FLORIDA COMPREHENSIVE ASSESSMENT TEST FOR READING AND MATHEMATICS

Technical Report For Test Administrations of FCAT 2003

Produced Jointly by
Human Resources Research Organization
(HumRRO)
Alexandria, Virginia

Under subcontract to and in cooperation with
Harcourt Educational Measurement
San Antonio, TX

Harcourt
Educational Measurement
A Harcourt Assessment Company

San Antonio, TX
October 28, 2003

Table of Contents

(Appendices may be purchased through the Florida Department of Education, Office of Assessment and School Performance.)

INTRODUCTION AND OVERVIEW

This report presents technical information about the measurement characteristics of the reading and mathematics assessments that were included in the Florida Comprehensive Assessment Test® (FCAT) for spring 2003. These characteristics provide an indication of the current quality of FCAT assessments in these two content areas.

Although this report is technical in nature, an attempt has been made to make it accessible to an audience familiar with basic testing concepts. Summary data is provided in the main body of the report while more detailed data may be found in the appendices.

Description of FCAT

As part of the student assessment and school accountability program of the Florida Department of Education (FDOE), FCAT assessments are designed to measure student achievement in specific reading and mathematics content as described by the Sunshine State Standards (SSS) Benchmarks (FDOE, 1996). Since 1998, the FCAT has included tests in reading for Grades 4, 8, and 10, and in mathematics for Grades 5, 8, and 10. In spring 2000, field-tests were administered in reading for Grades 3, 5, 6, 7, and 9 and in mathematics for students in Grades 3, 4, 6, 7, and 9. These new grade/subject test combinations for reading and mathematics became part of the FCAT in 2001. Since 2001, administration of the FCAT has included both reading and mathematics tests for Grades 3-10.

As seen in Table 1, the number of core items varied for mathematics tests across grades. For reading, however, the number of core items was identical for all grades.

Table 1. Number of Core Items by Subject and Grade

	Number of Core Items	
Grade	Mathematics	Reading
3	40	45
4	40	45
5	50	45
6	44	45
7	44	45
8	50	45
9	44	45
10	50	45

Test item formats vary depending on the subject and grade. The formats used are multiple choice (MC), gridded-response (GR), short-response (SR), and extended-response (ER) items. All tests include MC items. Mathematics tests in Grades 5 - 10 include GR items that require students to calculate numerical answers and fill in corresponding bubbles on an answer document. Both MC and GR items are machine-scored and are worth 1 point. Reading tests for Grades 4, 8, and 10 and mathematics tests for Grades 5, 8, and 10[1] also have performance or "constructed-response" tasks that require students to give a written response. The two types of performance tasks differ in the length of the response required and the number of points possible. The SR items are assigned 0, 1, or 2 points depending on the strength of the response. Similarly, student responses to ER items are assigned 0, 1, 2, 3, or 4 points. These items are hand-scored by trained raters using a process described later in this report.

In addition to core items on the FCAT, each test includes field-test items. To accommodate those items, ten separate test forms were constructed for each grade/subject combination. All forms within a grade/subject contained the same core items plus six to eight extra items. By having numerous forms for field-test items, the test allows a relatively large number of these items to be dispersed among subsets of students. Responses to field-test items do not contribute to students' scores.

Score reports consist of reading and mathematics scale scores, on a 100-to-500 scale, with subscores on performance category assignments and developmental scale scores. Performance category assignments are based on standard setting procedures that divide both the reading and mathematics scales into distinct levels of performance (FDOE, 1998, 2001 November 6). The FCAT reading tests report subscores in four reporting categories (also referred to as reading clusters):

- Words and Phrases
- Main Idea, Plot, and Purpose
- Comparisons and Cause/Effect
- Reference and Research

FCAT Mathematics tests provide subscores in five reporting categories (also referred to as mathematics strands):

- Number Sense, Concepts, and Operations
- Measurement
- Geometry and Spatial Sense
- Algebraic Thinking
- Data Analysis and Probability

The "developmental score" was created using vertical scaling techniques to place Grades 3 through 10 on a comparable metric, 0 to approximately 3000. Theoretically, students should receive higher scores as they move from grade-to-grade according to their increased achievement. To link an achievement scale from one grade to the next, a special data collection scheme was devised which incorporated the use of common items administered across more than one grade. These common items became the

[1] Grade/subjects that included performance tasks are referred to as "PT Grades" in this report.

basis for translating operational test results for all grades onto one scale (Hoffman, Wise, Thacker, and Ford, 2002).

Report Content

Test validity and reliability are key concerns for establishing the quality of an achievement test such as the FCAT. These two issues are intertwined, since measurement errors typically associated with the concept of reliability may also result in construct-irrelevant variance, one of the major threats to test validity (AERA, APA, NCME, 1999). Psychometric analysis, the major focus of this report, is fundamentally associated with relationships among test items as a means of examining item functioning and test reliability. This report presents test statistics as evidence of predictable patterns among test-item responses on several levels (item-level, test- or student-level, and state-level). Also included are background information on the process used to score the FCAT: item response theory or IRT, (Lord & Novick, 1968).

Summary statistics describe various technical attributes of the test. These attributes are illustrated in the report by the presentation of data about the calibration sample, traditional item statistics (*p*-values and item total correlations), IRT item statistics, a summary of the IRT test equating constants, IRT fit statistics, differential item functioning (DIF) statistics, test reliability, achievement scale unidimensionality, standard error of measurement, student classification accuracy and consistency, and intercorrelations among reporting categories and scale scores.

The FCAT is a continuous assessment system. While the essential structure and focus of the FCAT tests remain fairly fixed over time and student achievement results maintain a level of comparability across testing years, it must be stressed that the specific questions on a test administered in any given year show variability. In addition to variability of test questions administered on the “core” portion of the test (the portion of the test that actually contributes to reported student scores), it must also be recognized that every student will take some questions that do not count toward his or her ultimate score because the items are being field tested. Field-test items are newly-developed questions that are being tried out before they can be used on a future test. Field-test questions must be tried out at least one year before they are used to decide a student’s score. Although field-test items provide necessary data for the development of future tests, this technical report refers only to core-test items. A supplemental report (FDOE, 2003) presents summary data for the field-test items.

Although the bulk of this report concentrates on after-the-fact scoring and psychometric analyses, the success of the FCAT depends on the intense efforts required for item preparation, test assembly, and the hand scoring of performance-task items. Special sections will focus on these activities.

ITEM PREPARATION AND TEST ASSEMBLY

Prior to being included in the FCAT assessment, test items go though a three-phase development process. The first phase is drafting items to match the FCAT style and benchmarks.

Items are drafted by education professionals familiar with both the FCAT style and the intent of each of the SSS benchmarks. Draft items received by the contractor are subjected to a critical content and editorial review. Then items are forwarded to content staff at the Florida Department of Education Test Development Center (TDC) in Tallahassee, where they receive an additional review. Items submitted are typically accepted with no or minor edits, rejected as being inappropriate for the FCAT, or are returned to the contractor with comments regarding changes in style or focus that are necessary before the items can be moved further through the review process. A dialogue between the contractor and TDC staff on these "accept with revision" items assures that both the contractor and the TDC staff have deemed all items appropriate.

After this first phase of item writing, all FCAT items go through a rigorous review process before being considered for inclusion in a field test. The procedures used for item review for the FCAT 2003 field-test items are described in *Analysis of the FCAT Test Item Review Conducted by the Florida Department of Education and Harcourt Educational Measurement* (FDOE, 2003, May). Reviews were conducted by the following groups: (1) the FDOE for content, sensitivity/bias, match to benchmark, and FCAT style; (2) community sensitivity committees; (3) bias committees, with representatives from a variety of cultural backgrounds; and (4) content committees. The FDOE staff, as well as the committees representing the three other areas cited above, reviewed the reading passages on which the FCAT reading items were based. Item reviews were conducted following reading passage reviews. Similar procedures for passage and item reviews were followed in previous years for core items in the FCAT tests.

Once through the review process, these items are field tested during regular FCAT administrations. The items are quantitatively evaluated and placed in the item bank for possible use as core items in subsequent FCAT assessments.

Guided by both the content considerations required by the test blueprints for each content area and grade, as well as the statistical characteristics tied to each item, Harcourt and FDOE staff build forms through a process involving many steps. Typically, Harcourt content and psychometric staff propose draft forms of each grade and subject for TDC review. These draft forms are assembled according to the content guidelines documented for each test, as well as statistical guidelines documenting how well the proposed tests (whole tests as well as reportable strands/clusters) match the characteristics of previously administered versions of the FCAT.

CONSTRUCTED-RESPONSE SCORING PROCEDURES

Scorer Training

For some grade/content combinations, as has been noted earlier, students must provide handwritten responses to open-ended questions. These responses are judged by individual human scorers rather than by machines. Training of scorers is accomplished through the use of FDOE approved training materials that are agreed upon during the "Rangefinder Review" sessions held with state educators and members of the Test Development Center (TDC). Potential scorers are given an overview of the project and FDOE expectations and guidelines. They are shown several sets of training papers to ground them in the scoring rules. Scorers are then given "qualification sets" to ensure that a minimum agreement percentage can be met. Items are scored in groups of two or more (this process is known as the "rater item block" or RIB format), and the scorer must qualify on all items within the RIB to score the RIB. Only after the successful completion of the qualifying process are scorers allowed to assess actual student responses. To ensure consistency between training sessions (in the event that more than one group of scorers at separate times are trained on an item or group of items), papers are presented in the same order with the same comments. This is done so that each group of scorers will complete training using the same rules and information.

Year-to-Year Calibration

In order to ensure that an item scored in a previous administration is scored the same way in a current administration, all previous training materials are sent to the "Rangefinder Review" session and scoring rationales are discussed. Minimal changes are made to the training and validity sets, and the same scoring notes are used.

Read-Behinds

Read-behind is a process in which Team Leaders (and Scoring Directors, as needed) are required to review actual student responses which have been scored by members of their team (a team consists of no more than twelve scorers and one Team Leader). This process helps ensure that the scorers are assigning valid scores to student responses. At the beginning of the project, the Team Leaders are asked to spend their time doing read-behinds for everyone several times a day; this tends to identify the strength of individual scorers. Team Leaders ask scorers to review papers that have been incorrectly scored and help any scorer who has failed to adhere to the standards learn how his or her scoring has been in error. Throughout the project, read-behind is implemented for all scorers to ensure accuracy.

2003 FCAT Statistics

This section of the report presents psychometric analyses of the 2003 FCAT core assessments. Because of the requirements for rapid turnaround in score reporting, traditional item analyses and IRT analyses for the initial reporting period are conducted using a special calibration sample of students. Certain schools are chosen specifically for this purpose and those schools return their student responses on an early timeline. The general strategy is to select schools that provide a sample of students representative of the State's regions, ethnic diversity, and achievement scores in past years. Only standard curriculum students are used in the analyses: exceptional student education (ESE) students and students in the limited English proficiency (LEP) program for two or fewer years are excluded. In addition, students in the calibration sample have to meet criteria indicating they have attempted the test.[2] More details about the selection of this sample appear in *Plan for Selecting the Calibration Sample for the 2003 FCAT Administration* (FDOE, 2002, November).

Because of the importance of the calibration samples, this section begins with a comparison of the calibration samples to the State's total distributions of students. It is recognized that this presentation is out of chronological order, and was – in fact – conducted after all of the analyses were completed. However, the comparison is presented first to establish the credibility of the remaining analyses.

Calibration Sample Review

The tables on the following pages compare each grade/subject calibration sample with other statewide sets of students. One set of comparison students, labeled "total population," includes all students with FCAT records for 2003. Some of these students, however, did not receive FCAT scores because they failed the attemptedness criteria. A second set of students includes all standard curriculum students, again including those that did not receive test scores because of failing the attemptedness criteria. These two sets of students provide a basis for comparing the gender and ethnicity distributions of the calibration samples. Note also, that because of some missing ethnicity and gender information, the numbers of students across the respective categories do not sum to the totals listed.

In addition to the gender and ethnicity distributions, test scores for the calibration samples are compared to test scores for the total population that received scores and

[2] Test scores were computed only for students who met a criterion showing that they attempted to take the test. The criterion was that a student has at least six non-blank answers in each of two sessions.

for the total standard curriculum population that received test scores. Test score means for these groups are also disaggregated by ethnicity and gender.

