

New York State Testing Program

English Language Arts Grade 8

Technical Report 2004



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 © 2004 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

FOREWORD	1
TABLE OF CONTENTS	2
LIST OF TABLES	3
PART 1: TEST DESIGN	4
<i>The New York State Learning Standards for English Language Arts</i>	4
<i>Test Configuration</i>	4
<i>Writing Mechanics Score</i>	4
STUDENT PARTICIPATION AND TESTING ACCOMMODATIONS	6
<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
ITEM DEVELOPMENT	7
ITEM REVIEW PROCESS	7
<i>Documenting Content</i>	7
<i>Minimizing Bias</i>	7
<i>Minimizing Speededness</i>	8
TEST CONSTRUCTION AND PRE-EQUATING	8
<i>Calibration Samples</i>	8
<i>Answer Choice Information</i>	8
<i>Item Response Theory Models</i>	8
<i>Equating Method</i>	10
<i>Procedures for Eliminating Bias and Minimizing Differential Item Functioning</i>	11
PART 2: ITEM STATISTICS FOR THE OPERATIONAL DATA	13
DATA CLEANING	13
ITEM ANALYSIS	14
DIFFERENTIAL ITEM FUNCTIONING ANALYSIS OF OPERATIONAL DATA	15
PART 3: SCORING AND RELIABILITY	17
RAW SCORE TO SCALE SCORE CONVERSION	17
RELIABILITY	17
ESTIMATED CONDITIONAL STANDARD ERRORS OF SCALE SCORES	19
LOWEST AND HIGHEST OBTAINABLE SCALE SCORES	19
INTER-RATER AGREEMENT	19
EXPECTED SPI SCORES ON THE STANDARDS AT THE DECISION POINTS	23
PART 4: DESCRIPTIVE STATISTICS	24
SCALE-SCORE FREQUENCY DISTRIBUTIONS FOR THE STATE AND SUBGROUPS	24
G8 ELA SCALE SCORE MEANS AND STANDARD DEVIATIONS	25
G8 ELA PERFORMANCE LEVEL DISTRIBUTION	25
REFERENCES	26

List of Tables

Table 1. New York State Learning Standards for English Language Arts.....	4
Table 2. Points per item type for G8 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. G8 ELA Item Level Statistics	14
Table 6. Number of Students in each Gender or Ethnic Group	15
Table 7. The Numbers of Items Flagged for DIF in G8 ELA	16
Table 8. Raw Score to Scale Score with SE for G8 ELA 2004	18
Table 9. G8 ELA 2004 Inter-Rater Agreement: Public, Non-NYC.....	20
Table 10. G8 ELA 2004 Inter-Rater Agreement: Non-Public, Non-NYC.....	20
Table 11. G8 ELA 2004 Inter-Rater Agreement: Public, NYC	20
Table 12. Percentages of Inter-Rater Score Differences: Public, Non-NYC	21
Table 13. Percentages of Inter-Rater Score Differences: Non-Public, Non-NYC	21
Table 14. Percentages of Inter-Rater Score Differences: Public, NYC	21
Table 15. Reliability Indices of Hand Scoring: Public, Non-NYC.....	22
Table 16. Reliability Indices of Hand Scoring: Non-Public, Non-NYC.....	22
Table 17. Reliability Indices of Hand Scoring: Public, NYC	22
Table 18. G8 ELA 2004 Standard Performance Index Information	23
Table 19. G8 ELA 2003 Summary of Scale Score Information.....	24
Table 20. G8 ELA Statewide Scale Score Information	25
Table 21. G8 ELA Statewide Performance Level Information.....	25

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 8 English Language Arts (G8 ELA) assessment is written to test students in Standards 1, 2, and 3.

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

Standard 1	Students will read, write, listen, and speak for information and understanding.
Standard 2	Students will read, write, listen, and speak for literary response and expression.
Standard 3	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 G8 ELA test has the following configuration: the test is divided into two sessions. There are 25 multiple choice (MC) items worth a total of 25 points and there are 9 constructed response (CR) items worth a total of 18 points. The CR items may be short response (SR) or extended response (ER) items. The total number of items on the test is 34, and the maximum raw score total is 43 points.

Session 1

Session 1 is comprised of 25 MC items, together with 3 SR items and 1 ER item. Each MC item in session 1 addresses one of the three tested New York State Learning Standards for English Language Arts. The four CR items in session 1 follow a listening passage and make up the listening cluster mapped to Standard 1. These items are scored together to derive a listening score, which can range from zero to six points.

