New York State Testing Program

Mathematics Grade 4

Technical Report 2002

Developed and published under contract with New York State Department of Education by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc., 20 Ryan Ranch Road, Monterey, California 93940-5703. Copyright © 2003 by New York State Department of Education. 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 New York State Department of Education.

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**

THE NEW YORK STATE LEARNING STANDARDS FOR MATHEMATICS 4
TEST CONFIGURATION 4
STUDENT PARTICIPATION AND TESTING ACCOMMODATIONS 6
Students to be Tested *6*
Testing Accommodations *7*
ITEM DEVELOPMENT 7
ITEM REVIEW PROCESS 8
Documenting Content *8*
Minimizing Bias *8*
Minimizing Speededness *8*
TEST CONSTRUCTION AND PRE-EQUATING 8
Calibration Samples *8*
Answer Choice Information *9*
Item Response Theory Models *9*
Equating Method *10*
Item Selection Criteria and Process *11*
Procedures for Eliminating Bias and Minimizing Differential Item Functioning *12*

PART 2: ITEM STATISTICS FOR THE OPERATIONAL DATA **13**

DATA CLEANING 13
ITEM ANALYSIS 14
DIFFERENTIAL ITEM FUNCTIONING ANALYSIS OF OPERATIONAL DATA 17

PART 3: SCORING AND RELIABILITY **19**

RAW SCORE TO SCALE SCORE CONVERSION 19
RELIABILITY 19
ESTIMATED CONDITIONAL STANDARD ERRORS OF SCALE SCORES 21
LOWEST AND HIGHEST OBTAINABLE SCALE SCORES 22
INTER-RATER AGREEMENT 22
EXPECTED SPI SCORES ON THE STANDARDS AT THE DECISION POINTS 24

PART 4: DESCRIPTIVE STATISTICS **25**

SCALE-SCORE FREQUENCY DISTRIBUTIONS FOR THE STATE AND SUBGROUPS 25
G4 MA SCALE SCORE MEANS AND STANDARD DEVIATIONS 26

REFERENCES **27**

List of Tables

Table 1 Key Ideas for Grade 4 Mathematics 4
Table 2 G4 MA Test Design 4
Table 3 Condition Codes for the MA CR Items 5
Table 4 Steps Involved in Data Clean-up for Analysis Preparation 13
Table 5 G4 MA Item Level Statistics 15
Table 6 Number of Students in each Gender or Ethnic Group 17
Table 7 The Numbers of Items Flagged for DIF in G4 MA Summary 18
Table 8 Raw Score to Scale Score with SE for G4 MA 2002 20
Table 9 G4 MA 2002 Inter-Rater Agreement 22
Table 10 Percentages of Inter-Rater Score Differences 23
Table 11 Reliability Indices of Hand Scoring 23
Table 12 G4 MA 2002 Standard Performance Index Information 24
Table 13 G4 MA 2002 Summary of Scale Score Information 25
Table 14 G4 MA Statewide Scale Score Information 26

Part 1: Test Design

The New York State Learning Standards for Mathematics

The test measures progress toward the seven Key Ideas described in Standard 3 of the *Learning Standards for Mathematics, Science, and Technology* at: http://www.emsc.nysed.gov/ciai/pub/standards.pdf. The Grade 4 Mathematics test is written to test students in all seven Key Ideas and for each Key Idea students had the opportunity to demonstrate their knowledge both by selecting and generating responses. The seven Key Ideas are listed in Table 1 below with the approximate percent emphasis that is placed on each for Grade 4 Mathematics.

Table 1 Key Ideas for Grade 4 Mathematics

Key Ideas	Emphasis for Grade 4
Mathematical Reasoning	10–15%
Number and Numeration	15–25%
Operations	20–25%
Modeling/ Multiple Representation	5–10%
Measurement	15–20%
Uncertainty	5–10%
Patterns / Functions	10–15%

Test Configuration

Table 2 provides the test design for Grade 4 Mathematics, including the number of questions, question types, number of points, and time allotted for each testing session. Table 3 indicates the conditions codes used in scoring the responses to the CR items.