The first table on each of the following pages examines ethnicity distributions. These tables indicate that ethnicity representations of the calibration samples are a reasonable approximation of the State distributions, and the match tends to be better for the standard curriculum distributions. The second table on each page examines gender distributions. These indicate results for gender similar to the ethnicity distributions. The last table on each page presents FCAT score means and standard deviations. As expected, score means are lower and standard deviations are higher for the total population of students than for standard curriculum students only. Score means for the calibration sample closely match those for the full set of standard curriculum students. Gender and ethnicity differences in the total standard curriculum samples are also replicated by the calibration samples. Detailed description of sampling procedures is presented in Appendix E.

FCAT 2003 Grade 3 Reading

Table 2. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	65 (1.39%)	1,232 (29.31%)	993 (21.20%)	16 (0.34%)	143 (3.05%)	2,233 (47.68)	4,683
Standard curriculum students	3,035 (1.92%)	37,192 (23.59%)	32,230 (20.42%)	412 (0.26%)	4,603 (2.92%)	80,337 (50.91%)	157,868
All scored students	3,563 (1.87%)	45,158 (23.68%)	41,853 (21.94%)	517 (0.27%)	5,378 (2.82%)	98,850 (49.21%)	190,720

[a] Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 3. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,295 (49.01%)	2,386 (50.95%)	4,683
Standard curriculum students	78,133 (49.51%)	79,694 (50.49%)	157,868
All scored students	92,420 (48.46%)	98,075 (51.42%)	190,720

[a] Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 4. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	297.03	49.56	4,683	308.51	56.60	156,933	298.39	62.83	188,282
Male	294.50	52.00	2,295	305.77	58.20	77,624	293.43	65.20	96,738
Female	299.47	46.98	2,386	311.22	54.84	79,273	303.69	59.75	91,421
African American	274.47	48.94	1,232	281.51	51.72	36,915	272.63	56.99	44,499
Hispanic	293.12	50.66	993	295.11	55.54	32,071	282.24	62.38	41,358
White	310.84	44.74	2,233	325.19	53.24	79,885	316.53	59.86	92,832

Note: N's for Gender and Ethnicity categories will not sum to equal the N of "All." For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Grade 3 Mathematics

Table 5. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	66 (1.41%)	1,227 (26.18%)	997 (21.27%)	17 (0.36%)	141 (3.01%)	2,238 (47.75%)	4,687
Standard curriculum students	3,035 (1.92%)	37,192 (23.59%)	32,230 (20.42%)	412 (0.26%)	4,603 (2.92%)	80,337 (50.91%)	157,868
All scored students	3,563 (1.87%)	45,158 (23.68%)	41,853 (21.94%)	517 (0.27%)	5,378 (2.82%)	98,850 (49.21%)	190,720

[a]Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 6. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,302 (49.14%)	2,382 (50.82%)	4,687
Standard curriculum students	78,133 (49.51%)	79,694 (50.49%)	157,868
All scored students	92,420 (48.46%)	98,075 (51.42%)	190,720

[a]Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 7. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	313.14	60.11	4,687	317.11	61.80	157,172	307.50	67.39	188,660
Male	318.86	61.97	2,303	321.67	62.40	77,780	309.77	69.12	96,973
Female	307.67	57.74	2,382	312.66	60.88	79,353	305.15	65.42	91,551
African American	281.36	59.88	1,227	283.40	60.96	36,951	274.49	65.25	44,554
Hispanic	314.82	59.58	997	308.37	61.64	32,090	296.34	68.11	41,393
White	328.75	53.43	2,238	334.51	54.92	80,067	326.44	60.84	93,086

Note: N’s for Gender and Ethnicity categories will not sum to equal the N of “All.” For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Grade 4 Reading

Table 8. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	66 (1.42%)	1,269 (27.29%)	1,025 (22.04%)	14 (0.30%)	120 (2.58%)	2,155 (46.24%)	4,650
Standard curriculum students	3,051 (1.92%)	38,036 (23.95%)	31,903 (20.09%)	469 (0.30%)	4,073 (2.56%)	81,163 (51.11%)	158,695
All scored students	3,613 (1.84%)	47,729 (24.27%)	42,282 (21.50%)	587 (0.30%)	4,867 (2.47%)	96,995 (49.33%)	196,625

[a] Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 9. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,258 (48.56%)	2,390 (51.40%)	4,650
Standard curriculum students	77,579 (48.87%)	81,171 (51.13%)	158,795
All scored students	100,606 (51.17%)	95,692 (48.67%)	196,625

[a] Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 10. Mean Scale Scores for Different Student Groups

	Calibration Sample			**All Scored Standard Curriculum Students**			**All Scored Students**		
	$\bar{X}$	**SD**	**N**	$\bar{X}$	**SD**	**N**	$\bar{X}$	**SD**	**N**
All	314.65	50.45	4,650	317.14	50.81	157,919	305.15	60.45	193,610
Male	312.24	50.08	2,258	315.13	51.21	77,107	300.53	62.38	98,999
Female	316.95	50.71	2,390	319.08	50.36	80,770	310.07	57.91	94,414
African American	289.81	49.65	1,269	281.51	51.72	36,915	279.43	60.95	46,839
Hispanic	312.23	50.95	1,025	295.11	55.54	32,071	286.94	64.17	41,699
White	329.60	45.24	2,155	325.19	53.24	79,885	315.73	57.56	95,836

Note: N's for Gender and Ethnicity categories will not sum to equal the N of "All." For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Grade 4 Mathematics

Table 11. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	65 (1.41%)	1,233 (26.77%)	1,021 (22.17%)	15 (0.33%)	121 (2.63%)	2,147 (46.61%)	4,606
Standard curriculum students	3,051 (1.92%)	38,036 (23.95%)	31,903 (20.09%)	469 (0.30%)	4,073 (2.56%)	81,163 (51.11%)	158,695
All scored students	3,613 (1.84%)	47,729 (24.27%)	42,282 (21.50%)	587 (0.30%)	4,867 (2.47%)	96,995 (49.33%)	196,625

[a]Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 12. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,244 (48.72%)	2,362 (51.28%)	4,606
Standard curriculum students	77,579 (48.87%)	81,171 (51.13%)	158,795
All scored students	100,606 (51.17%)	95,692 (48.67%)	196,625

[a]Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 13. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	306.66	54.23	4,606	308.98	55.44	157,680	297.75	63.44	193,720
Male	310.75	55.87	2,244	313.60	56.27	76,992	299.36	66.23	99,082
Female	302.78	52.34	2,362	304.59	54.27	80,648	296.13	60.31	94,438
African American	277.80	51.87	1,233	267.43	60.95	46,839	278.84	53.12	37,592
Hispanic	304.93	54.14	1,021	286.94	64.17	41,699	300.95	55.12	31,731
White	323.35	48.75	2,147	315.73	57.56	95,836	324.81	50.04	80,711

Note: N's for Gender and Ethnicity categories will not sum to equal the N of "All." For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Grade 5 Reading

Table 14. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	82 (1.85%)	1,133 (25.54%)	973 (21.93%)	10 (0.23%)	110 (2.48%)	2,127 (47.95%)	4,436
Standard curriculum students	3,124 (2.00%)	36,123 (23.14%)	30,953 (19.82%)	444 (0.28%)	3,759 (2.41%)	81,670 (52.31%)	156,140
All scored students	3,631 (1.85%)	46,499 (23.73%)	41,455 (21.16%)	532 (0.27%)	4,503 (2.30%)	98,807 (50.43%)	195,922

[a] Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 15. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,154 (48.56%)	2,282 (51.44%)	4,436
Standard curriculum students	75,575 (48.40%)	80,538 (51.58%)	156,140
All scored students	100,135 (51.11%)	95,499 (48.74%)	195,922

[a] Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 16. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	300.39	50.40	4,436	303.08	51.85	155,185	290.42	60.62	193,133
Male	301.18	49.01	2,154	302.95	51.75	75,066	287.48	62.08	98,599
Female	299.64	51.67	2,282	303.21	51.94	80,098	293.55	58.88	94,380
African American	273.48	48.28	1,133	274.76	49.18	35,795	262.06	56.89	45,682
Hispanic	295.98	47.43	973	290.27	51.06	30,749	273.57	61.65	40,823
White	315.73	46.82	2,127	319.39	46.81	81,284	309.44	54.90	97,716

Note: N's for Gender and Ethnicity categories will not sum to equal the N of "All." For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

Table 17. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	83 (1.85%)	1,175 (26.23%)	977 (21.81%)	10 (0.22%)	111 (2.48%)	2,123 (47.40%)	4,479
Standard curriculum students	3,124 (2.00%)	36,123 (23.14%)	30,953 (19.82%)	444 (0.28%)	3,759 (2.41%)	81,670 (52.31%)	156,140
All scored students	3,631 (1.85%)	46,499 (23.73%)	41,455 (21.16%)	532 (0.27%)	4,503 (2.30%)	98,807 (50.43%)	195,922

[a] Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 18. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,175 (48.56%)	2,302 (51.40%)	4,479
Standard curriculum students	75,575 (48.40%)	80,538 (51.58%)	156,140
All scored students	100,135 (51.11%)	95,499 (48.74%)	195,922

[a] Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 19. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	329.33	48.42	4,479	332.12	48.41	155,296	319.87	59.27	192,942
Male	332.68	48.53	2,175	334.70	49.65	75,144	319.33	62.36	98,513
Female	326.20	48.11	2,302	329.71	47.09	80,127	320.51	55.79	94,231
African American	302.18	50.96	1,175	305.33	50.29	35,867	291.00	62.04	45,637
Hispanic	330.45	44.29	977	326.52	47.41	30,792	311.56	60.04	40,815
White	342.00	42.49	2,123	344.70	42.55	81,278	335.23	51.77	97,621

Note: N's for Gender and Ethnicity categories will not sum to equal the N of "All." For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Grade 6 Reading

Table 20. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	74 (1.66%)	1,133 (25.45%)	791 (17.77%)	7 (0.16%)	77 (1.73%)	2,369 (53.22%)	4,451
Standard curriculum students	3,325 (2.09%)	37,619 (23.68%)	31,323 (19.72%)	481 (0.30%)	3,042 (1.92%)	82,972 (52.23%)	158,849
All scored students	3,790 (1.91%)	47,813 (24.15%)	41,465 (20.94%)	618 (0.31%)	3,626 (1.83%)	100,262 (50.64%)	197,974

[a]Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 21. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,217 (49.81%)	2,234 (50.19%)	4,451
Standard curriculum students	77,036 (48.50%)	81,774 (51.48%)	158,849
All scored students	101,474 (51.26%)	96,331 (48.66%)	197,974

[a]Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 22. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	307.26	53.29	4,451	308.03	56.10	158,336	294.65	64.59	196,970
Male	306.81	54.54	2,217	305.71	57.75	76,757	289.21	67.42	100,849
Female	307.70	52. 03	2,234	310.24	54.41	81,541	300.42	60.94	95,960
African American	284.28	48.44	1,133	280.45	52.11	37,410	267.45	59.66	47,436
Hispanic	292.19	53.76	791	294.86	56.88	31,232	277.35	66.72	41,258
White	322.30	49.89	2,369	324.19	51.33	82,773	313.26	59.39	99,893

Note: N's for Gender and Ethnicity categories will not sum to equal the N of "All." For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Grade 6 Mathematics

Table 23. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	74 (1.67%)	1,130 (25.43%)	789 (17.75%)	7 (0.16%)	77 (1.73%)	2,367 (53.26%)	4,444
Standard curriculum students	3,325 (2.09%)	37,619 (23.68%)	31,323 (19.72%)	481 (0.30%)	3,042 (1.92%)	82,972 (52.23%)	158,849
All scored students	3,790 (1.91%)	47,813 (24.15%)	41,465 (20.94%)	618 (0.31%)	3,626 (1.83%)	100,262 (50.64%)	197,974

[a] Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 24. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,213 (49.80%)	2,231 (50.20%)	4,444
Standard curriculum students	77,036 (48.50%)	81,774 (51.48%)	158,849
All scored students	101,474 (51.26%)	96,331 (48.66%)	197,974

[a] Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 25. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	317.80	52.28	4,444	316.00	55.42	158,189	302.36	66.42	196,763
Male	321.58	53.96	2,213	318.89	56.87	76,669	301.53	70.11	100,700
Female	314.04	50.29	2,231	313.29	53.88	81,483	303.30	62.25	95,903
African American	288.92	49.31	1,130	285.29	55.63	37,388	270.13	67.11	47,365
Hispanic	308.94	51.07	789	308.20	54.61	31,190	291.86	66.78	41,220
White	333.31	47.26	2,367	331.18	48.99	82,702	320.21	58.99	99,798

