

# **New York State Testing Program**

## **English Language Arts and Reading**

### **Grade 4**

## **Technical Report 2005**



Developed and published under contract with New York State Education Department by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc., 20 Ryan Ranch Road, Monterey, California 93940-5703. Copyright © 2005 by New York State Education Department. Only State of New York educators and citizens may copy, download, and/or print the document located online at <http://www.emsc.nysed.gov/ciai/testing/pubs.html>. Any other use or reproduction of this document, in whole or in part, requires written permission of the New York State Education Department.

## Foreword

---

The technical information herein is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as described in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

# Table of Contents

---

<b>FOREWORD</b> .....	<b>1</b>
<b>TABLE OF CONTENTS</b> .....	<b>2</b>
<b>LIST OF TABLES</b> .....	<b>3</b>
<b>PART 1: TEST DESIGN</b> .....	<b>4</b>
<i>The New York State Learning Standards for English Language Arts</i> .....	4
<i>Test Configuration</i> .....	4
<i>Writing Mechanics Score</i> .....	5
<b>STUDENT PARTICIPATION AND TESTING ACCOMMODATIONS</b> .....	<b>6</b>
<i>Students to be Tested</i> .....	6
<i>Testing Accommodations</i> .....	6
<i>Students with Disabilities</i> .....	6
<i>Limited English Proficient (LEP) Students</i> .....	7
<i>Other Considerations</i> .....	7
<b>ITEM DEVELOPMENT</b> .....	<b>7</b>
<b>ITEM REVIEW PROCESS</b> .....	<b>7</b>
<i>Documenting Content</i> .....	7
<i>Minimizing Bias</i> .....	8
<i>Minimizing Speededness</i> .....	8
<b>TEST CONSTRUCTION AND PRE-EQUATING</b> .....	<b>8</b>
<i>Calibration Samples</i> .....	8
<i>Answer Choice Information</i> .....	8
<i>Item Response Theory Models</i> .....	9
<i>Equating Method</i> .....	10
<i>Item Selection Criteria and Process</i> .....	10
<i>Procedures for Eliminating Bias and Minimizing Differential Item Functioning</i> .....	11
<b>PART 2: ITEM STATISTICS FOR THE OPERATIONAL DATA</b> .....	<b>13</b>
DATA CLEANING .....	13
ITEM ANALYSIS .....	14
DIFFERENTIAL ITEM FUNCTIONING ANALYSIS OF OPERATIONAL DATA .....	16
<b>PART 3: SCORING AND RELIABILITY</b> .....	<b>18</b>
RAW SCORE TO SCALE SCORE CONVERSION .....	18
RELIABILITY .....	18
ESTIMATED CONDITIONAL STANDARD ERRORS OF SCALE SCORES .....	20
LOWEST AND HIGHEST OBTAINABLE SCALE SCORES .....	21
INTER-RATER AGREEMENT .....	21
EXPECTED SPI SCORES ON THE STANDARDS AT THE DECISION POINTS .....	26
<b>PART 4: DESCRIPTIVE STATISTICS</b> .....	<b>27</b>
SCALE SCORE FREQUENCY DISTRIBUTIONS FOR THE STATE AND SUBGROUPS .....	27
G4 ELA SCALE SCORE MEANS AND STANDARD DEVIATIONS .....	28
G4 ELA PERFORMANCE LEVEL DISTRIBUTION .....	28
<b>REFERENCES</b> .....	<b>29</b>

# List of Tables

---

Table 1. New York State Learning Standards for English Language Arts.....	4
Table 2. Points per item type for G4 ELA scores .....	5
Table 3. Condition Codes for the ELA CR items .....	5
Table 4. Steps Involved in Data Clean-up for Analysis Preparation.....	13
Table 5. G4 ELA Item Level Statistics .....	14
Table 6. G4 RD Item Level Statistics .....	15
Table 7. Number of Students in each Gender or Ethnic Group .....	16
Table 8. Numbers of Items Flagged for DIF in G4 ELA and RD.....	17
Table 9. Raw Score to Scale Score with SE for G4 ELA 2005 .....	19
Table 10. Raw Score to Scale Score with SE for G4 RD 2005.....	20
Table 11. G4 ELA & RD 2005 Inter-Rater Agreement: Public, Non-NYC, N=10,432 .....	22
Table 12. G4 ELA & RD 2005 Inter-Rater Agreement: Non-Public, Non-NYC, N=953 .....	22
Table 13. G4 ELA & RD 2005 Inter-Rater Agreement: Public, NYC, N=6,379.....	23
Table 14. Percentages of Inter-Rater Score Differences: Public, Non-NYC .....	23
Table 15. Percentages of Inter-Rater Score Differences: Non-Public, Non-NYC .....	24
Table 16. Percentages of Inter-Rater Score Differences: Public, NYC .....	24
Table 17. Reliability Indices of Hand Scoring: Public, Non-NYC .....	25
Table 18. Reliability Indices of Hand Scoring: Non-Public, Non-NYC.....	25
Table 19. Reliability Indices of Hand Scoring: Public, NYC .....	26
Table 20. G4 ELA 2005 Standard Performance Index Information .....	26
Table 21. G4 ELA 2005 Summary of Scale Score Information.....	27
Table 22. G4 ELA Statewide Scale Score Information .....	28
Table 23. G4 ELA Statewide Performance Level Information.....	28

## Part 1: Test Design

---

### The New York State Learning Standards for English Language Arts

The New York State *Learning Standards for English Language Arts* document is available from the New York State Education Department web site, at <http://www.emsc.nysed.gov/ciai/ela/pub/elalearn.pdf>. The four learning standards are listed in Table 1 below. The Grade 4 English Language Arts (G4 ELA) assessment is written to test students in Standards 1, 2, and 3.

**Table 1. New York State Learning Standards for English Language Arts**

<b>Standard 1</b>	Students will read, write, listen, and speak for information and understanding.
<b>Standard 2</b>	Students will read, write, listen, and speak for literary response and expression.
<b>Standard 3</b>	Students will read, write, listen, and speak for critical analysis and evaluation.
Standard 4	Students will read, write, listen, and speak for social interaction.

### Test Configuration

Similar to the 1999 through 2003 forms, the 2004 G4 ELA and Reading (RD) test has the following configuration. The test is divided into two sessions. There are 28 multiple choice (MC) items worth a total of 28 points; there are 8 constructed response (CR) items, with a total of 14 points that apply to the ELA score and 6 points that apply to the RD score. The CR items may be short response (SR) or extended response (ER) items. The total number of items on the test is 36, and the maximum raw score total is 42 points for ELA and 34 points for RD.

#### Session 1

Session 1 is comprised of 28 MC items, these items are scored and the right-wrong data from them is used towards the ELA raw score and towards the RD raw score. Each MC item addresses one of the three tested New York State Learning Standards for English Language Arts.

#### Session 2

Session 2 is comprised of two parts: Listening (Part 1), and Writing (Part 2). Part 1 contains linked information stimuli, accompanied by 2 SR items and 1 ER item which are scored together to derive an ELA listening cluster score (zero to four points) which addresses Standard 2. Part 2 consists of a stimulus and 1 ER item which contributes to the ELA independent writing score (zero to three points) that addresses Standard 2.

#### Session 3

Session 3 contains linked information stimuli, accompanied by 3 SR items and 1 ER item. All four items are scored together to obtain the ELA reading cluster score (zero to four points). Each SR item is scored individually (zero to two points) to provide the three analytic (RD) item scores.