Session 2

Session 2, Part 1, contains linked information stimuli, accompanied by three SR items and one ER item, which are scored together to derive a reading cluster score (zero to six points), that addresses Standard 3. Session 2 also contains an independent writing prompt addressing Standard 1. The prompt is followed by an ER item, which is scored to derive an independent writing score (zero to three points).

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 (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 8 ELA test.

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 8 ELA test.

Table 2. Points per item type for G8 ELA scores

Item Type or Cluster	Grade 8 ELA
Multiple choice (MC)	25 pts
Listening cluster	6 pts
Reading cluster	6 pts
Independent writing item	3 pts
Writing mechanics cluster	3 pts
Total points	43 pts
Total MC items and clusters	29 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

Condition Code	Meaning
A	Blank
F	Absent

Student Participation and Testing Accommodations

Students to be Tested

The New York State Testing Program (NYSTP) Grade 8 English Language Arts test must be administered to all public school students in Grade 8 and all ungraded students who are age-equivalent to students in Grade 8. This includes students who have been retained in Grade 8. 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://web.nysed.gov/vesid/sped/policy/changeaccomm.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 Department of Education, 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 Department of Education 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. Table 5 shows the omit rates for items on the G8 ELA 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 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 8 ELA scale.

Since these items are calibrated and on a common scale, the pool of available Grade 8 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 2004 operational NYSTP tests were constructed using items from the 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 θ 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 θ 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 α_j and γ_{ji} are the free parameters to be estimated from the data. Each item has $(m_j - 1)$ independent γ_{ji} parameters and one α_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 8 English Language Arts field test items were placed on the NYS G8 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 Department of Education. 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 calibration 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

Item analyses were conducted once CTB received data that met 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.

Initially, the state data set contained 252,127 cases. Table 4 below shows the data cleaning steps and the resulting size of the 85% sample used for conducting item analyses.

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

Steps Taken	# Cases Deleted	Ending N
Original Data		252,127
Duplicate Records	2	252,125
Grade Not Equal to 8	217	251,908
LEP5 Data	10,245	241,663
Invalid Data	5,998	235,665

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

As Table 4 shows, the following records were eliminated, in the order listed:

- Duplicated records
- Out-of-grade students, who were administered an 8th grade test despite not being 8th 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 scaling sample for the G8 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. Asterisked numbers indicate values for the correct response option.

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. G8 ELA Item Level Statistics

Raw Score Data		Test Administration Data						Reliability Feldt-Raju				P-Value Mean		
Mean	SD	Number of Items			Number of Students			ELA						
30.11	7.55	29			235,665			0.90				0.70		
Item	Omit	Pctsel0	Pctsel1	Pctsel2	Pctsel3	Pctsel4	Pctsel5	Pctsel6	p_bis1	p_bis2	p_bis3	p_bis4	Key	P-value
1	0.0002	0.02%	6.49%	*79.23%	11.80%	2.44%			-0.216	0.401	-0.269	-0.143	2	0.792
2	0.0003	0.01%	*76.85%	14.46%	4.06%	4.60%			0.447	-0.328	-0.120	-0.233	1	0.768
3	0.0005	0.02%	1.21%	12.44%	*78.64%	7.65%			-0.156	-0.237	0.368	-0.206	3	0.786
4	0.0004	0.01%	*92.63%	2.66%	1.18%	3.48%			0.304	-0.182	-0.151	-0.179	1	0.926
5	0.0009	0.02%	*79.43%	8.54%	6.80%	5.08%			0.440	-0.303	-0.235	-0.146	1	0.795
6	0.0013	0.01%	2.48%	11.33%	*74.84%	11.21%			-0.161	-0.187	0.401	-0.278	3	0.748
7	0.0009	0.02%	11.21%	33.11%	*46.23%	9.33%			-0.213	-0.114	0.332	-0.148	3	0.462
8	0.0008	0.02%	5.59%	3.65%	9.00%	*81.66%			-0.301	-0.205	-0.329	0.526	4	0.817
9	0.0020	0.02%	10.44%	*59.85%	15.92%	13.56%			-0.221	0.421	-0.140	-0.248	2	0.599
10	0.0012	0.03%	14.37%	2.14%	7.14%	*76.20%			-0.179	-0.237	-0.218	0.366	4	0.762
11	0.0013	0.02%	*50.56%	7.11%	17.08%	25.11%			0.339	-0.310	-0.178	-0.048	1	0.506
12	0.0010	0.02%	3.97%	*89.39%	4.58%	1.94%			-0.231	0.435	-0.276	-0.208	2	0.894