Note: N's for Gender and Ethnicity categories will not sum to equal the N of "All." For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Grade 7 Reading

Table 26. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	76 (1.73%)	1,105 (25.15%)	798 (18.17%)	13 (0.30%)	62 (1.41%)	2,339 (53.24%)	4,393
Standard curriculum students	3,358 (2.10%)	36,805 (23.07%)	31,975 (20.05%)	438 (0.27%)	2,145 (1.34%)	84,709 (53.11%)	159,512
All scored students	3,880 (1.95%)	47,170 (23.65%)	42,206 (21.16%)	558 (0.28%)	2,682 (1.34%)	102,424 (51.36%)	199,416

[a] Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 27. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,211 (50.33%)	2,182 (49.67%)	4,393
Standard curriculum students	76,979 (48.26%)	82,510 (51.73%)	159,512
All scored students	101,770 (51.03%)	97,469 (48.88%)	199,416

[a] Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 28. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	308.03	54.56	4,393	309.91	57.40	159,019	296.94	65.52	198,397
Male	306.96	55.40	2,211	308.12	57.99	76,701	292.25	67.38	101,115
Female	309.12	53.70	2,182	311.58	56.79	82,295	301.90	63.03	97,017
African American	286.473	51.05	1,105	281.87	54.27	36,632	268.41	62.15	46,809
Hispanic	289.59	57.92	798	296.16	58.41	31,871	280.13	67.23	41,989
White	323.68	49.55	2,339	326.15	52.32	84,505	315.84	59.63	102,027

Note: N's for Gender and Ethnicity categories will not sum to equal the N of "All." For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Grade 7 Mathematics

Table 29. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	76 (1.73%)	1,109 (25.25%)	799 (18.19%)	13 (0.30%)	61 (1.39%)	2,334 (53.14%)	4,392
Standard curriculum students	3,358 (2.10%)	36,805 (23.07%)	31,975 (20.05%)	438 (0.27%)	2,145 (1.34%)	84,709 (53.11%)	159,512
All scored students	3,880 (1.95%)	47,170 (23.65%)	42,206 (21.16%)	558 (0.28%)	2,682 (1.34%)	102,424 (51.36%)	199,416

[a] Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 30. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,211 (50.34%)	2,181 (49.66%)	4,392
Standard curriculum students	76,979 (48.26%)	82,510 (51.73%)	159,512
All scored students	101,770 (51.03%)	97,469 (48.88%)	199,416

[a] Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 31. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	307.19	50.23	4,392	307.86	53.35	158,849	295.34	62.45	198,120
Male	309.42	52.23	2,211	309.75	55.35	76,601	293.82	66.20	100,936
Female	304.93	48.04	2,181	306.11	51.35	82,225	297.00	58.22	97,017
African American	283.90	47.23	1,109	278.51	52.62	36,594	264.66	61.98	46,727
Hispanic	297.15	50.90	799	298.02	52.87	31,844	284.31	61.81	41,950
White	320.81	46.50	2,334	322.84	47.42	84,413	312.49	56.27	101,892

Note: N's for Gender and Ethnicity categories will not sum to equal the N of "All." For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Grade 8 Reading

Table 32. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	90 (2.01%)	1,144 (25.52%)	762 (17.00%)	10 (0.22%)	44 (0.98%)	2,431 (54.24%)	4,482
Standard curriculum students	3,488 (2.23%)	35,449 (22.71%)	30,499 19.54%)	450 (0.29%)	1,868 (1.20%)	84,250 (53.98%)	156,076
All scored students	4,017 (2.06%)	45,657 (23.42%)	40,441 (20.75%)	566 (0.29%)	2,380 (1.22%)	101,433 (52.04%)	194,931

[a] Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 33. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,188 (48.82%)	2,294 (51.18%)	4,482
Standard curriculum students	75,092 (48.11%)	80,940 (51.86%)	156,076
All scored students	99,225 (50.90%)	95,432 (48.96%)	194,931

[a] Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 34. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	314.04	47.04	4,482	312.64	48.86	155,396	300.54	57.50	193,630
Male	313.24	48.11	2,188	311.14	49.95	74,729	296.08	60.21	98,451
Female	314.81	45.98	2,294	314.05	47.78	80,627	305.26	54.10	94,935
African American	289.79	47.85	1,144	285.05	48.53	35,230	271.79	57.05	45,229
Hispanic	302.53	46.41	762	301.93	49.40	30,352	287.17	58.32	40,157
White	328.31	40.95	2,431	327.17	42.66	83,951	317.72	50.71	100,902

Note: N’s for Gender and Ethnicity categories will not sum to equal the N of “All.” For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Grade 8 Mathematics

Table 35. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	90 (2.01%)	1,145 (25.54%)	762 (16.99%)	10 (0.22%)	44 (0.98%)	2,432 (54.24%)	4,484
Standard curriculum students	3,488 (2.23%)	35,449 (22.71%)	30,499 (19.54%)	450 (0.29%)	1,868 (1.20%)	84,250 (53.98%)	156,076
All scored students	4,017 (2.06%)	45,657 (23.42%)	40,441 (20.75%)	566 (0.29%)	2,380 (1.22%)	101,433 (52.04%)	194,931

[a] Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 36. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,192 (48.88%)	2,292 (51.12%)	4,484
Standard curriculum students	75,092 (48.11%)	80,940 (51.86%)	156,076
All scored students	99,225 (50.90%)	95,432 (48.96%)	194,931

[a] Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 37. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	324.76	42.22	4,484	321.52	44.35	155,140	309.98	54.28	193,127
Male	325.73	42.75	2,192	322.20	45.25	74,559	307.78	56.93	98,091
Female	323.84	41.70	2,292	320.91	43.47	80,538	312.38	51.19	94,789
African American	300.31	43.16	1,145	294.40	45.63	35,153	280.33	57.42	45,070
Hispanic	317.23	39.62	762	313.47	43.34	30,288	300.66	53.29	40,050
White	337.72	36.69	2,432	334.51	37.94	83,843	325.58	46.44	100,688

Note: N's for Gender and Ethnicity categories will not sum to equal the N of "All." For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Grade 9 Reading

Table 38. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	145 (2.64%)	1,312 (23.88%)	1,149 (20.92%)	23 (0.42%)	52 (0.95%)	2,792 (50.81%)	5,495
Standard curriculum students	3,833 (2.24%)	40,356 (23.56%)	33,300 (19.44%)	859 (0.50%)	1,940 (1.13%)	90,781 (53.01%)	171,258
All scored students	4,343 (2.05%)	51,808 (24.43%)	43,533 (20.52%)	1,086 (0.51%)	2,486 (1.17%)	107,947 (50.89%)	212,099

[a] Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 39. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,708 (49.28%)	2,784 (50.66%)	5,495
Standard curriculum students	83,661 (48.85%)	87,510 (51.10%)	171,258
All scored students	109,236 (51.50%)	102,431 (48.29%)	212,099

[a] Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 40. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	295.83	53.36	5,495	301.10	55.04	169,497	290.16	60.81	209,145
Male	293.91	54.44	2,708	298.36	55.91	82,700	285.40	62.06	107,505
Female	297.74	52.24	2,784	303.74	54.05	86,721	295.36	58.92	101,272
African American	274.23	52.43	1,312	273.15	51.77	39,787	261.74	57.51	50,852
Hispanic	280.50	52.37	1,149	286.08	54.95	32,886	273.50	60.45	42,845
White	312.12	49.16	2,792	318.10	49.70	90,058	309.58	54.95	106,826

Note: N's for Gender and Ethnicity categories will not sum to equal the N of "All." For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Grade 9 Mathematics

Table 41. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	111 (2.45%)	1,087 (24.01%)	994 (21.96%)	19 (0.42%)	39 (0.86%)	2,255 (49.81%)	4,527
Standard curriculum students	3,833 (2.24%)	40,356 (23.56%)	33,300 (19.44%)	859 (0.50%)	1,940 (1.13%)	90,781 (53.01%)	171,258
All scored students	4,343 (2.05%)	51,808 (24.43%)	43,533 (20.52%)	1,086 (0.51%)	2,486 (1.17%)	107,947 (50.89%)	212,099

[a] Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 42. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,244 (49.57%)	2,280 (50.36%)	4,527
Standard curriculum students	83,661 (48.85%)	87,510 (51.10%)	171,258
All scored students	109,236 (51.50%)	102,431 (48.29%)	212,099

[a] Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 43. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	299.95	45.05	4,527	301.93	47.88	168,917	292.06	54.50	208,133
Male	302.89	46.02	2,244	304.44	48.72	82,382	292.10	56.53	106,877
Female	297.07	43.90	2,280	299.57	46.93	86,459	292.14	52.19	100,878
African American	277.93	44.46	1,087	274.81	45.88	39,592	263.58	53.67	50,470
Hispanic	289.42	43.23	994	290.06	48.31	32,741	280.06	54.23	42,616
White	314.36	40.86	2,255	316.98	41.67	89,849	309.12	47.64	106,454

Note: N's for Gender and Ethnicity categories will not sum to equal the N of "All." For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Grade 10 Reading

Table 44. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	118 (2.68%)	988 (22.46%)	832 (18.92%)	12 (0.27%)	44 (1.00%)	2,398 (54.52)	4,398
Standard curriculum students	3,397 (2.55%)	28,207 (21.17%)	24,587 (18.45%)	586 (0.44%)	1,417 (1.06%)	74,949 (56.24%)	133,267
All students	4,117 (2.32%)	40,359 (22.73%)	35,383 (19.93%)	787 (0.44%)	2,101 (1.18%)	93,612 (52.73%)	177,525

[a] Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 45. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	2,042 (46.43%)	2,352 (53.48%)	4,398
Standard curriculum students	70,936 (53.23%)	62,261 (46.72%)	133,267
All students	87,825 (49.47%)	88,854 (50.05%)	177,525

[a] Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 46. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	309.42	48.74	4,398	313.93	48.80	131,019	301.87	56.97	169,227
Male	307.73	50.17	2,042	313.50	49.31	61,167	299.11	58.89	83,486
Female	311.01	47.21	2,352	314.34	48.33	69,811	304.80	54.74	85,268
African American	285.52	46.39	988	285.46	47.66	27,493	272.50	55.69	37,811
Hispanic	299.00	50.87	832	300.44	52.02	24,109	285.11	61.04	33,564
White	323.13	43.13	2,398	328.21	41.87	74,000	319.85	48.25	90,457

Note: N's for Gender and Ethnicity categories will not sum to equal the N of "All." For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Grade 10 Mathematics

Table 47. Frequency Distributions for Different Student Groups, by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	115 (2.68%)	959 (22.36%)	814 (18.98%)	12 (0.28%)	41 (0.96%)	2,341 (54.59%)	4,288
Standard curriculum students	3,397 (2.55%)	28,207 (21.17%)	24,587 (18.45%)	586 (0.44%)	1,417 (1.06%)	74,949 (56.24%)	133,267
All scored students	4,117 (2.32%)	40,359 (22.73%)	35,383 (19.93%)	787 (0.44%)	2,101 (1.18%)	93,612 (52.73%)	177,525

[a]Total will not be equal to sum of ethnic group frequencies because a small percentage of students did not mark ethnicity.

Table 48. Frequency Distributions for Different Student Groups, by Gender

	Male	Female	Total[a]
Calibration sample	1,984 (46.27%)	2,299 (53.61%)	4,288
Standard curriculum students	70,936 (53.23%)	62,261 (46.72%)	133,267
All scored students	87,825 (49.47%)	88,854 (50.05%)	177,525

[a]Total will not be equal to sum of male and female groups because a small percentage of students did not mark gender.

Table 49. Mean Scale Scores for Different Student Groups

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	324.17	40.16	4,288	329.12	41.90	130,079	319.37	49.17	167,457
Male	328.70	39.56	1,984	333.61	41.61	60,524	321.49	50.58	82,275
Female	320.39	40.16	2,299	325.25	41.76	69,511	317.47	47.59	84,700
African American	299.47	41.69	959	301.07	44.09	27,301	289.29	52.37	37,438
Hispanic	317.24	39.40	814	320.06	41.02	23,771	310.03	47.89	32,783
White	335.99	34.43	2,341	341.46	35.14	73,625	334.21	41.30	89,890

Note: N's for Gender and Ethnicity categories will not sum to equal the N of "All." For gender, small percentages of students did not respond. For ethnicity, only the three most populous categories are shown.