## Writing Mechanics Score

As part of the ELA test, the three ER responses across sessions 1 and 2 are scored together to derive a writing mechanics cluster score (zero to three points). Although writing mechanics is not linked to any of the New York State Learning Standards for English Language Arts, it contributes to the overall ELA score.

Table 2 shows the numbers of score points by the item type or cluster, and the total numbers of items and clusters, for the Grade 4 ELA test.

**Table 2. Points per item type for G4 ELA scores**

<b>Item Type or Cluster</b>	<b>Grade 4 ELA</b>
Multiple choice (MC)	28 pts
Listening cluster	4 pts
Reading cluster	4 pts
Independent writing item	3 pts
Writing mechanics cluster	3 pts
Total points	42 pts
Total MC items and ELA clusters	32 items

In scaling and scoring, each of the clusters is treated as a constructed response (CR) item. The following condition codes were used in scoring the responses to the CR items:

**Table 3. Condition Codes for the ELA CR items**

<b>Condition Code</b>	<b>Meaning</b>
A	Blank
F	Absent

## **Student Participation and Testing Accommodations**

### **Students to be Tested**

The New York State Testing Program (NYSTP) Grade 4 English Language Arts test must be administered to all public school students in Grade 4 and all ungraded students who are age-equivalent to students in Grade 4. This includes students who have been retained in Grade 4. Nonpublic schools are strongly encouraged to administer the tests. The exceptions noted below apply to students in public and nonpublic schools participating in the NYSTP.

### **Testing Accommodations**

Accommodations were used in the NYSTP operational tests to provide equal access to assessments for students with disabilities. These accommodations are used to increase the validity of test scores by offsetting behavioral constraints due to the disability and retaining the essential features of the assessment. The following represents the policy of the New York State Education Department (NYSED) for the use of testing accommodations.

### **Students with Disabilities**

The Committee on Special Education (CSE) must decide for each student on a case-by-case basis, and document on the student's Individualized Education Program, whether the student will participate in the general State assessment, in a locally selected assessment, or in the New York State Alternate Assessment for Students with Severe Disabilities (NYSAA). The criteria that the CSE must use to determine eligibility for a locally selected assessment is available at <http://www.emsc.nysed.gov/deputy/Documents/disabilities-assess.htm>. The criteria to determine eligibility for the NYSAA is available on <http://www.vesid.nysed.gov/specialed/alterassessment/alterassess.htm>.

It is the responsibility of the principal to ensure that testing accommodations specified in the IEP or Section 504 Accommodation Plan (504 Plan) are provided to students with disabilities as long as they do not alter a construct being measured by the test. Students who have been declassified may continue to be provided testing accommodations if recommended by the local CSE at the time of declassification and in the student's declassification IEP. Testing accommodations that alter the construct being measured are not permitted on elementary- and intermediate-level State assessments. For more information, see <http://www.vesid.nysed.gov/specialed/publications/policy/testaccess/guide.htm>.

Principals may modify testing procedures for General Education students who incur an injury (for example, a broken arm) or experience the onset of a short- or long-term disability (for example, epilepsy) sustained or diagnosed within 30 days prior to the administration of State tests. In such cases, when sufficient time is not available for the development of an Individualized Education Program (IEP) or a 504 Plan, principals may authorize certain accommodations that will not significantly change the skills being tested.

Eligibility for such accommodations is based on the principal's professional discretion, but the principal may confer with members of the Committee for Special Education (CSE) or with other school personnel in making such a determination. Pursuant to Section 100.3 of the Regulations of the Commissioner of Education, building principals are responsible for administering State assessments and for maintaining the integrity of test content and programs in accordance with directions and procedures established by the Commissioner of Education.

## **Limited English Proficient (LEP) Students**

The No Child Left Behind (NCLB) Act requires that the English proficiency of all Limited English Proficient (LEP) students (as defined in Part 154 of the Regulations of the Commissioner of Education) be tested annually. New York State has introduced a new assessment of the English language proficiency of students for whom English is a second language. Effective Spring 2003, all LEP students, regardless of grade, must take the New York State English as a Second Language Achievement Test (NYSESLAT). LEP students must take this assessment even if they take the Grade 4 English Language Arts test.

Additional information concerning the inclusion of LEP students in State examinations in English Language Arts and Mathematics is provided on the Department's website <http://www.emsc.nysed.gov/osa>.

## **Other Considerations**

When determining who will participate in the NYSTP and who will participate in the Alternate Assessment, school administrators must consider those students who attend programs operated by the Board of Cooperative Educational Services (BOCES), or who are in approved private school placements, as well as in any other programs located outside the school district. Students who are absent during the testing administrations should be tested during the designated makeup period.

## **Item Development**

A staff of professional item writers researched, collected, and wrote the test material. All assessment materials were carefully reviewed for content and editorial accuracy. Artists and designers worked with the writers during development for graphic and textual consistency. With assistance from the New York State Education Department, all test items were developed to align with the content and measure the Learning Standards for English Language Arts. Standards Performance Index (SPI) scores are assigned to students for each of these reporting categories.

## **Item Review Process**

### **Documenting Content**

An integral part of the development process was documentation of content using New York State's Learning Standards. All items used on the New York State tests are reviewed for content by CTB Development staff, New York State Education Department staff, and New York State teachers. This procedure checks that items are sound in content and format, and targeted appropriately to the courses in which the associated concepts are typically taught.

## **Minimizing Bias**

The developers of the NYSTP tests gave careful attention to questions of possible ethnic, racial, gender, regional, and age bias. All materials were written and reviewed to conform to the company's editorial policies and guidelines for equitable assessment, as well as NYSED's guidelines for item development.

In addition, educators and other stakeholders from different parts of the state reviewed the items from their perspective as members of various ethnic groups. They identified assessment materials that might reflect possible bias in language, subject matter, or representation of people. Their comments and suggestions were considered carefully during the revision and selection of items for the operational tests. All materials were written to SED specifications and carefully checked by groups of trained New York community participants.

## **Minimizing Speededness**

Test developers also considered speededness in the development of the NYSTP tests. CTB believes that achievement tests should not be speeded; little or no useful instructional information can be obtained from the fact that a student did not finish a test, while a great deal can be learned from student responses to questions. For that reason, sufficient administration time limits were set for the NYSTP tests.

The Research Department at CTB routinely conducts additional speededness analyses based on actual test data. Tables 5 and 6 show the omit rates for items on the G4 ELA/RD test. All omit rates are sufficiently low enough ( $< 5\%$ ) to provide little evidence of speededness on these tests.

## **Test Construction and Pre-equating**

### **Calibration Samples**

Field test forms for the NYSTP tests were administered to students in public and private schools across the State in 2001, 2002, and 2003. Effort was made to select a sample of students representative of the State tested population. The field test items were calibrated and equated to the existing New York State Grade 4 ELA scale.

Since these items are calibrated and on a common scale, the pool of available Grade 4 English Language Arts items can be used to construct a test form and to produce a raw-score-to-scale-score table for that form. The 2005 operational NYSTP tests were constructed using items from that pool. What follows is an overview of the analysis of field test data that resulted in the calibration of items.