Table 5 continues

Table 5. G8 ELA Item Level Statistics (continued)

Item	Omit	Pctsel0	Pctsel1	Pctsel2	Pctsel3	Pctsel4	Pctsel5	Pctsel6	p_bis1	p_bis2	p_bis3	p_bis4	Key	p-value
13	0.0019	0.02%	12.89%	5.18%	2.79%	*78.94%			-0.305	-0.234	-0.212	0.472	4	0.789
14	0.0013	0.03%	*79.42%	4.74%	2.62%	13.07%			0.355	-0.229	-0.234	-0.162	1	0.794
15	0.0013	0.02%	7.75%	13.42%	6.46%	*72.22%			-0.090	-0.044	-0.173	0.189	4	0.722
16	0.0014	0.02%	2.08%	*64.56%	6.31%	26.89%			-0.217	0.149	-0.293	0.077	2	0.646
17	0.0033	0.02%	*60.09%	15.49%	6.87%	17.19%			0.402	-0.165	-0.165	-0.240	1	0.601
18	0.0025	0.02%	*66.53%	19.06%	7.73%	6.41%			0.438	-0.273	-0.220	-0.145	1	0.665
19	0.0028	0.03%	2.88%	26.69%	5.32%	*64.8%			-0.107	-0.244	-0.239	0.388	4	0.648
20	0.0029	0.02%	9.74%	*77.76%	3.82%	8.38%			-0.308	0.511	-0.240	-0.250	2	0.778
21	0.0036	0.02%	11.09%	11.08%	26.31%	*51.14%			-0.240	-0.218	0.005	0.297	4	0.511
22	0.0053	0.01%	3.63%	*89.84%	3.93%	2.07%			-0.268	0.414	-0.196	-0.198	2	0.898
23	0.0054	0.01%	*89.19%	5.34%	2.35%	2.58%			0.396	-0.220	-0.216	-0.202	1	0.892
24	0.0063	0.02%	2.68%	*86.98%	5.48%	4.21%			-0.229	0.459	-0.310	-0.185	2	0.870
25	0.0071	0.01%	*49.22%	9.35%	33.56%	7.16%			0.341	-0.247	-0.093	-0.172	1	0.492
26	0.0041	0.59%	4.93%	11.26%	19.74%	28.84%	23.63%	10.60%					CR	0.640 ⁺
27	0.0047	0.40%	4.10%	11.44%	24.68%	31.82%	20.04%	7.06%					CR	0.618 ⁺
28	0.0129	0.71%	13.07%	50.91%	34.02%								CR	0.723 ⁺
29	0.0048	0.72%	11.79%	49.65%	37.35%								CR	0.744 ⁺

+ average score divided by maximum score

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. For Grade 8, students bubbled in this information. 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 6 below.

Table 6. Number of Students in each Gender or Ethnic Group

Test	Female	Male	African-American	Asian-American	Hispanic-American
Grade 8 ELA	3,467	3,599	1,391	500	1,038

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 kth level of the matching variable,

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

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

The matching variable is raw score and the kth 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 G8 English Language Arts are summarized below.

Table 7. Numbers of Items Flagged for DIF in G8 ELA

Focal Group	Direction of DIF	G8 ELA
Female	In favor of	4 ¹
	Against	0
African-American	In favor of	1 ²
	Against	0
Asian-American	In favor of	0
	Against	0
Hispanic-American	In favor of	0
	Against	1 ³

¹ Items #26, #27, #28, 29 (D = .19, .19, .10 and .12)

² Item #16 (D = .10)

³ Item #9 (D = -.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 2004. 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 Table 8.

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 G8 ELA, and these are reported below in Table 8. 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 G8 ELA test. This index is appropriate to use when a test contains both MC and CR items. The Feldt-Raju index for the G8 ELA test was 0.90, a value comparable to that of 2003.