FCAT 2003 Item Analysis

This section contains traditional item analysis statistics: difficulty and item-total correlations. For each of the items on the 16 tests (2 subjects and 8 grades), item difficulties (*p*-values), item-total test correlations, and correlations between the item and reporting categories within each of the subject areas were computed. Complete results appear in Appendices A (Reading) and B (Mathematics).

Item Difficulty Summary

Tables 50-55 summarize the item analysis results by presenting the minimum, 25th-percentile, 50th- percentile, 75th-percentile, and maximum values for each grade/subject test (across all items).

For MC and GR items, *p*-values are simply the mean points across students. For these items, *p*-values also correspond to the proportion of students who answered the item correctly. To facilitate comparisons among all item types, item difficulties for the PT items were computed as the mean points achieved divided by total possible points.

Tables 50 and 51 illustrate the distribution of *p*-values for all reading and mathematics items, respectively. For a test to be effective, *p*-values should show that the items vary in difficulty, but they should not be too high (e.g., above 0.90) or too low (e.g., near chance, 0.20, for the multiple-choice items or less than 0.10 for the other item types). Tables 50 and 51 show that there were some high *p*-values, which were monitored during IRT processing, but more generally the item *p*-values were dispersed across a sufficient range to establish satisfactory measurement reliability across a wide range of achievement.

Table 50. Proportional* *p*-value Summary Data for All Reading Items

Grade	Number of Items	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
3	45	0.345	0.576	0.691	0.770	0.865
4	45	0.194	0.544	0.641	0.789	0.896
5	45	0.266	0.542	0.687	0.789	0.933
6	45	0.386	0.583	0.693	0.779	0.957
7	45	0.415	0.611	0.677	0.776	0.871
8	45	0.361	0.579	0.687	0.747	0.911
9	45	0.318	0.521	0.647	0.744	0.880
10	45	0.218	0.572	0.635	0.726	0.907

*Mean score divided by total possible score.

Table 51. Proportional* *p*-value Summary Data for All Mathematics Items

Grade	Number of Items	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
3	40	0.312	0.467	0.602	0.707	0.941
4	40	0.283	0.482	0.604	0.711	0.887
5	50	0.213	0.422	0.577	0.658	0.908
6	44	0.221	0.458	0.561	0.653	0.956
7	44	0.129	0.417	0.531	0.691	0.922
8	50	0.251	0.481	0.560	0.685	0.859
9	44	0.151	0.356	0.472	0.597	0.960
10	50	0.129	0.436	0.542	0.646	0.833

*Mean score divided by total possible score.

Pearson Item-Total Correlations

Tables 52 and 53 show the distribution of item-total raw score correlations and correlations between items and reporting category scores. These are computed as Pearson correlations. For the MC and GR items, these correlations are equivalent to point-biserial correlations between the dichotomous variable (right and wrong) and total score. Total scores and reporting category scores for these correlations are based on sums of the appropriate points per item—that is, the sum of all item scores for total scores, and the sum of items according to the reporting categories they represent. Distributions for the item-reporting category correlations include only correlations of items for the matching reporting categories. Correlations for all items are presented in Appendices A and B.

The most important criteria for the correlation statistics is that they are neither negative nor near zero. Items with negative correlations should not be used in IRT processing. Tables 52 and 53 show that no negative correlations were observed.

Biserial Item-Total Correlations

The point-biserial correlations produced for dichotomous items are restricted in possible range to the extent that the items are either very easy or very difficult. Biserial correlations adjust for item distributions and therefore offer an alternative statistic. Biserial correlations, however, which are presented in Tables 54 and 55, can exceed 1.

Table 52. Summary Data for Reading Item Total Correlations for All Items

Grade	Reporting Category	No. of Items	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
3	Total	45	0.267	0.390	0.442	0.490	0.570
	Word & Phrases	11	0.425	0.467	0.487	0.532	0.548
	Main Idea	20	0.309	0.389	0.469	0.517	0.585
	Relationships	10	0.389	0.458	0.514	0.542	0.581
	Research Ref.	4	0.607	0.618	0.637	0.645	0.646
4	Total	45	0.243	0.356	0.405	0.458	0.569
	Word & Phrases	7	0.501	0.534	0.547	0.581	0.595
	Main Idea	19	0.367	0.400	0.436	0.496	0.563
	Relationships	14	0.299	0.404	0.442	0.501	0.618
	Research Ref.	5	0.443	0.577	0.583	0.602	0.607
5	Total	45	0.197	0.354	0.397	0.451	0.520
	Word & Phrases	8	0.445	0.521	0.530	0.545	0.574
	Main Idea	16	0.272	0.391	0.452	0.477	0.525
	Relationships	16	0.297	0.417	0.449	0.497	0.563
	Research Ref.	5	0.429	0.492	0.512	0.569	0.588
6	Total	45	0.145	0.330	0.411	0.454	0.525
	Word & Phrases	9	0.322	0.404	0.464	0.517	0.547
	Main Idea	19	0.214	0.356	0.444	0.476	0.517
	Relationships	10	0.301	0.442	0.535	0.549	0.568
	Research Ref.	7	0.463	0.488	0.506	0.545	0.555
7	Total	45	0.226	0.352	0.411	0.478	0.566
	Word & Phrases	8	0.405	0.435	0.476	0.494	0.567
	Main Idea	17	0.367	0.416	0.466	0.481	0.534
	Relationships	11	0.413	0.446	0.499	0.527	0.599
	Research Ref.	9	0.355	0.487	0.505	0.597	0.605
8	Total	45	0.189	0.326	0.391	0.434	0.641
	Word & Phrases	6	0.468	0.492	0.509	0.528	0.531
	Main Idea	19	0.238	0.347	0.399	0.439	0.484
	Relationships	13	0.388	0.420	0.467	0.479	0.641
	Research Ref.	7	0.419	0.460	0.499	0.516	0.811
9	Total	45	0.229	0.320	0.391	0.455	0.529
	Word & Phrases	7	0.444	0.501	0.558	0.574	0.610
	Main Idea	18	0.290	0.366	0.426	0.488	0.540
	Relationships	10	0.427	0.467	0.496	0.533	0.592
	Research Ref.	10	0.374	0.378	0.447	0.479	0.488
10	Total	45	0.235	0.315	0.372	0.427	0.602
	Word & Phrases	9	0.369	0.416	0.513	0.516	0.578
	Main Idea	14	0.357	0.376	0.408	0.449	0.462
	Relationships	11	0.358	0.399	0.434	0.513	0.621
	Research Ref.	11	0.276	0.364	0.407	0.485	0.725

Table 53. Summary Data for Mathematics Item Total Correlations for All Items

Grade	Reporting Category	No. of Items	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
3	Total	40	0.198	0.333	0.415	0.476	0.561
	Number	12	0.312	0.449	0.521	0.562	0.598
	Measurement	8	0.373	0.435	0.527	0.550	0.558
	Geometry	7	0.467	0.477	0.480	0.494	0.507
	Algebra	6	0.485	0.505	0.521	0.567	0.574
	Data	7	0.377	0.518	0.537	0.574	0.625
4	Total	40	0.185	0.353	0.410	0.445	0.605
	Number	11	0.402	0.465	0.489	0.553	0.626
	Measurement	8	0.442	0.461	0.523	0.529	0.580
	Geometry	7	0.373	0.450	0.470	0.510	0.563
	Algebra	7	0.378	0.464	0.528	0.571	0.605
	Data	7	0.384	0.506	0.525	0.543	0.585
5	Total	50	0.221	0.372	0.450	0.524	0.616
	Number	12	0.403	0.426	0.493	0.581	0.639
	Measurement	11	0.396	0.533	0.576	0.595	0.653
	Geometry	9	0.310	0.371	0.444	0.479	0.777
	Algebra	10	0.412	0.477	0.507	0.579	0.643
	Data	8	0.360	0.394	0.482	0.527	0.756
6	Total	44	0.142	0.281	0.373	0.448	0.543
	Number	9	0.349	0.409	0.472	0.518	0.537
	Measurement	9	0.339	0.488	0.535	0.577	0.595
	Geometry	9	0.375	0.416	0.451	0.465	0.522
	Algebra	8	0.278	0.439	0.489	0.523	0.577
	Data	9	0.253	0.394	0.475	0.483	0.531
7	Total	44	0.121	0.343	0.411	0.453	0.551
	Number	9	0.360	0.491	0.500	0.506	0.564
	Measurement	9	0.385	0.471	0.514	0.540	0.551
	Geometry	8	0.298	0.453	0.471	0.556	0.559
	Algebra	9	0.440	0.447	0.511	0.539	0.602
	Data	9	0.201	0.478	0.507	0.538	0.583
8	Total	50	0.246	0.350	0.439	0.547	0.743
	Number	11	0.405	0.473	0.516	0.551	0.624
	Measurement	11	0.369	0.437	0.475	0.603	0.734
	Geometry	8	0.334	0.398	0.497	0.637	0.844
	Algebra	11	0.340	0.405	0.492	0.548	0.690
	Data	9	0.347	0.426	0.511	0.576	0.756
9	Total	44	0.224	0.318	0.396	0.498	0.634
	Number	8	0.422	0.431	0.469	0.534	0.567
	Measurement	7	0.420	0.475	0.576	0.610	0.627
	Geometry	11	0.336	0.392	0.561	0.620	0.652
	Algebra	10	0.421	0.437	0.481	0.508	0.580
	Data	8	0.299	0.459	0.488	0.576	0.591
10	Total	50	0.217	0.359	0.417	0.536	0.734
	Number	10	0.423	0.448	0.524	0.581	0.666
	Measurement	9	0.414	0.465	0.512	0.579	0.728
	Geometry	10	0.309	0.444	0.496	0.585	0.816
	Algebra	13	0.370	0.414	0.463	0.548	0.618
	Data	8	0.303	0.397	0.457	0.511	0.820

Table 54. Summary Data for Biserial Correlations for All Reading Items by Reporting Categories

Grade	Reporting Category	No. of Items	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
3	Total	45	0.334	0.495	0.610	0.648	0.746
	Word & Phrases	11	0.548	0.604	0.678	0.731	0.767
	Main Idea	20	0.388	0.541	0.623	0.683	0.755
	Relationships	10	0.490	0.577	0.691	0.707	0.736
	Research Ref.	4	0.846	0.849	0.863	0.879	0.885
4	Total	45	0.322	0.490	0.525	0.621	0.788
	Word & Phrases	7	0.707	0.782	0.797	0.840	0.857
	Main Idea	19	0.467	0.521	0.547	0.669	0.811
	Relationships	14	0.396	0.509	0.561	0.618	0.756
	Research Ref.	5	0.661	0.743	0.755	0.759	0.777
5	Total	45	0.251	0.470	0.550	0.626	0.790
	Word & Phrases	8	0.668	0.671	0.746	0.842	0.893
	Main Idea	16	0.346	0.493	0.590	0.644	0.741
	Relationships	16	0.400	0.574	0.619	0.671	0.726
	Research Ref.	5	0.666	0.679	0.694	0.716	0.753
6	Total	45	0.184	0.445	0.560	0.614	0.728
	Word & Phrases	9	0.506	0.580	0.682	0.716	0.763
	Main Idea	19	0.270	0.497	0.588	0.641	0.716
	Relationships	10	0.562	0.623	0.680	0.732	0.770
	Research Ref.	7	0.598	0.626	0.645	0.706	0.736
7	Total	45	0.286	0.458	0.558	0.652	0.789
	Word & Phrases	8	0.531	0.561	0.633	0.692	0.738
	Main Idea	17	0.503	0.529	0.607	0.679	0.768
	Relationships	11	0.521	0.591	0.646	0.727	0.828
	Research Ref.	9	0.451	0.619	0.650	0.807	0.818
8	Total	45	0.281	0.430	0.506	0.571	0.750
	Word & Phrases	6	0.588	0.651	0.667	0.690	0.698
	Main Idea	19	0.354	0.450	0.533	0.614	0.776
	Relationships	13	0.497	0.566	0.600	0.676	0.707
	Research Ref.	7	0.551	0.576	0.626	0.628	0.694
9	Total	45	0.288	0.427	0.552	0.607	0.769
	Word & Phrases	7	0.580	0.711	0.757	0.764	0.835
	Main Idea	18	0.364	0.491	0.594	0.666	0.703
	Relationships	10	0.536	0.588	0.648	0.720	0.746
	Research Ref.	10	0.472	0.497	0.574	0.617	0.646
10	Total	45	0.326	0.397	0.471	0.546	0.690
	Word & Phrases	9	0.531	0.640	0.648	0.691	0.775
	Main Idea	14	0.448	0.517	0.560	0.587	0.603
	Relationships	11	0.466	0.536	0.563	0.633	0.701
	Research Ref.	11	0.414	0.444	0.489	0.530	0.585