### **Answer Choice Information**

Statistical information about student performance is produced for each multiple choice item. Specifically, three statistics are examined for each item: (1) the proportion of students choosing each answer, (2) the point-biserial correlation between the answer choice and the number-correct score on the rest of the test, and (3) omit rates. For each constructed response item, the proportion of students at each score level, omit rates, and p-values (mean item score divided by the total number of points possible) are examined.

## Item Response Theory Models

Although useful, the differences in proportion of points received (p-values) limit the degree to which one can compare important characteristics of the test items. Item response theory (IRT) allows one to make better comparisons among items, even those from different test forms, by using a common scale for all items (i.e., as if there were a hypothetical test that contained items from all forms). The three-parameter logistic (3PL) model (Lord & Novick, 1968; Lord, 1980) was used to analyze item responses on the multiple choice items. For analysis of the constructed response items, the two-parameter partial credit model (2PPC) (Muraki, 1992; Yen, 1993) was used.

Item response theory is a statistical procedure that takes into account the fact that not all test items are alike and that all items do not provide the same amount of information in determining how much a student knows or can do. Computer programs that implement IRT models use actual students' data to estimate the characteristics of the items on a test, called "parameters." The parameter estimation process is called "item calibration."

IRT models typically vary according to the number of parameters estimated. For the New York State tests, three parameters are estimated: the discrimination parameter, the difficulty parameter(s), and, for multiple choice items, the guessing parameter. The discrimination parameter is an index of how well an item differentiates between high-performing and low-performing students. An item that cannot be answered correctly by low-performing students, but can be answered correctly by high-performing students, will have a high discrimination value. The difficulty parameter is an index of how easy or difficult an item is. The higher the difficulty parameter, the harder the item. The guessing parameter is the probability that a student with very low ability will answer the item correctly.

The scale score (SS) is the basic score for the New York State tests. It is used to derive other scores that describe test performance, such as the four performance levels and the standard-based performance index scores (SPIs). Scale scores can be obtained by one of two scoring methods: IRT item-pattern scoring, or number-correct scoring. Since 2002, scores on the New York State tests are determined using number-correct scoring.

Because the characteristics of MC and CR items are different, two IRT models were used in item calibration. The three-parameter logistic (3PL) model (Lord & Novick, 1968; Lord, 1980) was used in the analysis of MC items. In this model, the probability that a student with ability  $\theta$  responds correctly to item  $i$  is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]},$$

where  $a_i$  is the item discrimination,  $b_i$  is the item difficulty, and  $c_i$  is the probability of a correct response by a very low-scoring student.

For analysis of the constructed response items, the two-parameter partial credit (2PPC) model (Muraki, 1992; Yen, 1993) was used. The 2PPC model is a special case of Bock's (1972) nominal model. Bock's model states that the probability of an examinee with ability  $\theta$  having a score  $(k - 1)$  at the  $k$ -th level of the  $j$ -th item is

$$P_{jk}(\theta) = P(x_j = k - 1 | \theta) = \frac{\exp Z_{jk}}{\sum_{i=1}^{m_j} \exp Z_{ji}}, \quad k = 1 \dots m_j,$$

where

$$Z_{jk} = A_{jk}\theta + C_{jk}.$$

The  $m_j$  denotes the number of score levels for the  $j$ -th item, and typically the highest score level is assigned  $(m_j - 1)$  score points. For the special case of the 2PPC model used here, the following constraints were used:

$$A_{jk} = \alpha_j(k - 1),$$

and

$$C_{jk} = -\sum_{i=0}^{k-1} \gamma_{ji}, \text{ where } \gamma_{j0} = 0,$$

where  $\alpha_j$  and  $\gamma_{ji}$  are the free parameters to be estimated from the data. Each item has  $(m_j - 1)$  independent  $\gamma_{ji}$  parameters and one  $\alpha_j$  parameter; a total of  $m_j$  parameters are estimated for each item.

The IRT model parameters were estimated using CTB's PARDUX software (Burket, 1991). PARDUX estimates parameters simultaneously for MC and CR items using marginal maximum likelihood procedures implemented via the EM (expectation-maximization) algorithm (Bock & Aitkin, 1981; Thissen, 1982).

Simulation studies have compared PARDUX with MULTILOG (Thissen, 1991), PARSCALE (Muraki & Bock, 1991), and BIGSTEPS (Wright & Linacre, 1992). PARSCALE, MULTILOG, and BIGSTEPS are among the most widely known and used IRT programs. PARDUX was found to perform at least as well as these other programs (Fitzpatrick, 1990; Fitzpatrick, 1994; Fitzpatrick and Julian, 1996).

## Equating Method

After the item calibration, all of the Grade 4 English Language Arts field test items were placed on the NYS G4 ELA scale using the operational MC items as anchors. The equating was performed using the test characteristic curve method (Stocking & Lord, 1983) implemented by PARDUX. In previous years, operational data were used to re-calibrate items and re-equate them. NYSED, however, made a decision in 2002 to use the pre-equating model, which is similar to what is done for the New York State Regents program. This allows the production of scoring tables (see Part 3) ahead of the operational administration, once the operational form is selected.

## Item Selection Criteria and Process

Item selection for the NYSTP tests was based on the classical and IRT statistics of the test items. Selection was conducted by content experts from CTB and NYSED and reviewed by psychometricians at CTB. Final approval of the selected items was given by NYSED. Two criteria governed the item selection process. The first of these was to meet the content specifications provided by the New York State Education Department. Within the limits set by these requirements, developers selected items with the best psychometric characteristics from the field test item pool. Developers chose items that minimized measurement error throughout the range of expected achievement as indicated by the

reciprocal of the square root of the IRT information function (Lord, 1980, p. 71). Developers aimed to create forms with the content and psychometric properties of previous operational forms.

Item selection for the operational tests was facilitated using the Windows version of the program ITEMSYS (Burket, 1988). ITEMSYS creates an interactive connection between the developer selecting the test items and the item database. This program monitors the impact of each decision made during the item selection process and offers a variety of options for grouping, classifying, sorting, and ranking items to highlight key information as it is needed (see Green, Yen, & Burket, 1989).

ITEMSYS has three parts. The first part selects a working item pool of manageable size from the larger tryout pool. The second part of the program uses this selected item pool to perform the final test selection. In the third part of the program, a table shows both expected number correct and the standard error of ability estimate (a function of scale score), as well as statistical and graphic summaries on bias, fit, and the standard error of the final test. Any fault in the final selection becomes apparent as the final statistics are generated. Examples of possible faults that may occur are cases when the test is too easy or too difficult, contains items demonstrating differential item functioning (see below), does not meet the requirements to match a parallel form, or does not adequately measure part of the range of performance. A developer detecting any such problems can then return to the second stage of the program and revise the selection. The flexibility and utility of the program encourages multiple attempts at fine-tuning the item selection.

### **Procedures for Eliminating Bias and Minimizing Differential Item Functioning**

As part of the testing, the students reported their gender and ethnic background information. Using this self-reported information, statistical differential item functioning (DIF) analyses were conducted for male and female gender groups, and for the following ethnic groups: African-American, Hispanic-American, and Asian-American.

Three procedures were used to eliminate bias and minimize differential item functioning (DIF) in the New York State tests.

The first was based on the premise that careful editorial attention to validity is an essential step in keeping bias to a minimum. Bias can occur only if the test is differentially valid for a given group of test takers. If the test entails irrelevant skills or knowledge (however common), the possibility of DIF is increased. Thus, preserving content validity is essential.