Table 2 G4 MA Test Design

	Number of Questions	Number of Points	Time in Minutes
Session 1	30 MC	30	40
Session 2	7 SR 2 ER	14 6	50
Session 3	7 SR 2 ER	14 6	50

Table 3 Condition Codes for the MA CR Items

Condition Code	Meaning
A	Blank
B	Refusal
C	Illegible
D	Other language

Student Participation and Testing Accommodations

Students to be Tested

It is the policy of the New York State Department of Education (NYSED) that the NYSTP Grade 4 Mathematics test be administered to all public school students. Nonpublic schools are strongly encouraged to administer the tests. The exceptions noted below, which represent the policy of the NYSED, apply to students in public and nonpublic schools participating in the NYSTP.

Students with Disabilities

All students with disabilities must be provided full access to the tests to the extent that such testing is consistent with their individual needs. The Committee on Special Education (CSE) must decide for each student on a case-by-case basis and document, on the individual student's individualized education program (IEP), whether the student will participate in the NYSTP or in the New York State Alternate Assessment for Students with Severe Disabilities. Criteria that the CSE must use to determine if a student should participate in the Alternate Assessment are given in the NYSTP School Administrator's Manual (SAM).

Students in Ungraded Classes

Both students with disabilities and general education students who are in ungraded classes are required to take the NYSTP tests, unless they meet the criteria for LEP exemption or are eligible for the Alternate Assessment. The chronological ages of students in ungraded classes should be used to determine who must be tested. Ungraded students should be tested on the grade 4 assessments no later than the school year (July 1 - June 30) in which they reach their 11th birthday.

Limited English Proficient (LEP) Students

LEP students scoring at or above the 30th percentile on a norm-referenced English reading test or the publisher's recommended score on an approved measure of English as a Second Language (ESL) in reading are required to participate fully in the ELA test. Those students must take the ELA test in English. They may take the Mathematics examination in an alternative language or in English, whichever would be better for the student. Alternative language editions of the Mathematics examination are provided in Chinese, Haitian Creole, Korean, Russian and Spanish.

LEP students scoring below the 30th percentile on a norm-reference English reading test or the publisher's recommended score on an approved measure of ESL in reading may be exempted from taking the English Language Arts examination but must take the Mathematics examination if it is available in their native language.

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.

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 NYSED for the use of testing accommodations.

Students with Disabilities

It is the responsibility of the principal to ensure that the testing accommodations specified in the IEP or Section 504 Accommodation Plan are provided to students with disabilities. 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 for students with disabilities are discussed in detail in the NYSED's 1995 publication titled *Test Access and Modification for Individuals with Disabilities*.

School administrators are to indicate in writing on the test book whether the student received "deletion of spelling, punctuation, and/or paragraphing requirements" or "use of scribe, tape recorder, word processor, or typewriter." If a student uses a typewriter or word processor, administrators are to staple the printed pages to the test book. For student receiving "deletion of spelling, paragraphing, and/or punctuation," teachers are to cross out and correct misspelled words and/or provide correct paragraphing and/or punctuation. For students using scribes, tape recorder or large type or Braille editions, teachers are to transcribe the students' text onto regular test answer documents and test books exactly as dictated or recorded.

Students Who Incur Disabilities Shortly Before Test Administration

School principals may modify testing procedures for general education students who incur an injury (e.g., a broken arm) or experience the onset of a short- or long-term disability (e.g., epilepsy) sustained or diagnosed within 30 days prior to the administration of State tests. More details are available in the SAM.

Limited English Proficiency

Limited English Proficient (LEP) students are allowed extended test time and tests may be administered to LEP students individually or in small groups in a separate location. LEP students may use bilingual dictionaries or glossaries when taking State examinations in ELA and Mathematics. The bilingual dictionaries and glossaries must not provide definitions or other explanations, only word-for-word translations. In addition, LEP students may use both an English and alternative language edition of the test simultaneously.

Item Development

A staff of professional item writers researched, collected, and wrote the field test material. All assessment materials were carefully reviewed for content and editorial accuracy. Artists and designers worked with the writers during development to ensure 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 Key Ideas for Mathematics. Standards Performance Index (SPI) scores are assigned to students on 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 both CTB Development staff and by New York State Department of Education staff and New York State teachers. This procedure ensures that items would be 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 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 calibration 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 shows the omit rates for items on the G4 MA test. These results provide little evidence of speededness on these tests.