Table 55. Summary Data for Biserial Correlations for All Mathematics Items by Reporting Categories

Grade	Reporting Category	No. of Items	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
3	Total	40	0.264	0.443	0.548	0.608	0.727
	Number	12	0.501	0.606	0.659	0.719	0.757
	Measurement	8	0.482	0.610	0.686	0.713	0.774
	Geometry	7	0.598	0.607	0.619	0.633	0.656
	Algebra	6	0.618	0.632	0.690	0.738	0.752
	Data	7	0.503	0.660	0.709	0.750	0.783
4	Total	40	0.233	0.456	0.532	0.577	0.785
	Number	11	0.551	0.632	0.668	0.703	0.814
	Measurement	8	0.562	0.616	0.664	0.671	0.737
	Geometry	7	0.581	0.612	0.658	0.680	0.745
	Algebra	7	0.475	0.596	0.661	0.738	0.761
	Data	7	0.558	0.650	0.663	0.704	0.743
5	Total	50	0.287	0.461	0.576	0.662	0.788
	Number	12	0.505	0.576	0.636	0.755	0.801
	Measurement	11	0.539	0.678	0.743	0.770	0.836
	Geometry	9	0.426	0.520	0.575	0.604	0.660
	Algebra	10	0.554	0.620	0.658	0.693	0.785
	Data	8	0.464	0.497	0.549	0.659	0.685
6	Total	44	0.214	0.388	0.494	0.579	0.726
	Number	9	0.349	0.409	0.472	0.518	0.537
	Measurement	9	0.424	0.619	0.679	0.727	0.772
	Geometry	9	0.485	0.535	0.599	0.607	0.672
	Algebra	8	0.417	0.562	0.620	0.673	0.723
	Data	9	0.517	0.558	0.598	0.617	0.675
7	Total	44	0.213	0.476	0.531	0.590	0.777
	Number	9	0.490	0.618	0.630	0.653	0.709
	Measurement	9	0.566	0.640	0.692	0.714	0.820
	Geometry	8	0.548	0.586	0.613	0.696	0.706
	Algebra	9	0.552	0.650	0.672	0.760	0.791
	Data	9	0.354	0.623	0.660	0.674	0.755
8	Total	50	0.309	0.438	0.534	0.666	0.774
	Number	11	0.519	0.603	0.652	0.713	0.777
	Measurement	11	0.489	0.562	0.656	0.739	0.794
	Geometry	8	0.473	0.564	0.589	0.709	0.745
	Algebra	11	0.467	0.513	0.572	0.713	0.761
	Data	9	0.436	0.536	0.630	0.725	0.799
9	Total	44	0.331	0.426	0.534	0.627	0.874
	Number	8	0.535	0.549	0.605	0.670	0.711
	Measurement	7	0.579	0.663	0.739	0.777	0.953
	Geometry	11	0.477	0.531	0.704	0.799	0.907
	Algebra	10	0.553	0.595	0.621	0.666	0.735
	Data	8	0.620	0.632	0.678	0.743	0.773
10	Total	50	0.309	0.451	0.533	0.644	0.797
	Number	10	0.562	0.630	0.690	0.736	0.801
	Measurement	9	0.587	0.619	0.667	0.719	0.880
	Geometry	10	0.444	0.525	0.619	0.686	0.757
	Algebra	13	0.464	0.508	0.590	0.684	0.746
	Data	8	0.430	0.501	0.577	0.596	0.682

IRT Scaling

IRT Framework

FCAT scoring is built on item response theory (IRT). In essence, IRT assumes that test item-responses by students are the result of underlying achievement levels possessed by those students. IRT algorithms search for “item parameters” which capture a nonlinear relationship between achievement and the likelihood of correctly answering each item. Items that fit the IRT model will exhibit a pattern of lower probabilities of correct responses from low-ability students to higher probabilities of correct responses from high-ability students. This is reflected in an “item characteristic curve,” as depicted in Figure 1, for a multiple-choice item. Items differ in their difficulty such that the position of the point of inflection is higher or lower (to the right or to the left) along the achievement scale. For example, the point of inflection of the curve for the sample item in Figure 1 is centered at zero, the mean on the achievement index. An efficient test will be composed of items with test characteristics similar to that depicted, but with varying difficulties that are able to discriminate achievement along the entire scale, which is typically called “theta.” Item characteristic curves also differ in their lower asymptotes (related to how easy it is to get the item correct by guessing) and the gradient of their slopes at the inflection point.

While IRT modeling of performance tasks is conceptually similar, performance tasks require a more complex mathematical treatment. In the end, however, IRT modeling of a performance task captures the expected number of points that students should achieve on that performance task depending on their achievement level. The result is a curve similar to Figure 1 where the Y-axis represents expected points.

The three-parameter logistic (3PL) model (Lord & Novick, 1968) was used to process MC items, and the two-parameter partial credit (2PPC) model (Muraki, 1992) was used to process PT items. Figure 1 depicts an item characteristic curve using the 3PL model. For the PT items, student scores could fall into any of several different score categories (0, 1, or 2 for short-constructed response items and 0, 1, 2, 3, or 4 for extended-constructed response items). The 2PPC model captures probabilities for students receiving any of the possible points, depending on differences in their achievement. *FCAT 2003 Test Construction Specifications* (FDOE, 2002) presents the technical details of these models more fully. Multilog (Thissen, 1991) was used for the IRT analyses.

Gridded items receive a hybrid treatment. Initially, item parameters are computed using a two-parameter logistic model. Then they are converted to the 2PPC for subsequent processing and maintenance in the item data bank.[3]

[3] The 2PL “b” parameter is multiplied by the “a” parameter.

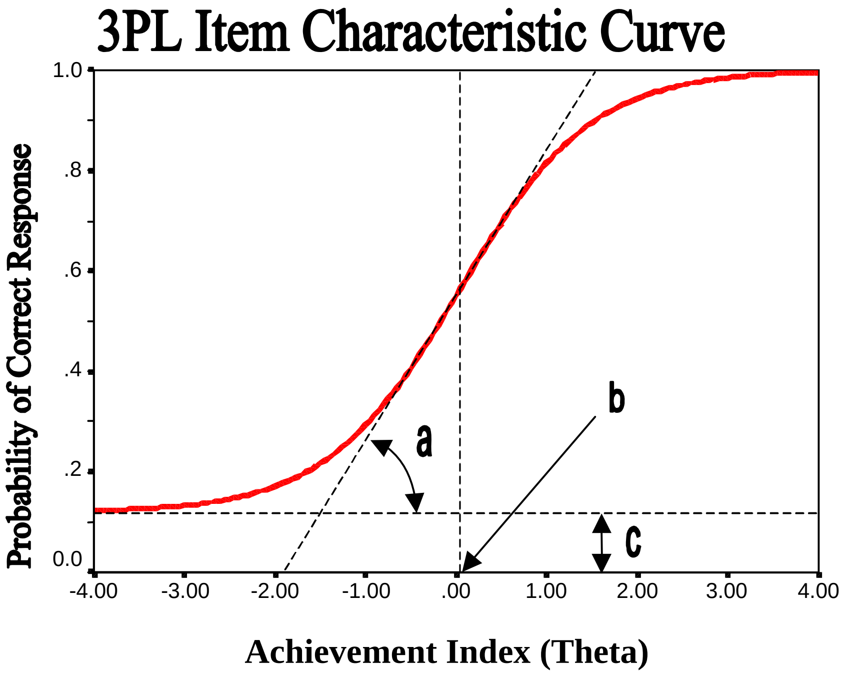

Figure 1. Item Characteristic Curve based on the three-parameter logistic trace line.

IRT item parameters provide the means for assigning achievement scores to individual students. Because the item parameters represent response probabilities, each student's achievement score is assigned as the level of achievement most likely to have created that student's observed responses.[4] Use of the sophisticated IRT model is advantageous for continuous testing programs, such as the FCAT, which must create a stable achievement scoring system given the reality that items included on the tests change from one year to the next.

IRT Results

Distributions of the three 3PL item parameters are presented in Tables 56 and 57 for MC items. The parameters are in the IRT traditional metric,[5] and the achievement scale can be interpreted as a standard scale with a true score mean of 0 and standard deviation of 1. The "A" parameter indicates the slope of the curve. The higher the slope, the more the item contributes to the estimation of achievement scores. "A" is similar to item-total correlation. For reference, the "A" for the sample curve in Figure 1 is 1.0. Items with lower slopes are useful when there are sufficient numbers of items.

[4] That is, scores are calculated using maximum likelihood estimation.

[5] A, B, and C are reported, where $P(\theta) = C + (1-C)/(1+ \exp(-1.7A(\theta-B)))$.

Tables 56 and 57 show that the “A” parameters are centered from 0.66 to 0.87 for reading and about 0.66 to 0.90 for mathematics. The results show that reading “A” parameters are slightly lower than mathematics “A” parameters.

The “B” parameter indicates the difficulty of the items by indicating where the item slope is centered along the achievement scale. “B” is conceptually similar to an item’s *p*-value. For reference, the “B” in Figure 1 is set at 0, indicating that the curve is centered at the population mean. “B” parameters should be spread across a wide range of achievement to accurately measure students at all levels of ability. That is, because of the way the curve flattens on the ends, an item centered in the middle of the achievement scale functions well only for students in the center of the achievement distribution. Items with higher and lower “B” parameters help to measure achievement for students in the upper and lower ends of the achievement distribution. Tables 56 and 57 show that in all cases the “B” parameters are spread across the scale.

Table 56. Multiple-Choice Item Parameter Summary Data—Traditional Metric—All Reading Items

Grade (No. of Items)	Parameter	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	A	0.450	0.690	0.870	1.020	1.570
(45)	B	-1.610	-1.000	-0.450	0.190	1.240
	C	0.050	0.110	0.160	0.230	0.350
4	A	0.420	0.690	0.780	0.960	1.540
(41)	B	-1.710	-0.910	-0.250	0.390	1.890
	C	0.060	0.130	0.190	0.290	0.440
5	A	0.360	0.610	0.770	0.950	1.360
(45)	B	-2.090	-1.290	-0.660	0.220	1.980
	C	0.050	0.120	0.170	0.220	0.540
6	A	0.250	0.540	0.760	0.940	1.350
(45)	B	-2.520	-1.230	-0.530	0.070	2.370
	C	0.070	0.130	0.170	0.230	0.520
7	A	0.240	0.570	0.800	1.010	1.380
(45)	B	-1.610	-0.890	-0.470	-0.010	1.170
	C	0.040	0.130	0.180	0.250	0.460
8	A	0.250	0.530	0.680	0.840	1.250
(41)	B	-3.090	-1.080	-0.480	0.070	1.220
	C	0.060	0.110	0.180	0.280	0.450
9	A	0.330	0.570	0.800	0.910	1.370
(45)	B	-1.900	-.0770	-0.200	0.560	1.860
	C	0.060	0.130	0.200	0.270	0.450
10	A	0.310	0.415	0.660	0.765	1.150
(41)	B	-2.760	-0.925	-0.385	0.075	1.550
	C	0.070	0.120	0.165	0.230	0.430

Table 57. Multiple-Choice Item Parameter Summary Data—Traditional Metric—All Mathematics Items

Grade (No. of Items)	Parameter	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	A	0.210	0.585	0.790	0.905	1.340
(40)	B	-2.660	-0.785	0.145	0.630	1.460
	C	0.030	0.075	0.150	0.220	0.460
4	A	0.340	0.605	0.745	1.000	1.540
(40)	B	-2.480	-0.615	-0.035	0.555	1.820
	C	0.030	0.125	0.170	0.230	0.400
5	A	0.440	0.660	0.900	1.110	1.550
(33)	B	-2.240	-0.530	-0.010	0.480	1.750
	C	0.060	0.110	0.170	0.240	0.460
6	A	0.230	0.530	0.660	0.890	1.620
(33)	B	-2.270	-0.210	0.320	1.010	2.040
	C	0.080	0.140	0.220	0.250	0.550
7	A	0.200	0.660	0.800	1.080	1.550
(33)	B	-6.100	-0.380	0.380	0.860	1.810
	C	0.040	0.140	0.200	0.300	0.430
8	A	0.370	0.590	0.855	1.080	1.730
(30)	B	-2.090	-0.490	0.355	0.750	1.520
	C	0.030	0.160	0.215	0.320	0.470
9	A	0.330	0.730	0.810	1.130	2.100
(29)	B	-2.590	-0.190	0.540	1.290	1.950
	C	0.070	0.120	0.190	0.250	0.350
10	A	0.310	0.540	0.770	1.140	1.500
(28)	B	-2.020	-0.565	0.010	0.485	1.010
	C	0.080	0.115	0.180	0.265	0.590

The 3PL "C" parameter factors in the effects of examinees not knowing the answer and still getting the item correct. This is also called the "pseudo-guessing" parameter. Notice in Figure 1 that the curve asymptotes at a lower value of about 0.2. For MC items with four possible responses, without knowing anything about the item content, the chances of responding correctly are at that lower bound value. Typically, "C" values should be around 0.2. Higher values may signal poorly functioning distractors. Tables 56 and 57 show that the "C" parameters tend to fall in the expected range, but that there are also a few items with high "C" parameters.