The second step was to follow the item writing guidelines established by NYSED. Developers reviewed NYSTP materials with these guidelines in mind. These internal editorial reviews were done by at least four separate people: the content editor, who directly supervises the item writers; the project director; a style editor; and a proofreader. The final test built from the tryout materials was reviewed by at least these same people.

In the third procedure, New York State educational community professionals who represent various ethnic groups reviewed all tryout materials. These professionals were asked to consider and comment on the appropriateness of language, subject matter, and representation of people.

It is believed that these three procedures improved the quality of the New York State tests and reduced bias. However, current evidence suggests that expertise in this area is no substitute for data; reviewers are often wrong about which items work to the disadvantage of a group, apparently because some of their

ideas about how students will react to items may be faulty (Sandoval & Mille, 1979; Jensen, 1980). Thus, an empirical approach is desirable.

A fourth procedure provides the empirical approach recommended to supplement expert, yet subjective, judgment methods. Statistical methods were used to identify items exhibiting possible DIF. Items flagged for DIF in the field test stage are closely examined for content bias.

## Part 2: Item Statistics for the Operational Data

---

### Data Cleaning

Typically, item analyses are conducted once CTB receives data that meets the following requirements established by NYSED:

- Comprises at least 85% of the estimated number of students in the State
- Includes New York City and Buffalo
- Includes at least one of the cities of Rochester, Syracuse, or Yonkers, and
- Includes at least two of the cities of Mount Vernon, Albany, Binghamton, Schenectady, or New Rochelle.

The data received by CTB in 2005 contained 100% of the cases. A number of cases were excluded from the data analysis. Initially, the state data set contained 231,707 cases. Table 4, below, shows the data cleaning steps and the resulting size (93%) of the cases used for conducting item analysis.

**Table 4. Steps Involved in Data Clean-up for Analysis Preparation**

Steps Taken	# Cases Deleted	Ending N
Original Data		231,707
Duplicate Records	32	231,675
Grade Not Equal to 4	223	231,452
LEP5 Data	12,212	219,240
Invalid Data	3,106	<b>216,134</b>

Students whose LEP status = 5 are not required to take the test.

As Table 4 shows, the following records were eliminated:

- Duplicated records
- Out-of-grade students, who were administered a 4<sup>th</sup> grade test despite not being 4<sup>th</sup> grade students
- Students whose limited English proficient (LEP) status was "5," indicating that they scored below the threshold percentile on a norm-referenced English reading test or the publisher's recommended score on an approved measure of English as a second language in Reading
- Students who did not have a valid attempt in each of three sections as determined by the application of CTB's Invalidation / Omission / Suppression rules (approved by NYSED).

## Item Analysis

Table 5 presents the results of item analyses conducted using the total population operational data (except cases excluded during the data cleaning) for the G4 ELA test. The labels for the variables denote the following:

- ITEM** Item number.
- OMIT** Proportion of students who had a blank response or double marks on MC items, or condition codes on the CR items.
- PCTSEL\*** For MC items, this is the percentage of students who chose the first through the fourth answer option (or double-marked, for Pctsel0). For CR items, it is the percentage of students who received a score of 0 through the maximum number of points possible.
- P\_BIS** Point-biserial correlations for each response option.
- KEY** The correct response option, for MC items.
- P\_VAL** Item difficulty after omitted responses are converted to 0s (wrong). For MC items, p-value is the proportion of students responding correctly. For a CR item, p-value is the mean raw score divided by the maximum number of score points for an item.

**Table 5. G4 ELA Item Level Statistics**

Raw Score Data			Test Administration Data				Reliability Feldt-Raju				P-Value Mean	
Mean		SD	Number of Items		Number of Students							
30.52		7.28	32		216,134		0.90				0.73	
ITEM	OMIT	PCTSEL0	PCTSEL1	PCTSEL2	PCTSEL3	PCTSEL4	P_BIS1	P_BIS2	P_BIS3	P_BIS4	KEY	P_VAL
1	0.0001	0.01%	92.42%	2.40%	0.77%	4.39%	0.35	-0.16	-0.14	-0.27	1	0.924
2	0.0002	0.02%	4.71%	83.11%	5.00%	7.14%	-0.31	0.48	-0.31	-0.18	2	0.831
3	0.0001	0.02%	2.87%	0.67%	0.59%	95.83%	-0.27	-0.14	-0.12	0.33	4	0.958
4	0.0002	0.01%	1.28%	93.41%	3.67%	1.61%	-0.16	0.34	-0.23	-0.18	2	0.934
5	0.0005	0.04%	5.05%	2.92%	86.06%	5.89%	-0.22	-0.28	0.43	-0.23	3	0.861
6	0.0004	0.02%	3.18%	1.04%	2.43%	93.29%	-0.21	-0.20	-0.28	0.41	4	0.933
7	0.0004	0.01%	88.92%	1.68%	8.08%	1.26%	0.36	-0.16	-0.24	-0.24	1	0.889
8	0.0007	0.03%	6.52%	3.99%	8.21%	81.17%	-0.31	-0.19	-0.17	0.41	4	0.812
9	0.0007	0.03%	4.13%	89.56%	1.81%	4.40%	-0.25	0.41	-0.18	-0.23	2	0.896
10	0.0011	0.03%	4.86%	81.87%	6.13%	7.00%	-0.27	0.48	-0.27	-0.23	2	0.819
11	0.0008	0.02%	4.95%	3.23%	5.71%	86.00%	-0.31	-0.21	-0.27	0.49	4	0.860
12	0.0010	0.02%	8.44%	7.35%	83.21%	0.88%	-0.35	-0.29	0.51	-0.16	3	0.832
13	0.0012	0.02%	6.49%	3.68%	9.39%	80.29%	-0.24	-0.26	-0.25	0.46	4	0.803
14	0.0011	0.02%	62.92%	16.66%	2.23%	18.06%	0.45	-0.29	-0.16	-0.21	1	0.629
15	0.0011	0.01%	76.12%	4.85%	6.42%	12.49%	0.46	-0.28	-0.22	-0.24	1	0.761

Table 5 continues

**Table 5. G4 ELA Item Level Statistics (continued)**

ITEM	OMIT	PCTSEL0	PCTSEL1	PCTSEL2	PCTSEL3	PCTSEL4	P_BIS1	P_BIS2	P_BIS3	P_BIS4	KEY	P_VAL
16	0.0012	0.02%	6.06%	24.83%	59.90%	9.07%	-0.25	-0.17	0.41	-0.22	3	0.599
17	0.0016	0.04%	17.07%	19.29%	14.53%	48.91%	-0.19	-0.03	-0.18	0.31	4	0.489
18	0.0017	0.03%	8.80%	1.35%	1.14%	88.50%	-0.29	-0.19	-0.15	0.39	4	0.885
19	0.0019	0.03%	6.97%	77.05%	2.49%	13.28%	-0.27	0.40	-0.19	-0.19	2	0.770
20	0.0020	0.03%	12.21%	69.43%	11.42%	6.71%	-0.34	0.39	-0.16	-0.06	2	0.694
21	0.0034	0.03%	12.09%	4.10%	22.02%	61.43%	-0.27	-0.22	-0.08	0.36	4	0.614
22	0.0042	0.03%	7.26%	8.13%	75.71%	8.46%	-0.33	-0.17	0.46	-0.21	3	0.757
23	0.0046	0.03%	7.17%	4.10%	74.97%	13.27%	-0.27	-0.24	0.40	-0.14	3	0.750
24	0.0086	0.04%	14.60%	6.89%	71.33%	6.29%	-0.23	-0.19	0.43	-0.22	3	0.713
25	0.0089	0.03%	74.89%	9.86%	5.62%	8.72%	0.41	-0.15	-0.31	-0.17	1	0.749
26	0.0106	0.04%	6.75%	68.34%	10.65%	13.15%	-0.25	0.37	-0.17	-0.12	2	0.683
27	0.0123	0.03%	63.68%	10.96%	15.53%	8.57%	0.37	-0.16	-0.09	-0.29	1	0.637
28	0.0128	0.03%	6.47%	19.83%	63.59%	8.81%	-0.21	-0.15	0.42	-0.26	3	0.636
29	0.0012	0.66%	10.56%	38.24%	37.27%	13.15%					CR	0.629
30	0.0049	4.30%	23.39%	43.56%	28.26%						CR	0.651
31	0.0020	1.02%	19.67%	50.18%	28.93%						CR	0.689
32	0.0031	2.10%	17.85%	38.60%	32.62%	8.52%					CR	0.567