Table 8. Raw Score to Scale Score with SE for G8 ELA 2004

No. Correct (RS)	ELA	
	Scale Score	SE
0	527	113
1	527	113
2	527	113
3	527	113
4	527	113
5	527	113
6	591	49
7	610	31
8	620	21
9	627	17
10	633	14
11	638	12
12	642	11
13	646	10
14	650	10
15	653	9
16	656	9
17	659	8
18	662	8
19	665	8
20	667	7
21	670	7
22	672	7
23	675	7
24	677	7
25	680	7
26	682	7
27	685	7
28	687	7
29	690	7
30	693	8
31	696	8
32	700	8
33	703	9
34	708	9
35	712	10
36	718	11
37	724	12
38	732	14
39	742	16
40	755	19
41	774	24
42	804	32
43	830	41

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.

Table 8 lists 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 8 ELA SS is 670, we are 68% confident that his or her true score lies within the range 670 plus or minus 7, that is, between 663 and 677.

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 G8 ELA test, LOSS and HOSS values were set at 527 and 830.

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 9-17.

Table 9. G8 ELA 2004 Inter-Rater Agreement: Public, Non-NYC, N=7746

Inter-Rater Agreement (Read 1 : Non-NYC public school teachers; Read 2 : MI readers)								
CR Item	Score Points	Agreement (%)			RS Mean		RS SD	
		Exact	Approx.	TOTAL	Read 1	Read 2	Read 1	Read 2
Listening	6	35.4	47.2	82.5	3.8	3.4	1.38	1.17
Reading	6	35.2	46.7	82.0	3.7	3.2	1.30	1.10
Ind Writing	3	61.0	38.0	98.9	2.1	2.0	0.74	0.66
Writ Mech	3	56.4	42.4	98.9	2.2	2.0	0.72	0.65

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 10. G8 ELA 2004 Inter-Rater Agreement: Non-Public, Non-NYC, N=487

Inter-Rater Agreement (Read 1 : Non-NYC public school teachers; Read 2 : MI readers)								
CR Item	Score Points	Agreement (%)			RS Mean		RS SD	
		Exact	Approx.	TOTAL	Read 1	Read 2	Read 1	Read 2
Listening	6	32.4	46.6	79.1	4.2	3.7	1.32	1.11
Reading	6	31.6	48.9	80.5	4.0	3.3	1.24	1.10
Ind Writing	3	66.1	33.3	99.4	2.3	2.3	0.70	0.64
Writ Mech	3	60.8	38.2	99.0	2.3	2.2	0.71	0.66

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 11. G8 ELA 2004 Inter-Rater Agreement: Public, NYC, N=6636

Inter-Rater Agreement (Read 1 : Non-NYC public school teachers; Read 2 : MI readers)								
CR Item	Score Points	Agreement (%)			RS Mean		RS SD	
		Exact	Approx.	TOTAL	Read 1	Read 2	Read 1	Read 2
Listening	6	39.6	47.6	87.2	3.4	3.2	1.50	1.26
Reading	6	42.0	44.9	86.9	3.4	3.0	1.36	1.19
Ind Writing	3	61.6	37.2	98.8	2.0	1.8	0.80	0.70
Writ Mech	3	60.2	38.7	98.9	2.0	1.8	0.76	0.67

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

Reader 1 (Non-NYC public school teachers) minus Reader 2 (MI readers)										
CR Item	-4	-3	-2	-1	0	1	2	3	4	5
Listening	0.05	0.26	2.61	14.81	35.37	32.35	12.54	1.92	0.06	0.03
Reading		0.15	1.70	12.61	35.24	34.10	13.56	2.53	0.09	0.01
Independent Writing			0.25	10.95	60.97	27.02	0.79	0.03		
Writing Mechanics			0.15	9.23	56.44	33.19	0.93	0.05		

Table 13. Percentages of Inter-Rater Score Differences: Non-Public, Non-NYC

Reader 1 (Non-NYC public school teachers) minus Reader 2 (MI readers)										
CR Item	-4	-3	-2	-1	0	1	2	3	4	5
Listening			2.87	15.61	32.44	31.01	15.40	2.46	0.21	
Reading			0.82	10.88	31.62	37.99	15.20	3.49		
Independent Writing			0.41	12.11	66.12	21.15	0.21			
Writing Mechanics			0.82	13.14	60.78	25.05	0.21			

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

Reader 1 (Non-NYC public school teachers) minus Reader 2 (MI readers)										
CR Item	-4	-3	-2	-1	0	1	2	3	4	5
Listening	0.03	0.42	3.39	17.3	39.56	30.32	8.21	0.69	0.06	0.02
Reading		0.29	2.08	15.58	42.03	29.32	9.31	1.28	0.11	
Independent Writing			0.30	9.98	61.6	27.20	0.89	0.03		
Writing Mechanics		0.02	0.11	9.13	60.16	29.60	0.98	0.02		

