	0.00	0.03	0.90	0.07	0.00	0.00		
	5			0.00	0.00	0.05	0.90	0.05	0.00	0.00		
26	Overall			0.00	0.00	0.05	0.87	0.08	0.00	0.00		
	1			0.00	0.00	0.05	0.86	0.09	0.00	0.00		
	2			0.00	0.00	0.04	0.85	0.10	0.01	0.00		
	3			0.00	0.00	0.04	0.87	0.09	0.00	0.00		
	4			0.00	0.00	0.03	0.88	0.08	0.00	0.00		
	5			0.00	0.00	0.07	0.86	0.07	0.00	0.00		
27	Overall			0.00	0.01	0.12	0.74	0.13	0.01	0.00		
	1			0.00	0.00	0.10	0.75	0.13	0.01	0.00		
	2			0.00	0.00	0.09	0.75	0.15	0.00	0.00		
	3			0.00	0.01	0.11	0.74	0.14	0.01	0.00		
	4			0.00	0.00	0.11	0.76	0.12	0.00	0.00		
	5			0.00	0.01	0.15	0.73	0.11	0.00	0.00		

Table G–4. NYS Public Schools (Without NYC) Grade 6 ELA Operational Test 2009: Percentages 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	0.09	0.43	0.40	0.07	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.00	0.09	0.43	0.41	0.07	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.10	0.24	0.56	0.10	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.01	0.12	0.43	0.37	0.06	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.01	0.08	0.39	0.44	0.08	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.01	0.08	0.44	0.40	0.07	0.00	0.00	0.00
31–34	Overall	0.00	0.00	0.00	0.02	0.16	0.43	0.31	0.07	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.02	0.15	0.44	0.32	0.06	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.05	0.29	0.56	0.10	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.04	0.21	0.45	0.25	0.04	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.02	0.15	0.41	0.31	0.10	0.01	0.00	0.00
	5	0.00	0.00	0.00	0.02	0.14	0.43	0.33	0.08	0.00	0.00	0.00
30&34	Overall	0.00	0.00	0.00	0.01	0.22	0.62	0.15	0.00	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.01	0.21	0.62	0.16	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.02	0.10	0.56	0.32	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.01	0.22	0.60	0.16	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.01	0.17	0.63	0.19	0.01	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.01	0.24	0.63	0.12	0.00	0.00	0.00	0.00

Table G–5. NYS Public Schools (Without NYC) Grade 7 ELA Operational Test 2009: Percentages 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	0.11	0.61	0.27	0.01	0.00		
	1			0.00	0.01	0.12	0.62	0.24	0.01	0.00		
	2			0.00	0.01	0.21	0.64	0.14	0.00	0.00		
	3			0.00	0.00	0.14	0.64	0.22	0.00	0.00		
	4			0.00	0.00	0.08	0.61	0.29	0.01	0.00		
	5			0.00	0.00	0.12	0.58	0.28	0.01	0.00		
28	Overall			0.00	0.00	0.11	0.70	0.18	0.01	0.00		
	1			0.00	0.01	0.14	0.69	0.16	0.00	0.00		
	2			0.00	0.00	0.10	0.76	0.14	0.00	0.00		
	3			0.00	0.01	0.12	0.70	0.17	0.00	0.00		
	4			0.00	0.00	0.06	0.70	0.23	0.01	0.00		
	5			0.00	0.00	0.12	0.70	0.18	0.01	0.00		
33	Overall			0.00	0.00	0.04	0.83	0.13	0.00	0.00		
	1			0.00	0.00	0.05	0.84	0.11	0.00	0.00		
	2			0.00	0.00	0.05	0.83	0.13	0.00	0.00		
	3			0.00	0.00	0.07	0.80	0.13	0.00	0.00		
	4			0.00	0.00	0.03	0.84	0.13	0.00	0.00		
	5			0.00	0.00	0.03	0.82	0.14	0.00	0.00		
34	Overall			0.00	0.00	0.10	0.77	0.13	0.01	0.00		
	1			0.00	0.00	0.12	0.75	0.12	0.01	0.00		
	2			0.00	0.00	0.14	0.71	0.13	0.02	0.00		
	3			0.00	0.00	0.10	0.69	0.20	0.01	0.00		
	4			0.00	0.00	0.07	0.82	0.11	0.00	0.00		
	5			0.00	0.00	0.09	0.78	0.12	0.00	0.00		
35	Overall			0.00	0.00	0.11	0.79	0.09	0.00	0.00		
	1			0.00	0.00	0.11	0.79	0.09	0.00	0.00		
	2			0.00	0.00	0.06	0.88	0.06	0.00	0.00		
	3			0.00	0.01	0.12	0.79	0.07	0.00	0.00		
	4			0.00	0.00	0.10	0.78	0.11	0.01	0.00		
	5			0.00	0.00	0.12	0.79	0.08	0.00	0.00		