**Table 6. G4 RD Item Level Statistics**

Raw Score Data		Test Administration Data						Reliability				P-Value	
Mean	SD	Number of Items			Number of Students			Feldt-Raju				Mean	
26.54	6.24	31			216,134			0.89				0.78	
ITEM	OMIT	PCTSEL0	PCTSEL1	PCTSEL2	PCTSEL3	PCTSEL4	P_BIS1	P_BIS2	P_BIS3	P_BIS4	KEY	P_VAL	
1	0.0001	0.01%	92.42%	2.40%	0.77%	4.39%	0.36	-0.17	-0.14	-0.28	1	0.924	
2	0.0002	0.02%	4.71%	83.11%	5.00%	7.14%	-0.32	0.49	-0.31	-0.18	2	0.831	
3	0.0001	0.02%	2.87%	0.67%	0.59%	95.83%	-0.28	-0.15	-0.12	0.34	4	0.958	
4	0.0002	0.01%	1.28%	93.41%	3.67%	1.61%	-0.16	0.35	-0.24	-0.19	2	0.934	
5	0.0005	0.04%	5.05%	2.92%	86.06%	5.89%	-0.22	-0.29	0.44	-0.23	3	0.861	
6	0.0004	0.02%	3.18%	1.04%	2.43%	93.29%	-0.22	-0.21	-0.30	0.42	4	0.933	
7	0.0004	0.01%	88.92%	1.68%	8.08%	1.26%	0.37	-0.16	-0.25	-0.25	1	0.889	
8	0.0007	0.03%	6.52%	3.99%	8.21%	81.17%	-0.32	-0.19	-0.17	0.42	4	0.812	
9	0.0007	0.03%	4.13%	89.56%	1.81%	4.40%	-0.26	0.42	-0.19	-0.24	2	0.896	
10	0.0011	0.03%	4.86%	81.87%	6.13%	7.00%	-0.28	0.49	-0.27	-0.24	2	0.819	
11	0.0008	0.02%	4.95%	3.23%	5.71%	86.00%	-0.32	-0.21	-0.28	0.50	4	0.860	
12	0.0010	0.02%	8.44%	7.35%	83.21%	0.88%	-0.36	-0.30	0.53	-0.17	3	0.832	
13	0.0012	0.02%	6.49%	3.68%	9.39%	80.29%	-0.24	-0.26	-0.25	0.46	4	0.803	
14	0.0011	0.02%	62.92%	16.66%	2.23%	18.06%	0.45	-0.30	-0.17	-0.20	1	0.629	
15	0.0011	0.01%	76.12%	4.85%	6.42%	12.49%	0.47	-0.28	-0.22	-0.25	1	0.761	
16	0.0012	0.02%	6.06%	24.83%	59.90%	9.07%	-0.25	-0.17	0.40	-0.22	3	0.599	
17	0.0016	0.04%	17.07%	19.29%	14.53%	48.91%	-0.18	-0.03	-0.18	0.29	4	0.489	
18	0.0017	0.03%	8.80%	1.35%	1.14%	88.50%	-0.29	-0.20	-0.15	0.40	4	0.885	
19	0.0019	0.03%	6.97%	77.05%	2.49%	13.28%	-0.27	0.40	-0.20	-0.19	2	0.770	
20	0.0020	0.03%	12.21%	69.43%	11.42%	6.71%	-0.34	0.39	-0.16	-0.06	2	0.694	
21	0.0034	0.03%	12.09%	4.10%	22.02%	61.43%	-0.27	-0.22	-0.08	0.36	4	0.614	

Table 6 continues

**Table 6. G4 RD Item Level Statistics (continued)**

ITEM	OMIT	PCTSEL0	PCTSEL1	PCTSEL2	PCTSEL3	PCTSEL4	P_BIS1	P_BIS2	P_BIS3	P_BIS4	KEY	P_VAL
22	0.0042	0.03%	7.26%	8.13%	75.71%	8.46%	-0.33	-0.17	0.46	-0.21	3	0.757
23	0.0046	0.03%	7.17%	4.10%	74.97%	13.27%	-0.28	-0.24	0.40	-0.13	3	0.750
24	0.0086	0.04%	14.60%	6.89%	71.33%	6.29%	-0.23	-0.19	0.43	-0.22	3	0.713
25	0.0089	0.03%	74.89%	9.86%	5.62%	8.72%	0.41	-0.15	-0.32	-0.17	1	0.749
26	0.0106	0.04%	6.75%	68.34%	10.65%	13.15%	-0.26	0.37	-0.18	-0.12	2	0.683
27	0.0123	0.03%	63.68%	10.96%	15.53%	8.57%	0.35	-0.16	-0.07	-0.29	1	0.637
28	0.0128	0.03%	6.47%	19.83%	63.59%	8.81%	-0.21	-0.14	0.41	-0.26	3	0.636
29	0.0041	5.17%	19.10%	75.32%							CR	0.849
30	0.0059	7.94%	20.80%	70.68%							CR	0.811
31	0.0046	5.65%	37.71%	56.19%							CR	0.750

### Differential Item Functioning Analysis of Operational Data

To assess DIF for the New York State tests, students were identified as African-American, White, Hispanic, or Asian-American. These ethnic groups were chosen for DIF analyses because these populations are the largest in the State. Gender analyses were also conducted.

Developers strive to produce tests that minimize DIF. The DIF results reported here are those obtained when scoring students on the operational test using the pre-equated field test parameters. Thus, they may differ from DIF results obtained at the time of the field test administration.

Using demographic information, statistical DIF analyses were conducted for various ethnic groups and for males and females. A random sample was drawn from the final state GRT. Next, the sample was augmented by randomly selecting additional cases for any group of students whose count in the sample was less than 500 in an attempt to enhance the reliability of the DIF analyses. The numbers of cases for the groups are reported in Table 7 below.

**Table 7. Number of Students in each Gender or Ethnic Group**

Test	Female	Male	African-American	Asian-American	Hispanic-American
Grade 4 ELA	3,470	3,584	1,383	500	1,091

The standardized mean difference (SMD) statistic (Zwick, Donoghue, & Grima, 1993) was used to examine DIF on the operational data. The SMD statistics can provide DIF information for both multiple choice and constructed response items. The SMD takes into account the natural ordering of the response levels of the items and has the desirable property of being based on those ability levels where members of the focal group are present. The standardized mean difference output results in a single statistic for each item.

$$SMD = \sum p_{Fk} m_{Fk} - \sum p_{Rk} m_{Rk},$$

where  $p_{Fk}$  is the proportion of focal group members who are at the  $k$ th level of the matching variable,

$m_{Fk}$  is the mean item score for the focal group at the  $k$ th level, and

$m_{Rk}$  is the analogous value for the reference group.