Table 15. Reliability Indices of Hand Scoring: Public, Non-NYC

CR Item	Intra-Class Correlation ¹	Weighted Kappa ²
Listening	0.81	0.62
Reading	0.78	0.58
Independent Writing	0.79	0.58
Writing Mechanics	0.75	0.52
<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 16. Reliability Indices of Hand Scoring: Non-Public, Non-NYC

CR Item	Intra-Class Correlation ¹	Weighted Kappa ²
Listening	0.77	0.55
Reading	0.76	0.55
Independent Writing	0.80	0.60
Writing Mechanics	0.77	0.55

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

CR Item	Intra-Class Correlation ¹	Weighted Kappa ²
Listening	0.86	0.73
Reading	0.84	0.69
Independent Writing	0.82	0.64
Writing Mechanics	0.79	0.60

Expected SPI Scores on the Standards at the Decision Points

The current New York State Grades 4 and 8 Score Reports for students report a Standard Performance Index (SPI) score for each of the standards or Key Ideas. The SPI for a student, for a given Key Idea, 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 Key Idea. 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 Key Idea and a low SPI on another Key Idea 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 Key Idea tend to be easy, while items measuring another Key Idea tend to be hard.

To better understand the relation between a given SPI score and performance on a Key Idea, 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 66 and the expected SPI for the Level 3 student is 64, the student's 66, 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 2004 Grade 8 English Language Arts exam are listed in Table 18.

Table 18. G8 ELA 2004 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=658	At SS=697	At SS=737
1	13	20	40	72	88
2	8	8	34	82	98
3	7	12	34	64	85

Part 4: Descriptive Statistics

Scale-Score Frequency Distributions for the State and Subgroups

Table 19 summarizes the scale-score frequency distributions for the state and the following groups of students:

- public schools
- non-public schools
- two groups of limited English proficient (LEP) students
- non-disabled students, and
- students with disabilities.

The public vs. non-public distinction was identified by the 9th character 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 10th, 25th, 50th, 75th, and 90th percentiles is provided below. No interpolation was employed in computing the percentiles. As an example, in the row of Statewide Inclusive at the 25th percentile, the number 680 represents the highest scale score achieved by the lowest 25 percent of the population.

Table 19. G8 ELA 2003 Summary of Scale Score Information

Sub Groups - Percentages	10th	25th	50th	75th	90th
Statewide Inclusive	665	680	696	718	742
LEP = 5	627	646	662	677	690
LEP = not 5	665	682	696	718	742
Public, LEP not 5	665	680	696	718	742
Non-Public, LEP not 5	670	685	703	718	742
Disabled, LEP not 5	638	653	670	685	700
Visually Impaired, LEP not 5	646	656	682	700	732
Non-Disabled, LEP not 5	672	685	700	718	742

G8 ELA Scale Score Means and Standard Deviations

The scale score means, standard deviations, and the total number of students in the statewide final general research file are shown in the table below.

Table 20 G8 ELA Statewide Scale Score Information

Population Sub Grouping	Number of Students (N)	Scale Score Mean	Scale Score Standard Deviation
All Students	228,619	697.85	30.00
LEP = 5	5744	658.67	31.74
LEP = not 5	235665	700.69	33.07
Public, LEP not 5	210231	700.24	32.88
Non-Public, LEP not 5	25434	704.40	34.37
Disabled, LEP not 5	31041	688.80	29.65
Visually Impaired, LEP not 5	82	680.90	35.62
Non-Disabled, LEP not 5	204624	705.52	30.79

G8 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 five previous years are also included.

Table 21. G8 ELA Statewide Performance Level Information

Year	Population Sub Grouping	Number of Students (N)	PCT in PL1	PCT in PL2	PCT in PL3	PCT in PL4
2004	All Students	241546	6.94	44.76	37.04	11.26
2003	All Students	236490	9.05	44.86	38.30	7.79
2002	All Students	228849	6.78	47.94	34.88	10.40
2001	All Students	218235	12.94	41.06	34.89	11.11
2000	All Students	220405	12.58	41.18	36.24	9.99
1999	All Students	214735	8.76	41.65	40.68	8.91

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 English Language Arts*. 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.