Test Construction and Pre-equating

Calibration Samples

Three field test forms for the NYSTP tests were administered to students in public and private schools across the State in 2001. Efforts were made to ensure that the sample of students was representative of the state tested population. The 2001 field test items were calibrated and equated to the existing scale. Thus, parameters for these items were already on the appropriate New York State scale (one each for grade 4 ELA, grade 4 Mathematics, grade 8 ELA, and grade 8 Mathematics).

Since these items are calibrated and on a common scale, the pool of available grade 4 Mathematics items can be used to construct a test form and to produce a raw-score-to-scale-score table for that form. The 2002 operational NYSTP tests were constructed using the items in the pool. What follows is an overview of the analysis of field test data which results 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 low-performing students cannot answer correctly, but high-performing students can, 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. Starting in 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

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

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 4 Mathematics field test items were placed on the NYS G4 MA scale using the 2001 operational items as anchors. This was possible because the operational items were taken by the same students who took the field test items within the same testing window. 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 items selected was given by NYSED. Two criteria governed the item selection process. The first of these was to meet the content specifications specified by the New York State Department of Education. Within the limits set by these requirements, developers selected from the pool of field test items those with the best psychometric characteristics. Developers selected 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 immediately apparent as the final statistics are generated. Examples of possible faults that may occur are when the test is too easy or difficult, contains items demonstrating differential item functioning or DIF (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

Statistical differential item functioning (DIF) analyses were conducted for gender groups and such ethnic groups as African-American, Hispanic-American and Asian-American in the sample.

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.

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 244,121 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		244,121
Duplicate Completely Identical Records	661	243,460
Duplicate Identical Personal Info	62	243,398
Grade Not Equal to 4	0	243,460
LEP3 Data	8,284	235,114
Non-LEP3 Data		235,114
Non-LEP3 Data after Exclusion Rules	4,159	230,955

Students whose LEP status = 3 are not required to take the test.

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

- Duplicated records,

- Students whose limited English proficient (LEP) status was "3," indicating that they scored below the thirtieth 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. These students are not required to take the Mathematics test, unless a version of the test in their native language is available, and were consequently removed from consideration,

- All students who took the test in Chinese, Haitian-Creole, Russian, Korean, or Spanish, plus all students who took the English version AND whose LEP status is not 3, and

- 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 G4 MA test. The labels for the variables denote the following:

ITEM Item number.

OMIT Proportion of students who had 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 none or the first through the fourth answer option. 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.

PTBIS* Point-biserial correlations for each response option. Asterisked numbers indicate values for the correct response option.