Table G–6. NYS Public Schools (Without NYC) Grade 8 ELA Operational Test 2009: Percentages 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	0.10	0.41	0.39	0.09	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.02	0.15	0.46	0.32	0.05	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.04	0.48	0.40	0.07	0.01	0.00	0.00
	3	0.00	0.00	0.00	0.01	0.10	0.43	0.38	0.08	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00	0.09	0.42	0.41	0.08	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00	0.07	0.38	0.41	0.12	0.01	0.00	0.00
31–34	Overall	0.00	0.00	0.00	0.02	0.16	0.46	0.30	0.05	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.02	0.17	0.48	0.28	0.04	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.19	0.41	0.34	0.06	0.01	0.00	0.00
	3	0.00	0.00	0.00	0.03	0.17	0.44	0.28	0.07	0.01	0.00	0.00
	4	0.00	0.00	0.00	0.01	0.15	0.49	0.30	0.05	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.02	0.15	0.46	0.31	0.05	0.00	0.00	0.00
30&34	Overall	0.00	0.00	0.00	0.00	0.09	0.57	0.32	0.02	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.00	0.09	0.61	0.28	0.01	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.17	0.65	0.17	0.01	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.00	0.13	0.58	0.27	0.02	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00	0.05	0.49	0.44	0.02	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00	0.08	0.58	0.31	0.03	0.00	0.00	0.00

Appendix H

Item Level Differences for ELA Including New York City Schools Only

Table H–1. NYC Public Schools Grades 3–8 ELA Operational Test 2009: Percentages 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 Overall			0.00	0.00	0.10	0.77	0.12	0.00	0.00		
26	NYC Overall			0.00	0.00	0.01	0.99	0.00	0.00	0.00		
27	NYC Overall			0.00	0.00	0.11	0.78	0.11	0.00	0.00		
28	NYC Overall			0.00	0.00	0.03	0.86	0.10	0.01	0.00		
Grade 4												
Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
29–31	NYC Overall		0.00	0.00	0.00	0.16	0.54	0.28	0.02	0.00	0.00	
32–35	NYC Overall		0.00	0.00	0.01	0.19	0.53	0.25	0.03	0.00	0.00	
31&35	NYC Overall		0.00	0.00	0.01	0.16	0.59	0.23	0.01	0.00	0.00	
Grade 5												
Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
21	NYC Overall			0.00	0.00	0.06	0.86	0.08	0.00	0.00		
26	NYC Overall			0.00	0.00	0.05	0.84	0.11	0.00	0.00		
27	NYC Overall			0.00	0.01	0.13	0.71	0.13	0.01	0.00		
Grade 6												
Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
27–30	NYC Overall	0.00	0.00	0.00	0.01	0.10	0.42	0.40	0.08	0.00	0.00	0.00
31–34	NYC Overall	0.00	0.00	0.00	0.03	0.19	0.42	0.29	0.06	0.00	0.00	0.00
30&34	NYC Overall	0.00	0.00	0.00	0.02	0.27	0.59	0.13	0.00	0.00	0.00	0.00
Grade 7												
Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
27	NYC Overall			0.00	0.00	0.11	0.63	0.24	0.01	0.00		
28	NYC Overall			0.00	0.00	0.10	0.67	0.22	0.01	0.00		
33	NYC Overall			0.00	0.00	0.06	0.85	0.10	0.00	0.00		
34	NYC Overall			0.00	0.00	0.12	0.77	0.10	0.00	0.00		
35	NYC Overall			0.00	0.01	0.09	0.78	0.11	0.00	0.00		
Grade 8												
Item #	Scoring Model	-5	-4	-3	-2	-1	0	1	2	3	4	5
27–30	NYC Overall	0.00	0.00	0.00	0.02	0.16	0.46	0.31	0.05	0.00	0.00	0.00
31–34	NYC Overall	0.00	0.00	0.01	0.06	0.25	0.43	0.21	0.04	0.00	0.00	0.00
30&34	NYC Overall	0.00	0.00	0.00	0.01	0.13	0.59	0.26	0.01	0.00	0.00	0.00