The item parameters for the 2PPC model used to score GR and PT items are conceptually more difficult to translate graphically. Therefore, Table 58 presents only distributions of "A" parameters for these items. The "A" parameters for GR and PT items tend to be higher than those for MC items. Algebraically, we should be able to make a direct comparison. Because IRT processing is trying to fit the same achievement construct to all items, this is evidence of the convergence or similarity between the knowledge and skills required for the different item types. (Note that there are only two ER items in any one mathematics test, and they are indicated as the minimum and maximum values. For reading, the single ER item is indicated as the median value.)

Table 58. "A" Parameter Summary Data—Gridded Items and Performance Tasks

Grade	Item Type (No. of Items)	Minimum	25th Percentile	Median	75th Percentile	Maximum
Reading						
4	SR (3)	0.760		0.890		1.190
	ER (1)			0.830		
8	SR (3)	1.100		1.210		1.240
	ER (1)			0.860		
10	SR (4)	0.840	0.910	1.020	1.075	1.090
	ER (1)			0.640		
Mathematics						
5	GR (11)	0.960	1.130	1.450	1.765	2.250
	SR (4)	0.620	0.695	0.775	0.885	0.990
	ER (2)	0.450				0.560
6	GR (11)	0.300	0.960	1.150	1.230	1.780
7	GR (11)	1.010	1.080	1.380	1.545	2.170
8	GR (14)	0.860	1.260	1.420	1.610	1.950
	SR (4)	1.080	1.110	1.205	1.370	1.470
	ER (2)	0.940				1.240
9	GR (15)	0.680	1.030	1.610	1.865	2.550
10	GR (16)	0.640	0.985	1.440	1.605	2.190
	SR (4)	0.800	0.855	1.085	1.390	1.520
	ER (2)	0.800				1.050

Scale Conversion and Test Equating

IRT scaling produces item parameters for an achievement scale targeted to a true score mean of 0 and true score standard deviation of 1. The FCAT, however, reports scores on a scale that runs from 100 to 500. Therefore, a transformation is needed for the IRT item parameters in order for them to produce the appropriate scores. Figure 2 shows a sample item characteristic curve after conversion to the associated 100-500 scale.

In addition to the need for student scores to be placed on an appropriate scale, there is also the need for those scores to be comparable to scores from past years. Students from 2003 are expected to perform differently (presumably better) than students in previous years. To report scores in 2003 on the 100–500 FCAT scale and make those scores comparable to scores from past years, the data output by IRT processing needed to be altered by an equating process. This process involves (1) repeating the 2003 test "anchor items" that had been used in previous FCAT administrations, and (2) applying the Stocking/Lord (1983) procedure using those anchor items to adjust for the difference between students in 2002 and students in 1998 (2001 for tests that became operational that year). The anchor items and the Stocking/Lord procedure are used to equate 2003 test scores to the test scores originally reported in 1998 (or 2001). The procedure, with different anchor items, has been conducted every year since 1998 (or 2001).

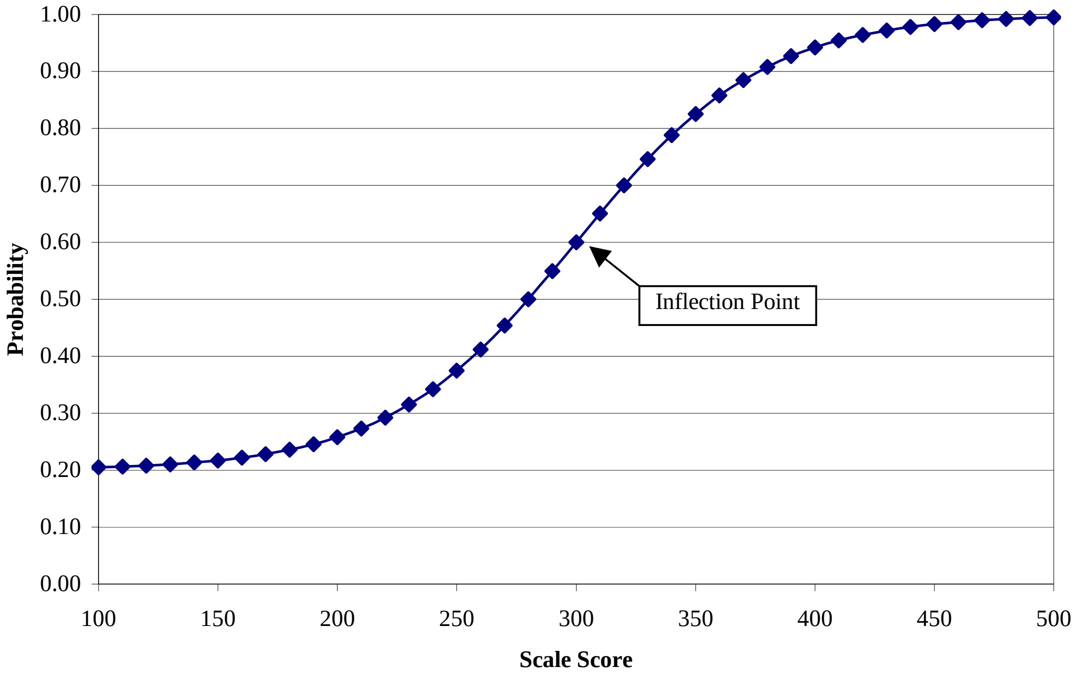

Figure 2. Sample item characteristic curve after conversion to the 100-500 scale (3PL IRT model with A=0.015, B=300, and C=0.2).

With the completion of the 2003 scaling, the anchor items have two sets of item parameters: new parameters on the mean = 0, standard deviation = 1 scale produced parameters of the current year and the old parameters that were transformed during their previous use. The old parameters are on the original 1998 (or 2001) scale. The Stocking/Lord procedure uses the old item parameters to locate the achievement scale and then searches for a transformation multiplier and additive constant that can combine to make the new parameters replicate the 1998 (or 2001) achievement scale as closely as possible. This is done by attempting to match test characteristic curves (which are summations of item characteristic curves, such as in Figure 1 on page 32) produced by the old parameters with test characteristic curves formed by transformations of new parameters. Since the items are the same, the same scale is expected to result.

Appendix C documents the item-level reviews that were conducted during the equating process. Specifically, items with questionable parameter estimates (low, high, or at variance with their prior parameter estimates) were reviewed for use in the equating process. In several instances, intended linking items were dropped from the equating process. This year, Item 27 from Grade 5 Mathematics, Item 13 from Grade 8 Mathematics, and Item 11 from Grade 10 Mathematics were dropped from equating. In addition to HumRRO and Harcourt Educational Measurement, NCS/Pearson and the Florida Department of Education also participated in these reviews. In previous years, this procedure was conducted by examining each set of corresponding item parameters separately. Last year, HumRRO introduced a computational procedure that produces a metric indicating the difference between the shapes of the item characteristic curves produced by the current versus base-year item parameters. This metric takes all item parameters into account. The items with the largest differences were identified for further review and possible elimination from equating. A more complete description of this procedure, as well as a list of items eliminated from equating, is presented in Appendix C.

Table 59 indicates the number of anchor parameters used in equating and the transformation constants that were derived to replicate the base-year FCAT scale. The M2 additive constant projects the change in average true score achievement level expected for standard curriculum students. Thus, while an average standard curriculum student would be expected to have a score of 300 for Grade 4 Reading in 1998, the average standard curriculum student in 2003 would be expected to have a score of approximately 315, the value of M2 for Grade 4 Reading.

Table 59. Equating Multiplicative and Additive Constants

Grade	Anchor Item Type and Number	M1 Multiplier	M2 Additive Constant
Reading			
3	15 MC	48.428	303.864
4	16 MC, 1 SR	44.581	314.796
5	14 MC	45.063	299.689
6	14 MC	46.740	306.288
7	13 MC	46.846	307.191
8	12 MC, 1 SR	41.791	313.904
9	14 MC	46.143	295.662
10	12 MC, 1 SR	42.706	307.812
Mathematics			
3	15 MC	52.381	313.920
4	13 MC	46.204	307.651
5	9 MC, 4 GR	42.662	330.857
6	11 MC, 4 GR	44.216	319.242
7	11 MC, 3 GR	44.892	308.670
8	9 MC, 4 GR	37.920	325.128
9	8 MC, 7 GR	39.860	301.336
10	10 MC, 8 GR	34.893	324.297

Note: For computation of mathematics results in Grades 5, 8, and 10, one short response (SR) item was not included in scoring, scaling, and equating. The items were removed by FDOE for content reasons.

[a] Anchor item 27 was dropped.

[b] Anchor item 13 was dropped.

[C] Anchor item 11 was dropped, and four additional items were added.

IRT Fit Statistics

Again, IRT scaling algorithms attempt to find item parameters (numerical characteristics) that create a match between observed patterns of item responses and theoretical response patterns defined by the selected IRT models. The Q1 statistic (Yen, 1981) may be used as an index for how well theoretical item curves are found which match observed item responses. Q1 is computed by first conducting an IRT item parameter estimation, then estimating students' achievement using the estimated item parameters, and – finally – by using students' achievement scores in combination with estimated item parameters to compute expected performance on each item. Differences between expected item performance and observed item performance are then compared at selected intervals across the range of student achievement. Q1 is computed as a ratio involving expected and observed item performance and is, therefore, interpretable as a chi-square statistic.

Because the different types of items have different numbers of IRT parameters, Q1 for each item type has a different number of degrees of freedom. Therefore, Q1 is not directly comparable across item types. An adjustment (translation to a z-score, or ZQ) is made for different numbers of item parameters and sample sizes to create a more general statistic. The FCAT has a set of standards for a minimum ZQ for an item to be labeled as having "acceptable" versus "poor" fit (FDOE, 1998).[6] Complete Q1 results are in the Appendices. Tables 60 and 61 present the distributions of ZQs and Table 62 presents the numbers of poorly fitting items by item type. The low proportion of poorly fitting items is consistent with the previously reported patterns of strong point-biserials and strong "A" parameters. The exception is perhaps Grade 3 Reading. Table 60 shows higher ZQ values for Grade 3 than for the other grades. This may have resulted from the fact that three reading items were rated "poor" fit (see Table 62). The only other grade to have a poor fitting item was Grade 6 (only one), while the remaining grades had none.

Table 60. Z Transformation of Q1 Statistic, Summary Data—All Reading Items

Grade	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	1.977	3.566	5.711	8.478	14.760
4	-1.300	-0.065	0.842	1.966	5.578
5	-.0918	0.291	0.869	1.521	6.634
6	-1.050	0.103	1.023	2.101	14.656
7	-1.097	-0.017	0.137	1.033	7.606
8	-1.568	-0.467	0.381	2.453	10.224
9	-1.101	-0.404	0.902	2.418	14.379
10	-1.330	-0.026	0.873	2.667	9.027

[6] If ZQ > (n*4/1500) where n=sample size, then fit is rated as "poor."