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 MA Item Level Statistics

Raw Score Data			Test Administration Data					Reliability Feldt-Raju			P-Value Mean	
Mean		SD	Number of Items		Number of Students							
46.19		14.08	48		230,955			0.934			0.6598	
Item	Omit	Pctsel0	Pctsel1	Pctsel2	Pctsel3	Pctsel4	Ptbis1	Ptbis2	Ptbis3	Ptbis4	Key	p-value+
1	0.001		3.69%	5.72%	*89.54%	0.96%	-0.26	-0.27	*0.36	-0.12	3	0.895
2	0.001		4.11%	*85.11%	7.83%	2.86%	-0.18	*0.37	-0.29	-0.21	2	0.851
3	0.001		1.95%	2.13%	3.88%	*91.94%	-0.15	-0.26	-0.21	*0.33	4	0.919
4	0.001		2.78%	3.56%	*90.18%	3.34%	-0.16	-0.15	* 0.32	-0.28	3	0.902
5	0.001		10.52%	*85.15%	2.23%	2.03%	-0.35	*0.42	-0.16	-0.23	2	0.852
6	0.014		8.46%	7.38%	6.30%	*76.48%	-0.26	-0.19	-0.26	*0.42	4	0.765
7	0.001		8.00%	14.19%	*73.39%	4.35%	-0.27	-0.30	* 0.45	-0.23	3	0.734
8	0.013		11.26%	*74.66%	7.14%	5.61%	-0.20	*0.34	-0.20	-0.20	2	0.747
9	0.001		7.31%	1.08%	*90.52%	1.00%	-0.26	-0.13	*0.28	-0.13	3	0.905
10	0.002		11.94%	7.99%	*58.72%	21.16%	-0.21	-0.27	* 0.42	-0.25	3	0.587
11	0.001		*74.88%	10.36%	10.13%	4.50%	*0.31	-0.18	-0.25	-0.14	1	0.749
12	0.001		3.42%	2.95%	4.92%	*88.57%	-0.26	-0.20	-0.34	*0.45	4	0.886
13	0.002		2.29%	1.92%	*83.17%	12.47%	-0.17	-0.17	*0.43	-0.39	3	0.832
14	0.001		5.18%	*76.47%	16.12%	2.11%	-0.17	*0.26	-0.20	-0.18	2	0.765
15	0.001		11.23%	7.12%	*66.05%	15.51%	-0.23	-0.13	*0.44	-0.37	3	0.660
16	0.001		3.95%	*77.20%	7.11%	11.64%	-0.26	*0.42	-0.16	-0.33	2	0.772
17	0.001		19.88%	*62.79%	7.23%	9.96%	-0.33	*0.49	-0.20	-0.28	2	0.628
18	0.005		13.59%	7.30%	*69.89%	8.69%	-0.33	-0.21	*0.42	-0.19	3	0.699
19	0.002		27.31%	3.88%	*64.46%	4.20%	-0.25	-0.19	*0.35	-0.23	3	0.645
20	0.003		15.08%	24.28%	8.02%	*52.32%	-0.34	-0.17	-0.24	*0.45	4	0.523
21	0.006		8.42%	24.08%	*49.09%	17.80%	-0.19	-0.17	*0.40	-0.29	3	0.491
22	0.004		12.78%	8.98%	10.01%	*67.82%	-0.35	-0.29	-0.21	*0.51	4	0.678
23	0.006		10.52%	8.50%	17.65%	*62.77%	-0.26	-0.24	-0.30	*0.48	4	0.628
24	0.006		21.38%	*62.52%	5.11%	10.36%	-0.26	*0.54	-0.24	-0.40	2	0.625
25	0.008		4.17%	8.87%	9.36%	*76.85%	-0.26	-0.28	-0.29	*0.48	4	0.768
26	0.009		24.80%	3.41%	4.68%	*66.23%	-0.23	-0.22	-0.20	*0.34	4	0.662
27	0.010		*60.51%	17.85%	5.92%	14.76%	*0.54	-0.31	-0.19	-0.35	1	0.605
28	0.014		22.95%	*47.9%	9.57%	18.17%	0.01	*0.37	-0.28	-0.36	2	0.479
29	0.015		*55.55%	18.60%	16.92%	7.44%	*0.41	-0.16	-0.29	-0.23	1	0.555
30	0.015		10.31%	13.48%	24.90%	*49.8%	-0.22	-0.17	-0.22	*0.38	4	0.498
31	0.001	6.37%	36.37%	57.16%							CR	0.753
32	0.001	3.87%	41.11%	54.93%							CR	0.755
33	0.002	10.38%	10.64%	78.82%							CR	0.841
34	0.001	6.35%	37.09%	56.47%							CR	0.750
35	0.002	30.25%	33.69%	35.86%							CR	0.527
36	0.003	40.82%	24.46%	34.43%							CR	0.467
37	0.003	3.28%	10.45%	39.28%	46.65%						CR	0.763
38	0.005	36.43%	17.14%	45.90%							CR	0.545
39	0.017	20.75%	18.93%	21.26%	37.37%						CR	0.579
40	0.002	7.38%	22.22%	70.24%							CR	0.814

(Table 5 continues)

Table 5 G4 MA Item Level Statistics *(continued)*

Item	Omit	Pctsel0	Pctsel1	Pctsel2	Pctsel3	Pctsel4	Ptbis1	Ptbis2	Ptbis3	Ptbis4	Key	p-value+
41	0.014	19.14%	33.56%	45.87%							CR	0.627
42	0.004	22.03%	21.75%	20.34%	35.46%						CR	0.563
43	0.005	22.56%	29.22%	47.77%							CR	0.624
44	0.003	24.74%	44.89%	30.09%							CR	0.525
45	0.005	25.34%	28.77%	45.40%							CR	0.598
46	0.002	14.48%	41.92%	43.36%							CR	0.643
47	0.018	36.61%	39.67%	21.95%							CR	0.418
48	0.007	35.13%	17.82%	16.98%	29.37%						CR	0.466

+ 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 4, students bubble 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 4 Mathematics	3,558	3,719	1,406	500	1,284

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.