The matching variable is raw score and the  $k$ th level refers to each successive raw score point.

A moderate amount of practically significant DIF, for or against the focal group, is represented by an SMD with an absolute value between .10 and .19, inclusive. A large amount of practically significant DIF is represented by an SMD with an absolute value of .20 or greater. SMD DIF results using operational data for G4 English Language Arts and Reading are summarized below.

**Table 8. Numbers of Items Flagged for DIF in G4 ELA and RD**

Focal Group	Direction of DIF	G4 ELA	G4 RD
Female	In favor of	3 <sup>1</sup>	0
	Against	0	0
African-American	In favor of	0	1 <sup>2</sup>
	Against	0	0
Asian-American	In favor of	1 <sup>3</sup>	0
	Against	0	0
Hispanic-American	In favor of	1 <sup>4</sup>	1 <sup>5</sup>
	Against	1 <sup>6</sup>	1 <sup>7</sup>

<sup>1</sup> Items #29, #30, #31 (SMD = .11, .13, and .10)

<sup>2</sup> Item #30 (SMD = .15)

<sup>3</sup> Item #30 (SMD = .10)

<sup>4</sup> Item #30 (SMD = .10)

<sup>5</sup> Item #30 (SMD = .14)

<sup>6</sup> Item #10 (SMD = -.10)

<sup>7</sup> Item #10 (SMD = -.10)

## Part 3: Scoring and Reliability

---

### Raw Score to Scale Score Conversion

To facilitate ease of interpretation and implementation, number-correct scoring was used on the New York State tests in 2005. In number-correct scoring, a student's scale score is derived directly from his or her raw, or number-correct, score. The relationship between raw scores and their corresponding scale scores is expressed in a raw score to scale score (RS-SS) table.

In IRT, all the item characteristic curves for the items on a test can be added together to yield a function, the test characteristic curve (TCC), that shows the expected raw score for each given scale score. By inverting the TCC, an expected scale score can be computed for each raw score. This new function, the inverse of the TCC, can be summarized in an RS-SS table. An advantage of RS-SS tables is that they make scoring relatively straightforward. With number-correct scoring, it is sufficient to know how many raw score points a student obtained on the test to determine a student's scale score. The RS-SS conversion tables for both content areas appear in Tables 9 and 10.

### Reliability

The reliability of measurement refers to the reproducibility or consistency of an individual's test score. The two most frequently reported indices of reliability are the standard error of measurement and the reliability coefficient.

The standard error of measurement is a measure of the extent to which an individual's scores vary over numerous parallel tests. We computed a *conditional* error, the standard error (SE) for each scale score for G4 ELA, and these are reported below in Table 9. See also the section on estimated conditional standard errors of scale scores.

The reliability coefficient is the correlation coefficient between scores on parallel tests and is an index of how well scores on one parallel test predict scores from another parallel test. The Feldt-Raju index was calculated to estimate the reliability of the G4 ELA test. This index is appropriate to use when a test contains both MC and CR items. The Feldt-Raju index for the G4 ELA test was 0.90, and the index for the RD test was 0.89, values comparable to those of 2004.

**Table 9. Raw Score to Scale Score with SE for G4 ELA 2005**

No. Correct (RS)	ELA	
	Scale Score	SE
0	455	122
1	455	122
2	455	122
3	455	122
4	455	122
5	455	122
6	498	79
7	529	48
8	546	33
9	558	26
10	567	22
11	575	19
12	582	16
13	588	15
14	593	13
15	598	12
16	603	11
17	607	11
18	610	10
19	614	10
20	617	9
21	621	9
22	624	9
23	627	9
24	631	9
25	634	9
26	638	9
27	641	9
28	645	9
29	649	10
30	653	10
31	658	11
32	663	11
33	669	12
34	675	13
35	683	14
36	691	15
37	701	16
38	712	18
39	727	21
40	747	24
41	774	30
42	800	42

**Table 10. Raw Score to Scale Score with SE for G4 RD 2005**

No. Correct (RS)	ELA	
	Scale Score	SE
0	455	139
1	455	139
2	455	139
3	455	139
4	455	139
5	455	139
6	520	74
7	552	42
8	567	28
9	577	21
10	584	18
11	591	15
12	596	14
13	601	12
14	605	11
15	609	11
16	613	10
17	617	10
18	620	9
19	624	9
20	627	9
21	630	9
22	634	9
23	637	9
24	641	9
25	645	9
26	649	10
27	653	10
28	658	11
29	664	12
30	670	13
31	679	15
32	690	18
33	709	26
34	800	116

**Estimated Conditional Standard Errors of Scale Scores**

Each student's scale score is based on a sample of the student's performance at a given time and inherently contains some measurement error. The classical SEM presumes the amount of measurement error is constant throughout the range of student ability. However, this is not realistic. Measurement error is less, and reliability greater, when more items exist and items are more informative. Item response theory lends itself to the calculation of a standard error for each scale score.

Tables 9 and 10 list standard errors for selected scale scores. These standard errors are "constrained" so that the upper and lower limits of one standard error band around a scale score are below the upper and

lower limits of the band for the next higher scale score. Typically, only standard errors on extreme ends are constrained. Because more items exist in the middle range of scale scores, the standard error is typically the smallest in the middle. A SS plus and minus one SE constitutes a 68% confidence interval. For example, for a student whose Grade 4 ELA SS is 621, we are 68% confident that his or her true score lies within the range 621 plus or minus 9, that is, between 612 and 630.

### **Lowest and Highest Obtainable Scale Scores**

A maximum likelihood procedure cannot produce scale score estimates for students with zero or perfect scores. Scale score estimates below the level expected by guessing are unreliable and subsequently not reported. Also, while maximum likelihood estimates may be available for students with extreme scores other than a perfect score, occasionally these estimates have standard errors that are very large, and differences between these extreme values have little meaning. Therefore, scores are established for these students based on a rational but necessarily non-maximum likelihood procedure. These values are called the lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS). The same LOSS and HOSS values are used for either number-correct or item-pattern scoring. For the New York State G4 ELA test, LOSS and HOSS values were set at 455 and 800.

### **Inter-Rater Agreement**

In order to monitor the reliability of scoring among the teachers who scored the student responses, approximately 10% of the student papers were submitted to a second group of raters provided by Measurement Incorporated. Note that the teachers were trained by Measurement Incorporated. The results of the inter-rater agreement analyses for public schools outside of New York City, non-public schools outside of New York City, and public schools within New York City, are provided in Tables 11-19.