Table 61. Z Transformation of Q1 Statistic, Summary Data—All Mathematics Items

Grade	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	-0.808	0.099	1.379	2.387	6.495
4	-1.404	-0.312	0.586	1.401	3.352
5	-1.378	-0.167	1.023	2.490	14.707
6	-1.479	-0.021	0.760	2.408	6.566
7	-0.961	0.208	1.202	3.593	9.547
8	-1.269	-0.013	1.103	3.046	30.620
9	-0.711	-0.005	1.008	2.439	6.360
10	-1.235	-0.225	1.001	2.373	8.132

Table 62. Number of Poorly Fitting Items According to Q1 Statistics—All Items

Grade	Reading			Mathematics			
	MC	SR	ER	MC	GR	SR	ER
3	3/45			0/40			
4	0/41	0/3	0/1	0/40			
5	0/45			0/33	0/11	1/4	0/2
6	1/45			0/33	0/11		
7	0/45			0/33	0/11		
8	0/41	0/3	0/1	0/30	1/14	0/4	1/2
9	0/45			0/29	0/15		
10	0/41	0/3	0/1	0/28	0/16	0/4	0/2

Note: Numbers shown are – Number of items with “poor fit”/Total number of items

Achievement Scale Unidimensionality

By fitting all items simultaneously to the same achievement scale, IRT is operating under the assumption that there is a strong, single construct that underlies the performance of all items. Under this assumption, performance on the items should be related to achievement (as depicted by Figure 1), and additionally, any relationship of performance between pairs of items should be “explained” or “accounted for” by variance in student levels of achievement. This is the “local dependence” assumption of unidimensional IRT and suggests a relatively straightforward test for unidimensionality, called the Q3 statistic (Yen, 1984).

Computation of the Q3 statistic begins in the same manner as the Q1 statistic: expected student performance on each item is calculated using item parameters and estimated achievement scores. Then, for each student and each item, the difference between expected and observed item performance is calculated. The difference can be thought of as the residual in performance after accounting for underlying achievement. If performance on the items is driven by a single achievement construct, then not only will the residuals be small (as tested by the Q1 statistic), but correlations between residuals of the pairs of items will also be small. These correlations are analogous to partial correlations, which can be interpreted as the relationship between two variables (items) after the effects of a third variable (underlying achievement) are held constant or “accounted for.” The correlation among IRT residuals is the Q3 statistic.

With *n* items, there are *n*(*n*-1)/2 Q3 statistics. For example, for Grade 3 Reading with 45 items, there are 990 Q3 values. All Q3 values should be small. To summarize Q3 data, Tables 63 and 64 present the minimum, 5th percentile, median, 95th percentile, and maximum values for each FCAT grade/subject combination. To add perspective to the meaning of the Q3 distributions, the average zero-order correlations among item responses are also indicated. If the achievement construct is “accounting for” the relationships among the items, Q3 values should be much smaller than the zero-order correlations. These tables indicate that, for all grades/subjects, at least 90 percent of the items have Q3 values that are expectedly small, showing Q3 values between -.07 and .03. These data, coupled with the Q1 data above, indicate that the unidimensional IRT model provides a very reasonable solution for capturing the essence of student achievement defined by the carefully selected set of items for each grade and subject.

Table 63. **Q3 Statistic, Summary Data—All Reading Items**

		Q3 Distribution				
Grade	Average Correlation	Minimum	5th Percentile	Median	95th Percentile	Maximum
3	0.168	-0.137	-0.074	-0.018	0.032	0.250
4	0.154	-0.111	-0.058	-0.021	0.022	0.125
5	0.137	-0.107	-0.060	-0.020	0.016	0.097
6	0.134	-0.099	-0.057	-0.020	0.021	0.110
7	0.151	-0.116	-0.060	-0.020	0.016	0.118
8	0.130	-0.113	-0.061	-0.020	0.022	0.110
9	0.134	-0.092	-0.055	-0.020	0.015	0.130
10	0.123	-0.115	-0.058	-0.018	0.014	0.174

Table 64. **Q3 Statistic, Summary Data—All Mathematics Items**

		Q3 Distribution				
Grade	Average Correlation	Minimum	5th Percentile	Median	95th Percentile	Maximum
3	0.142	-0.097	-0.065	-0.022	0.018	0.163
4	0.141	-0.096	-0.059	-0.022	0.012	0.187
5	0.178	-0.105	-0.061	-0.019	0.024	0.290
6	0.114	-0.100	-0.056	-0.019	0.018	0.073
7	0.141	-0.106	-0.057	-0.019	0.024	0.093
8	0.186	-0.124	-0.061	-0.016	0.023	0.155
9	0.152	-0.089	-0.059	-0.019	0.021	0.160
10	0.180	-0.099	-0.056	-0.016	0.017	0.100

Item Bias Analyses

FCAT test items receive intensive, qualitative reviews by expert panels before being placed into field tests, including review for possible gender or ethnicity bias (FDOE, 2002, May). In addition, items are examined after each use for quantitative evidence of differential performance by various subgroups of examinees, representing gender/racial/ethnic groups, whose achievement levels are assumed to be comparable. The differential item functioning (DIF) analyses are conducted for gender (Males vs. Females) and ethnicity (Caucasians vs. African Americans and Caucasians vs. Hispanics.)

Analyses of DIF were done using two methods that are described by Zwick, Donoghue, and Grima (1993). Both methods compare performance on each item with performance on the test as a whole. For any given achievement level, as defined by the FCAT scale score, performance on each item should be the same for females and males. Similarly, at any given level of overall achievement, performance on each item should be similar for African Americans or Hispanics when compared with the Caucasian population. The Mantel (1963) statistic (a version of the common Mantel-Haenszel (1959) statistic that accommodates performance task items) is a chi-square statistic that tests the statistical significance (or probability) of differences in item performance. Standardized mean difference (SMD) looks at the size of the difference and is particularly useful because with large sample sizes, such as those found in the FCAT calibration samples, a statistically significant difference may appear for a comparison done on groups responding to an item; however, that difference – on examination by educators and policymakers – may not be deemed large enough to cause concern from a practical testing and decision-making perspective. An SMD rating system, which was put into place (FDOE, 1998), groups each item into one of seven categories according to its demonstrated differential functioning for or against any of the identified comparison groups. Complete Mantel-Haenszel and SMD results are presented in Appendices A and B. Tables 65 and 66 present the distribution of SMD summary ratings. Given the review through which these items had already passed, including field-test use in previous years, the low incidence of large DIF ratings is not surprising.

Table 65. Item DIF Rating Summary—All Reading Items

	Overall Standardized Mean Difference Rating						
Grade	**1 – Low DIF**	**2**	**3**	**4**	**5**	**6**	**7 – High DIF**
3	44		1				
4	42	3					
5	42	3					
6	41	4					
7	43	1	1				
8	38	5		2			
9	43	1	1				
10	41	3	1				

Table 66. Item DIF Rating Summary—All Mathematics Items

	Overall Standardized Mean Difference Rating						
Grade	**1 – Low DIF**	**2**	**3**	**4**	**5**	**6**	**7 – High DIF**
3	38	2					
4	37	3					
5	43	5	1		1		
6	41	3					
7	44						
8	43	4	3				
9	43	1					
10	45	3	2				

Test Reliability and Standard Error of Measurement

The previous discussion pointed to FCAT test items for each test converging on a common achievement scale. Two additional views of this convergence – conditional standard errors of measurement and reliability – are presented in this section.

Test reliability concerns the concept that a test score results from some true level of achievement plus measurement error. For a population of students, reliability is a ratio of variation in true achievement compared with variation in observed test scores. The less measurement error contaminates test scores, the closer the ratio is to 1. Under classical test theory, measurement error is assumed to be the same at all levels of achievement, and one reliability coefficient can be estimated to acknowledge that error. Within the IRT framework, however, measurement error is not assumed to be constant across the range of ability. Rather, measurement error, that is, the standard error of measurement (SEM), is a function of how well a student's pattern of item responses matches the expected response pattern uncovered by the IRT modeling processes. In other words, with IRT modeling, score assignment is more accurate for a student who correctly answers the easy items and misses the difficult items than for a student who gets as many easy items correct as difficult items. Furthermore, score assignment tends to be more accurate for students toward the center of the distribution than for students with more extreme scores.

Conditional standard error curves, depicted in Figures 3 and 4 (Reading and Mathematics, respectively) on the following pages, are one method for depicting test reliability. The curves plot the average SEM extracted from student score records as a function of achievement level. SEM is like a standard deviation so that approximately two-thirds of the students with a given level of achievement will have observed test scores within 1 SEM of the given true score. For example, in Figure 3, the Grade 3 Reading SEM plots show that students whose true achievement level is 200 will have a SEM of approximately 25. That means that approximately two-thirds of those students will have test scores between 175 and 225. The remaining one-third of the students with a true achievement level of 200 will have test scores more than 25 points away from 200. As expected, SEM is larger at the tails of the achievement level distribution and smaller in the center. Most students, however, are in the center of the distribution. Cut points, used to determine student performance categories (1-5), are located in the center of the distribution as well (see Tables 67 and 68).

It is possible to synthesize an overall reliability system from the standard error curves by using the average SEM for all students to compute a “marginal” reliability. These values, which can be interpreted like traditional reliability statistics, such as Cronbach’s alpha, are presented in Table 69.

While marginal reliability estimates were computed using only the calibration sample, it is important to note that the SEM curves and reliability estimates were computed using all students who received scores, including the non-standard curriculum students. This was done in order to make reliability data consistent across grades and subjects and not confounded by any differences in calibration samples. In addition, these estimates are consistent with the application of the FCAT: they characterize test results for all students who receive scores.

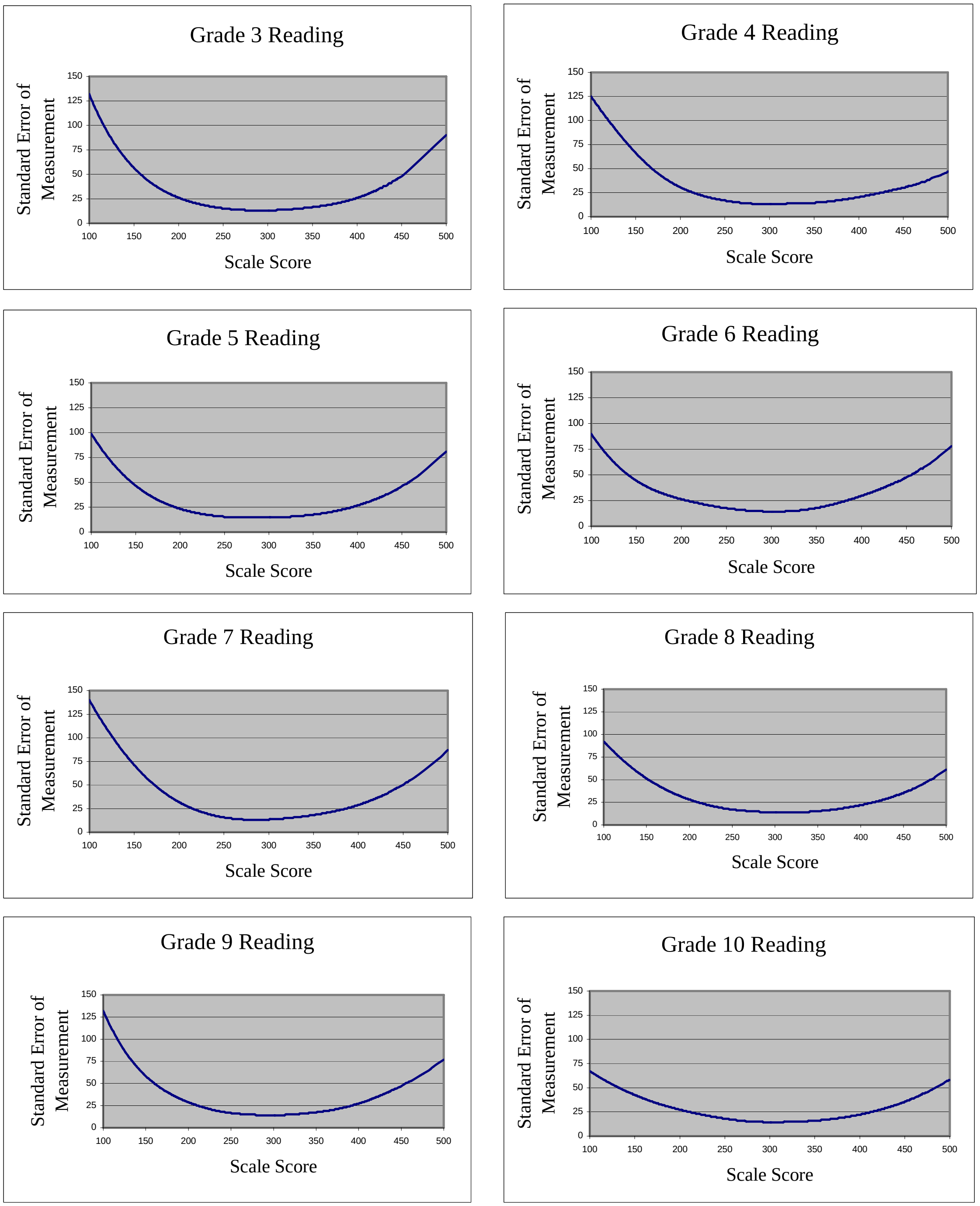

Figure 3. Standard error of measurement plots for 2003 FCAT Reading, by grade.