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

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 the 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 Mathematics are summarized below.

Table 7 The Numbers of Items Flagged for DIF in G4 MA Summary

Focal Group	Direction of DIF	Number of Items
Female	In favor of	1[1]
	Against	0
African-American	In favor of	2[2]
	Against	0
Asian-American	In favor of	2[3]
	Against	1[4]
Hispanic-American	In favor of	3[5]
	Against	0

[1] Item #37 (D =.12)

[2] Items #33 and #38 (D = .10 and .12)

[3] Items #26 and #39 (D =.11, .13)

[4] Item #36 (D = -.10)

[5] Items #26, #38 and #42 (D = .10, .11 and .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 2002. 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 tests scores. 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 MA and these are reported below in Table 8. See also the section on estimated conditional standard errors of scale scores, below.

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. Among several ways to estimate the reliability of a test, Cronbach's alpha (Cronbach, Schönemann, & McKie, 1965) probably is the most frequently used. Cronbach's alpha is a measure of internal consistency (i.e., how homogeneous test items are) that is appropriate for a test containing only MC items. Since the G4 MA test contains MC and CR items, Cronbach's alpha would underestimate reliability because of the effect of variance attributable to item types. A more appropriate index of internal consistency, the Feldt-Raju index, was used to estimate the reliability of the G4 MA test. It was 0.934, a value comparable to that for 2001.

Table 8 Raw Score to Scale Score with SE for G4 MA 2002

No. Correct (RS)	2002 G4 Mathematics			
	English, Russian, Haitian-Creole, Chinese, Korean		Spanish	
	Scale Score[1]	SE[1]	Scale Score[2]	SE[2]
0	448	115	448	116
1	448	115	448	116
2	448	115	448	116
3	448	115	448	116
4	448	115	448	116
5	448	115	448	116
6	465	98	488	76
7	506	57	517	47
8	525	38	533	33
9	537	29	544	27
10	547	24	553	22
11	555	21	560	20
12	561	18	566	18
13	567	17	572	16
14	572	15	577	15
15	576	14	581	14
16	580	13	585	13
17	584	13	589	12
18	588	12	593	12
19	591	11	596	11
20	594	11	599	11
21	597	11	602	11
22	600	10	605	10
23	602	10	608	10
24	605	10	611	10
25	607	9	613	10
26	610	9	616	9
27	612	9	618	9
28	614	9	621	9
29	616	9	623	9
30	618	8	625	9
31	621	8	627	9
32	623	8	630	8
33	625	8	632	8
34	626	8	634	8
35	628	8	636	8
36	630	8	638	8
37	632	8	640	8
38	634	8	642	8
39	636	8	644	8
40	638	7	647	8
41	639	7	649	8
42	641	7	651	8
43	643	7	653	8
44	645	7	655	8
45	647	7	658	9
46	648	7	660	9

(*Table 8 continues)*

Table 8 Raw Score to Scale Score with SEM for G4 MA 2002 *(continued)*

No. Correct (RS)	2002 G4 Mathematics			
	English, Russian, Haitian-Creole, Chinese, Korean		Spanish	
	Scale Score[1]	SE[1]	Scale Score[2]	SE[2]
47	650	8	663	9
48	652	8	665	9
49	654	8	668	10
50	656	8	671	10
51	658	8	674	10
52	660	8	678	11
53	662	8	682	12
54	665	8	686	12
55	667	9	691	13
56	669	9	697	15
57	672	9	704	17
58	675	9	713	20
59	678	10	727	25
60	681	10	751	38
61	684	11	810	90
62	688	11		
63	692	12		
64	697	13		
65	703	15		
66	710	17		
67	720	20		
68	733	25		
69	757	37		
70	810	81		

1 Scale scores for students who took the English, Russian, Haitian-Creole, Korean, and Chinese versions of the Grade 4 MA exam were computed using all of the items.