**Table 11. G4 ELA & RD 2005 Inter-Rater Agreement: Public, Non-NYC, N=10,432**

Inter-Rater Agreement (Read 1: NYS school teachers; Read 2: MI readers)								
CR Item	Score Points	Agreement (%)			RS Mean		RD SD	
		Exact	Approx.	TOTAL	Read 1	Read 2	Read 1	Read 2
Listening	4	52.8	43.4	96.2	2.6	2.3	0.85	0.75
Independent Writing	3	52.0	42.8	94.8	2.0	1.9	0.84	0.87
Writing Mechanics	3	60.3	38.3	98.6	2.1	2.0	0.72	0.73
Reading (ELA)	4	51.4	43.7	95.1	2.4	2.1	0.92	0.80
#1 Reading Analytic	2	86.5	13.2	99.6	1.7	1.7	0.53	0.57
#2 Reading Analytic	2	78.8	20.3	99.0	1.7	1.6	0.61	0.65
#3 Reading Analytic	2	73.6	25.9	99.5	1.5	1.5	0.59	0.60
Approximate agreement (%) is the percent of pairs of reads that differ by one score point. Total agreement (%) is the sum of exact and approximate agreement percents.								

**Table 12. G4 ELA & RD 2005 Inter-Rater Agreement: Non-Public, Non-NYC, N=953**

Inter-Rater Agreement (Read 1: NYS school teachers; Read 2: MI readers)								
CR Item	Score Points	Agreement (%)			RS Mean		RD SD	
		Exact	Approx.	TOTAL	Read 1	Read 2	Read 1	Read 2
Listening	4	51.9	42.6	94.5	2.6	2.3	0.84	0.74
Independent Writing	3	49.3	44.7	94.0	2.0	1.8	0.87	0.84
Writing Mechanics	3	59.2	38.9	98.1	2.1	2.0	0.73	0.74
Reading (ELA)	4	47.4	46.0	93.4	2.4	2.0	0.92	0.79
#1 Reading Analytic	2	83.5	16.0	99.6	1.7	1.6	0.54	0.57
#2 Reading Analytic	2	77.5	21.5	99.1	1.6	1.6	0.63	0.65
#3 Reading Analytic	2	74.3	25.3	99.6	1.5	1.5	0.60	0.60
Approximate agreement (%) is the percent of pairs of reads that differ by one score point. Total agreement (%) is the sum of exact and approximate agreement percents.								

**Table 13. G4 ELA & RD 2005 Inter-Rater Agreement: Public, NYC, N=6,379**

Inter-Rater Agreement (Read 1: NYS school teachers; Read 2: MI readers)								
CR Item	Score Points	Agreement (%)			RS Mean		RD SD	
		Exact	Approx.	TOTAL	Read 1	Read 2	Read 1	Read 2
Listening	4	57.8	39.4	97.2	2.4	2.3	0.88	0.76
Independent Writing	3	52.6	42.4	95.0	1.9	1.8	0.84	0.89
Writing Mechanics	3	59.5	39.1	98.7	2.0	2.0	0.72	0.75
Reading (ELA)	4	54.0	43.2	97.2	2.0	2.0	0.91	0.82
#1 Reading Analytic	2	85.2	14.6	99.8	1.6	1.6	0.60	0.62
#2 Reading Analytic	2	76.4	22.7	99.1	1.6	1.5	0.66	0.70
#3 Reading Analytic	2	72.1	27.4	99.5	1.4	1.3	0.63	0.62

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

**Table 14. Percentages of Inter-Rater Score Differences: Public, Non-NYC**

Reader 1 (NYS school teachers) minus Reader 2 (MI readers)									
CR Item	-4	-3	-2	-1	0	1	2	3	4
Listening		0.01	0.55	11.82	52.81	31.60	3.13	0.06	0.02
Independent Writing		0.22	1.84	17.09	52.00	25.69	2.78	0.37	
Writing Mechanics		0.02	0.46	14.97	60.30	23.33	0.91	0.01	
Reading (ELA)		0.01	0.78	14.11	51.41	29.56	3.94	0.16	0.03
#1 Reading Analytic			0.04	3.48	86.45	9.71	0.33		
#2 Reading Analytic			0.26	8.13	78.75	12.15	0.72		
#3 Reading Analytic			0.10	8.91	73.62	17.01	0.36		

**Table 15. Percentages of Inter-Rater Score Differences: Non-Public, Non-NYC**

<b>Reader 1 (NYS school teachers) minus Reader 2 (MI readers)</b>									
<b>CR Item</b>	<b>-4</b>	<b>-3</b>	<b>-2</b>	<b>-1</b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>
Listening			0.42	7.24	51.94	35.36	4.72	0.21	0.10
Independent Writing		0.31	1.68	16.16	49.32	28.54	3.25	0.73	
Writing Mechanics			0.31	14.27	59.18	24.66	1.47	0.10	
Reading (ELA)			0.73	10.91	47.43	35.05	5.46	0.31	0.10
#1 Reading Analytic			0.10	6.19	83.53	9.86	0.31		
#2 Reading Analytic			0.63	10.49	77.54	11.02	0.31		
#3 Reading Analytic			0.21	13.22	74.29	12.07	0.21		

**Table 16. Percentages of Inter-Rater Score Differences: Public, NYC**

<b>Reader 1 (NYS school teachers) minus Reader 2 (MI readers)</b>									
<b>CR Item</b>	<b>-4</b>	<b>-3</b>	<b>-2</b>	<b>-1</b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>
Listening		0.05	0.85	17.12	57.77	22.29	1.88	0.03	0.02
Independent Writing		0.06	2.10	19.75	52.56	22.65	2.54	0.33	
Writing Mechanics		0.02	0.74	18.59	59.51	20.55	0.53	0.06	
Reading (ELA)		0.03	1.27	21.65	53.96	21.54	1.52	0.03	
#1 Reading Analytic			0.03	5.89	85.17	8.68	0.22		
#2 Reading Analytic			0.24	8.21	76.38	14.50	0.67		
#3 Reading Analytic			0.19	8.76	72.13	18.61	0.31		

**Table 17. Reliability Indices of Hand Scoring: Public, Non-NYC**

CR Item	Intra-Class Correlation <sup>1</sup>	Weighted Kappa <sup>2</sup>
Listening	0.78	0.57
Independent Writing	0.77	0.55
Writing Mechanics	0.79	0.59
Reading (ELA)	0.79	0.58
#1 Reading Analytic	0.88	0.76
#2 Reading Analytic	0.85	0.70
#3 Reading Analytic	0.80	0.61
<p>1 Agresti, A. (1990). Categorical data analysis (pp.366-367). New York: Wiley. Intra-class correlation is the percent of overall score variance accounted for by the variance of mean response scores.</p> <p>2 Weighted kappa is a measure of association in contingency tables, and is 1 when agreement is perfect and 0 when agreement is what would be expected by chance.</p>		

**Table 18. Reliability Indices of Hand Scoring: Non-Public, Non-NYC**

CR Item	Intra-Class Correlation	Weighted Kappa
Listening	0.75	0.52
Independent Writing	0.75	0.50
Writing Mechanics	0.78	0.57
Reading (ELA)	0.75	0.52
#1 Reading Analytic	0.86	0.72
#2 Reading Analytic	0.84	0.69
#3 Reading Analytic	0.81	0.62

**Table 19. Reliability Indices of Hand Scoring: Public, NYC**

CR Item	Intra-Class Correlation	Weighted Kappa
Listening	0.81	0.62
Independent Writing	0.79	0.57
Writing Mechanics	0.79	0.59
Reading (ELA)	0.82	0.64
#1 Reading Analytic	0.90	0.79
#2 Reading Analytic	0.86	0.72
#3 Reading Analytic	0.82	0.63

**Expected SPI Scores on the Standards at the Decision Points**

The current New York State Grade 4 ELA Score Reports for students report a Standard Performance Index (SPI) score for each of the learning standards. The SPI for a student, for a given learning standard, is an estimate of the percent of maximum raw score that the student would get if he or she took a large sample of items in that learning standard. The SPI is a diagnostic tool since it provides a profile of the student's relative strengths and weaknesses in terms of the content standards. However, just because a student has a high SPI on one learning standard and a low SPI on another learning standard does not necessarily mean that he or she is strong on the former standard and weak on the latter. This can occur if items measuring one learning standard tend to be easy, while items measuring another learning standard tend to be hard.