2 Scale scores for students who took (or had access to) the Spanish version of the Grade 4 MA exam were computed using all but items 6,8,39,41, and 47 due to translation errors.

Estimated Conditional Standard Errors of Scale Scores

Each student's scale score is based on a sample of student 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, where 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 who took the Chinese, Haitian-Creole, Russian, Korean, or English versions and whose grade 4 MA SS is 643, we are 68% confident that his or her true score lies within the range 643 plus or minus 7, that is, between 636 and 650.

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 MA test, LOSS and HOSS values were set at 448 and 810.

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 and outside of New York City are provided in Tables 9-11. Additional results for public schools in New York City and non-public school will be reported as they become available.

Table 9 G4 MA 2002 Inter-Rater Agreement

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
Item 31	2	98.58	1.18	99.76	1.57	1.57	0.57	0.56
Item 32	2	98.77	1.23	100.00	1.60	1.59	0.53	0.54
Item 33	2	95.70	3.50	99.20	1.71	1.71	0.64	0.64
Item 34	2	96.17	3.73	99.90	1.57	1.57	0.57	0.57
Item 35	2	93.00	6.81	99.81	1.16	1.13	0.78	0.80
Item 36	2	81.03	18.54	99.57	1.02	0.96	0.86	0.89
Item 37	3	82.07	16.60	98.67	2.40	2.34	0.68	0.74
Item 38	2	74.65	22.70	97.35	1.13	1.01	0.90	0.87
Item 39	3	80.61	19.02	99.63	1.85	1.93	1.12	1.11
Item 40	2	96.69	3.08	99.77	1.70	1.70	0.54	0.54
Item 41	2	90.54	9.23	99.77	1.36	1.34	0.71	0.74
Item 42	3	73.84	22.89	96.73	1.81	1.74	1.17	1.22
Item 43	2	75.73	23.66	99.39	1.37	1.35	0.77	0.79
Item 44	2	83.30	16.51	99.81	1.16	1.16	0.71	0.69
Item 45	2	85.34	14.33	99.67	1.31	1.32	0.79	0.80
Item 46	2	67.98	31.64	99.62	1.38	1.46	0.64	0.66
Item 47	2	79.99	19.63	99.62	0.93	0.93	0.77	0.76
Item 48	3	70.39	26.58	96.97	1.53	1.47	1.23	1.20

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 Percentages of Inter-Rater Score Differences

Reader 1 (Non-NYC public school teachers) minus Reader 2 (MI readers)							
CR Item	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**
Item 31		0.19	0.71	98.58	0.47	0.05	
Item 32			0.52	98.77	0.71		
Item 33		0.47	1.66	95.70	1.84	0.33	
Item 34		0.05	1.84	96.17	1.89	0.05	
Item 35		0.14	2.08	93.00	4.73	0.05	
Item 36		0.28	6.05	81.03	12.49	0.14	
Item 37		0.33	5.72	82.07	10.88	0.95	0.05
Item 38		0.57	6.43	74.65	16.27	2.08	
Item 39	0.09	0.14	13.34	80.61	5.68	0.05	0.09
Item 40		0.14	1.28	96.69	1.80	0.09	
Item 41		0.19	3.36	90.54	5.87	0.05	
Item 42		1.42	8.75	73.84	14.14	1.80	0.05
Item 43		0.24	11.12	75.73	12.54	0.38	
Item 44		0.14	8.47	83.30	8.04	0.05	
Item 45		0.24	7.33	85.34	7.00	0.09	
Item 46		0.19	20.10	67.98	11.54	0.19	
Item 47		0.24	9.89	79.99	9.74	0.14	
Item 48		1.37	10.50	70.39	16.08	1.47	0.19

Table 11 Reliability Indices of Hand Scoring

CR Item	Intra-Class Correlation[1]	Weighted Kappa[2]
Item 31	0.73	0.48
Item 32	0.96	0.92
Item 33	0.89	0.78
Item 34	0.86	0.71
Item 35	0.86	0.71
Item 36	0.93	0.87
Item 37	0.84	0.68
Item 38	0.97	0.94
Item 39	0.89	0.79
Item 40	0.93	0.87
Item 41	0.88	0.76
Item 42	0.96	0.92
Item 43	0.88	0.75
Item 44	0.92	0.84
Item 45	0.88	0.76
Item 46	0.96	0.92
Item 47	0.81	0.64
Item 48	0.93	0.86

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.

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.

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 in the sense that 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 necessary mean that she or he 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.

What teachers and students seem to need in order to better understand the SPIs are 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 Key Idea 4 is 60 and the expected SPI for the Level 3 Student is 53, the student's 60, although seemingly low compared with the perfect 100, is still higher than what is expected for the Level 3 Student on the key idea. Expected SPIs for the 2002 Grade 8 Mathematics exam are listed in Table 12.

Table 12 G4 MA 2002 Standard Performance Index Information

Key Idea	Expected Percent of the Max. Raw Score at each of the Cut Points for the **English, Russian, Haitian-Creole, Chinese, Korean** versions of the test.					Expected Percent of the Max. Raw Score at each of the Cut Points for the **Spanish** version of the test.				
	# Items	**Max Pts.**	**Level 2 At SS=602**	**Level 3 At SS=637**	**Level 4 At SS=678**	**# Items**	**Max Pts.**	**Level 2 At SS=602**	**Level 3 At SS=637**	**Level 4 At SS=678**
1	6	9	27	56	89	4	*5	31	62	92
2	11	14	38	61	87	10	*13	37	60	86
3	10	15	34	60	87	9	*13	35	61	88
4	4	7	36	53	78	4	7	36	53	78
5	6	11	39	62	84	6	11	39	62	84
6	5	6	29	56	90	5	6	29	56	90
7	6	9	21	42	74	5	*7	25	47	80

* Due to translation errors the following Key Ideas have fewer scored items on the Spanish version of the test: Key idea 1 (2 items worth 2 points each), 2 (1 item worth 1 point), 3 (1 item worth 1 point) and 7 (1 item worth 1 point).

Part 4: Descriptive Statistics

Scale-Score Frequency Distributions for the State and Subgroups

Table 13 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 LEA code for each school. The non-disabled vs. disabled distinction was identified in the final state dataset. Additionally, two groups of LEP students are defined as those who have either "2" or "3" in the appropriate column of the final state dataset. The "LEP2" group is identified as having limited English proficiency and scored at or above either the 30th percentile on a norm-referenced English reading test or the publisher's recommended score on an approved measure of English as a Second Language (ESL) in reading. Similarly, the "LEP3" group is identified as having limited English proficiency and scored below either the 30th percentile on a norm-referenced English reading test or the publisher's recommended score on an approved measure of English as a Second Language (ESL) in reading.

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 630 represents the highest scale score achieved by the lowest 25 percent of the population.

Table 13 G4 MA 2002 Summary of Scale Score Information

Sub Groups - Percentages	10th	25th	50th	75th	90th
Statewide Inclusive	610	630	652	675	692
LEP = 2	597	618	641	662	681
LEP = 3	567	594	616	636	656
Public	607	630	652	675	697
Non-Public	614	632	652	672	688
Disabled	572	602	628	650	669
Non-Disabled	614	634	654	675	697

G4 MA Scale Score Means and Standard Deviations

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. Statistics for the three previous years are also included.

Table 14 G4 MA Statewide Scale Score Information

Year	Population Sub Grouping	Number of Students (N)	PCT in PL1	PCT in PL2	PCT in PL3	PCT in PL4
2002	All Students	245,022	6.88	25.20	45.49	22.44
2001	All Students	249,119	8.16	22.57	43.14	26.14
2000	All Students	249,797	8.83	26.03	46.68	18.45
1999	All Students	245,358	9.61	23.59	43.15	23.64

References

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.

Cronbach, L. J., Schönemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, 25, 291–312.

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, Science, and Technology*. Available at http://www.emsc.nysed.gov/ciai/pub/standards.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.