To better understand the relation between a given SPI score and performance on a learning standard, teachers and students should refer to the SPIs expected of students who are just at each of the New York State decision points. These expected SPIs at the decision points can be used as "reference points" against which each student's SPIs are compared. For example if a student's SPI on Standard 3 is 75 and the expected SPI for the Level 3 student is 74, the student's 75, although seemingly low compared with the perfect 100, is still higher than what is expected for the Level 3 student on the Standard. Expected SPIs for the 2005 Grade 4 English Language Arts exam are listed in Table 20.

**Table 20. G4 ELA 2005 Standard Performance Index Information**

Standard	Expected Percent of the Max. Raw Score at each of the Cut Points				
	# Items	Max Pts.	Level 2	Level 3	Level 4
			At SS=603	At SS=645	At SS=692
1	14	14	45	79	96
2	12	17	37	65	83
3	5	8	29	52	74

## Part 4: Descriptive Statistics

---

### Scale Score Frequency Distributions for the State and Subgroups

Table 21 summarizes the scale-score frequency distributions for the state and the following groups of students:

- public schools
- non-public schools
- limited English proficiency (LEP=5) students
- non-disabled students, and
- students with disabilities.

The public vs. non-public distinction was identified by the 7<sup>th</sup> and 8<sup>th</sup> characters of the BEDs code for each school. The non-disabled vs. disabled distinction was identified in the final state dataset. LEP students are defined as those who have "5" in the appropriate column of the final state dataset. The "LEP5" group is identified as limited English proficient and scored below a State-designated level of proficiency on the Language Assessment Battery-Revised (LAB-R) or the New York State English as a Second Language Achievement Test (NYSESLAT).

A summary table of the scale score frequency distributions containing the SSs at the 10<sup>th</sup>, 25<sup>th</sup>, 50<sup>th</sup>, 75<sup>th</sup>, and 90<sup>th</sup> percentiles is provided below. No interpolation was employed in computing the percentiles. As an example, in the row of Statewide Inclusive at the 25<sup>th</sup> percentile, the number 638 represents the highest scale score achieved by the lowest 25 percent of the population.

**Table 21. G4 ELA 2005 Summary of Scale Score Information**

<b>Sub Groups - Percentages</b>	<b>10<sup>th</sup></b>	<b>25<sup>th</sup></b>	<b>50<sup>th</sup></b>	<b>75<sup>th</sup></b>	<b>90<sup>th</sup></b>
Statewide Inclusive	617	638	663	691	712
LEP = 5	567	603	627	653	683
LEP = not 5	617	641	663	691	727
Public, LEP not 5	617	641	663	691	727
Non-Public, LEP not 5	617	641	663	691	712
Disabled, LEP not 5	567	598	627	649	675
Visually Impaired, LEP not 5	582	614	645	675	701
Non-Disabled, LEP not 5	627	645	669	691	727

## G4 ELA Scale Score Means and Standard Deviations

The scale score means, standard deviations, and the total number of students with valid scores in the clean data file are shown in the table below.

**Table 22. G4 ELA Statewide Scale Score Information**

Population Sub Grouping	Number of Students (N)	Scale Score Mean	Scale Score Standard Deviation
All Students	220,736	665.08	43.71
LEP = 5	4,602	625.76	48.68
LEP = not 5	216,134	665.92	43.21
Public, LEP not 5	191,358	665.87	43.18
Non-Public, LEP not 5	24,776	666.33	43.47
Disabled, LEP not 5	26,623	623.64	45.25
Visually Impaired, LEP not 5	88	643.53	46.44
Non-Disabled, LEP not 5	189,511	671.86	39.44

## G4 ELA Performance Level Distribution

The total number of students and the percent of students in each performance level in the statewide final general research file are shown in the table below. Full descriptions of the performance level assignments are posted on the NYSDE website, at:

<http://www.emsc.nysed.gov/osa/elaei/elaeiarch/elascorguide03.pdf>. Students in the Performance Level 1 (PL1) category exhibited only basic knowledge and skills in ELA on the assessment. Students in the Performance Level 2 (PL2) category demonstrated partial skills and knowledge that do not meet proficiency. Students in the Performance Level 3 (PL3) category are considered to be proficient and students in the Performance Level 4 (PL4) category are believed to possess advanced knowledge and skills in ELA. Statistics for the six previous years are also included.

**Table 23. G4 ELA Statewide Performance Level Information**

Year	Number of Students (N)	PCT in PL1	PCT in PL2	PCT in PL3	PCT in PL4
2005	220949	5.11	24.03	49.64	21.22
2004	231622	5.54	31.38	48.23	14.86
2003	238545	5.63	29.58	42.87	21.92
2002	242172	7.72	30.00	41.57	20.70
2001	245019	9.94	29.43	43.65	16.98
2000	247301	9.37	31.40	43.63	15.60
1999	240539	10.84	40.50	43.65	5.01

## References

---

- Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York, NY pp.366-367.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association, Inc.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Burket, G. R. (1988). *ITEMSYS* [Computer program]. Unpublished.
- Burket, G. R. (1991). *PARDUX* [Computer program]. Unpublished.
- Fitzpatrick, A. R. (1990). *Status Report on the results of Preliminary Analysis of Dichotomous and Multi-Level Items Using the PARMATE Program*. Unpublished manuscript
- Fitzpatrick, A. R. (1994). *Two studies comparing parameter estimates produced by PARDUX and BIGSTEPS*. Unpublished manuscript.
- Fitzpatrick, A. R., & Julian, M. W. (1996). *Two studies comparing the parameter estimates produced by PARDUX and PARSCALE*. Monterey, CA: CTB/McGraw-Hill.
- Green, D. R., Yen, W. M., & Burket, G. R. (1989). Experiences in the application of item response theory in test construction. *Applied Measurement in Education*, 2, 297-312.
- Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter Scaling of Rating Data* [Computer program]. Chicago, IL: Scientific Software, Inc.
- New York State Department of Education. (1995). *Test Access and Modification for Individuals with Disabilities*. Available at <ftp://unix2.nysed.gov/pub/education.dept.pubs/vesid/oses/test.access.mod/testacce.txt>

- New York State Department of Education. (2003). *Learning Standards for Mathematics*. Available at <http://www.emsc.nysed.gov/ciai/ela/pub/elalearn.pdf>
- Sandoval, J. H., & Mille, M. P. (1979, August). *Accuracy of judgments of WISC-R item difficulty for minority groups*. Paper presented at the annual meeting of the American Psychological Association, New York.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
- Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, 47, 175-186.
- Thissen, D. (1991). *MULTILOG* [Computer program]. Chicago, IL: Scientific Software, Inc.
- Wright, B. D., & Linacre, J. M. (1992). *BIGSTEPS Rasch Analysis* [Computer program]. Chicago, IL: MESA Press.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213.
- Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 36, 233-25.