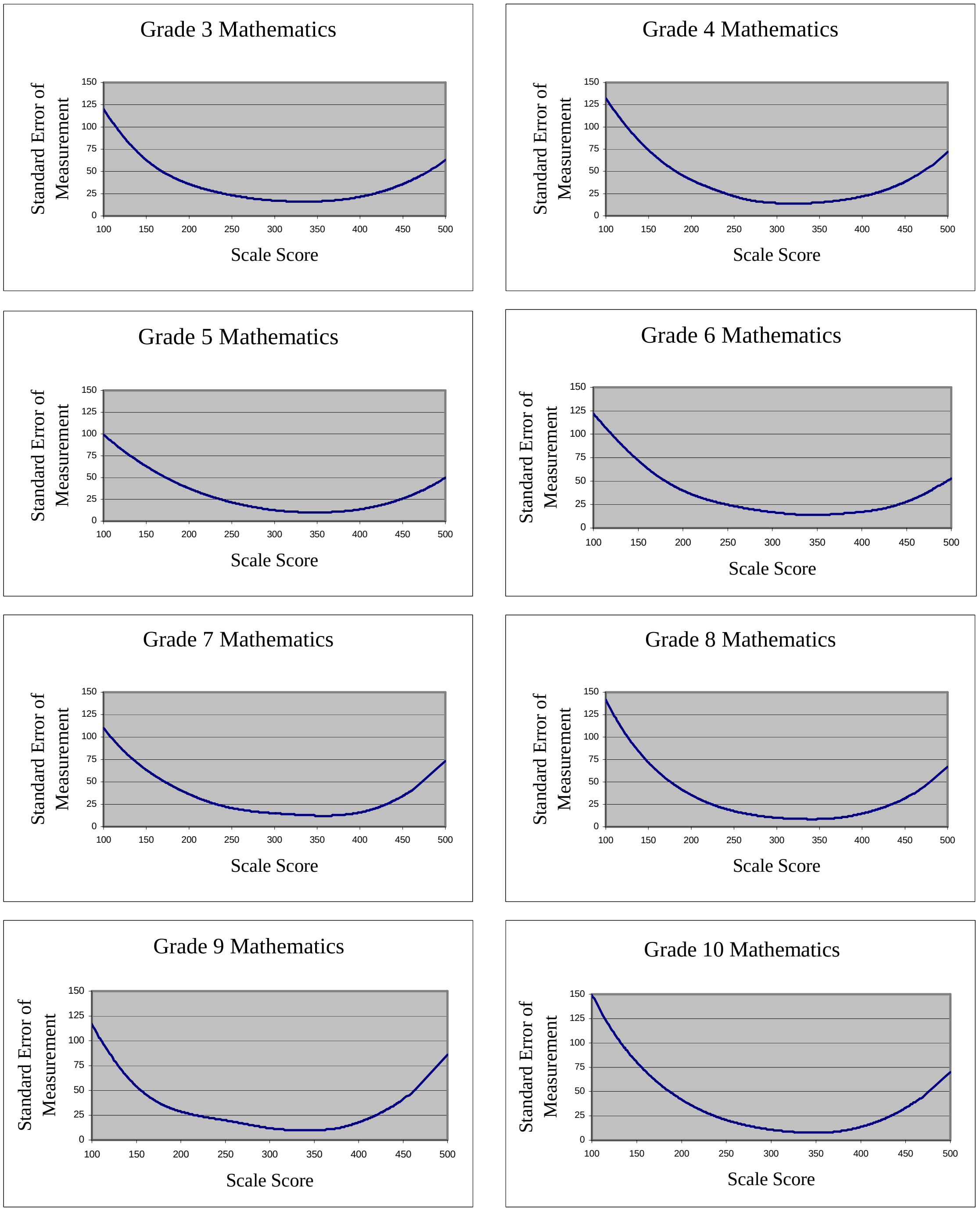

Figure 4. Standard error of measurement plots for 2003 FCAT Mathematics, by grade.

Table 67. Reading SEM at Cut points for Achievement Levels 1-5 (Scores at or above cut points are in higher category).

Grade	Cut points	SEM
3	259	14
	284	13
	332	15
	394	24
4	275	14
	299	13
	339	14
	386	18
5	256	15
	286	15
	331	16
	384	23
6	265	16
	296	14
	339	16
	387	26
7	267	14
	300	13
	344	17
	389	26
8	271	15
	310	14
	350	15
	394	21
9	285	14
	322	15
	354	18
	382	23
10	287	15
	327	15
	355	16
	372	18
PASS (10 only)	300	14

Table 68. Mathematics SEM at Cut points for Achievement Levels 1-5 (Scores at or above cut points are in higher category).

Grade	Cut points	SEM
3	253	23
	294	18
	346	16
	398	21
4	260	19
	298	15
	347	15
	394	21
5	288	14
	326	11
	355	10
	395	13
6	283	19
	315	15
	354	14
	391	16
7	275	17
	306	15
	344	13
	379	13
8	280	12
	310	9
	347	8
	371	10
9	261	18
	296	12
	332	10
	367	11
10	287	13
	315	9
	340	8
	375	9
PASS (10 only)	300	11

Viewing both the reliability and SEM data is important. The marginal reliabilities indicate that FCAT scores have reliabilities similar to those of other standardized and statewide tests. The SEM curves indicate that individuals near the center of the distribution will have test scores that vary by chance by less than 20 points (that is, plus or minus the lowest SEM). Individual test scores will vary more toward the upper and lower portions of the distribution. Rogosa (1994, 2000) explored the implication of failing to note both reliability and SEM estimates when interpreting test data for programs such as the FCAT. While reliabilities around 0.90 are typically viewed positively, test scores can fluctuate randomly, as noted by SEM.

Table 69 also shows traditional Cronbach's alpha reliability statistics. These estimates are based on raw scores only and have been calculated for the total set of items and for the items that comprise each of the separate reporting categories. Lower reliabilities for the reporting categories reflect the reality that fewer numbers of items are associated with each of these subtests. The numbers of items are in parentheses.

Table 69. IRT Marginal Reliabilities and Cronbach's Alpha

Cronbach's Alpha							
Reading	***IRT Marginal r_{ii}***	***Total***	***Word and Phrases***	***Main idea***	***Recognizing Relationships***	***Research Reference***	
Grade 3	0.91	.912	.718 (11)	.820 (20)	.689 (10)	.546 (4)	
4	0.91	.904	.681 (7)	.808 (19)	.727 (14)	.476 (5)	
5	0.90	.897	.692 (8)	.742 (16)	.769 (16)	.376 (5)	
6	0.90	.894	.598 (9)	.776 (19)	.695 (10)	.563 (7)	
7	0.91	.905	.539 (8)	.793 (17)	.717 (11)	.695 (9)	
8	0.90	.894	.488 (6)	.745 (19)	.757 (13)	.614 (7)	
9	0.89	.885	.616 (7)	.756 (18)	.683 (10)	.541 (10)	
10	0.88	.882	.626 (9)	.656 (14)	.678 (11)	.649 (11)	
Mathematics	***IRT Marginal r_{ii}***	***Total***	***Number Sense, Concepts, Operations***	***Measure-ment***	***Geometry and Spatial Sense***	***Algebraic Thinking***	***Data Analysis/ Probability***
Grade 3	0.88	.881	.744 (12)	.598 (8)	.479 (7)	.489 (6)	.605 (7)
4	0.88	.880	.738 (11)	.607 (8)	.489 (7)	.538 (7)	.556 (7)
5	0.93	.919	.746 (12)	.790 (11)	.625 (9)	.729 (10)	.596 (8)
6	0.87	.866	.554 (9)	.672 (9)	.548 (9)	.541 (8)	.508 (9)
7	0.89	.888	.623 (9)	.623 (9)	.542 (8)	.672 (9)	.636 (9)
8	0.93	.929	.740 (11)	.762 (11)	.682 (8)	.717 (11)	.692 (9)
9	0.90	.894	.548 (8)	.576 (7)	.740 (11)	.658 (10)	.614 (8)
10	0.92	.920	.721 (10)	.704 (9)	.713 (10)	.742 (13)	.586 (8)

Intercorrelations among Reporting Categories and Scale Scores

Tables 70 through 85 present intercorrelations among IRT derived scale scores, total raw scores, and the FCAT reporting categories. As expected, correlations between total raw scores and IRT scale scores are high (0.92 to 0.98). Comparisons of the correlations among reporting category scales themselves are affected by differences in scale reliabilities (see Table 69) that result from differences in numbers of items in the categories. For example, in Table 70 observed correlations with the Research and Reference reporting category would be expected to be lower than the other correlations because Research and Reference is measured with only three items for

Grade 3. This means that all the correlations among the reporting categories are underestimated due to lower reliabilities of corresponding subscores.

Tables for Reading

Table 70. Grade 3 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N = 4,683

	Total Raw Score (45)	Word & Phrases (11)	Main Ideas (20)	Relation-ship (10)	Research Ref. (4)
Scale Score	0.970	0.838	0.920	0.826	0.685
Total Raw Score		0.866	0.942	0.864	0.697
Word & Phrases			0.728	0.668	0.555
Main Ideas				0.738	0.594
Relationships					0.524

Table 71. Grade 4 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N= 4,650

	Total Raw Score (45)	Word & Phrases (7)	Main Ideas (19)	Relation-ship (14)	Research Ref. (5)
Scale Score	0.972	0.770	0.910	0.869	0.659
Total Raw Score		0.772	0.930	0.906	0.693
Word & Phrases			0.653	0.613	0.472
Main Ideas				0.747	0.550
Relationships					0.565

Table 72. Grade 5 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N=4,436

	Total Raw Score (45)	Word & Phrases (8)	Main Ideas (16)	Relation-ship (16)	Research Ref. (5)
Scale Score	0.975	0.772	0.868	0.893	0.615
Total Raw Score		0.787	0.897	0.908	0.644
Word & Phrases			0.608	0.630	0.444
Main Ideas				0.712	0.473
Relationships					0.508

Table 73. **Grade 6 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N=4,451**

	Total Raw Score (45)	Word & Phrases (9)	Main Ideas (19)	Relationship (10)	Research Ref. (7)
Scale Score	0.965	0.762	0.879	0.837	0.757
Total Raw Score		0.790	0.916	0.858	0.785
Word & Phrases			0.629	0.582	0.547
Main Ideas				0.696	0.609
Relationships					0.608

Table 74. **Grade 7 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N=4,393**

	Total Raw Score (45)	Word & Phrases (8)	Main Ideas (17)	Relationship (11)	Research Ref. (9)
Scale Score	0.959	0.713	0.884	0.821	0.807
Total Raw Score		0.746	0.915	0.860	0.845
Word & Phrases			0.586	0.553	0.529
Main Ideas				0.693	0.698
Relationships					0.652

Table 75. **Grade 8 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N=4,482**

	Total Raw Score (45)	Word & Phrases (6)	Main Ideas (19)	Relationship (13)	Research Ref. (7)
Scale Score	0.977	0.655	0.854	0.888	0.812
Total Raw Score		0.683	0.888	0.895	0.822
Word & Phrases			0.506	0.554	0.449
Main Ideas				0.688	0.631
Relationships					0.658

Table 76. **Grade 9 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N=5,495**

	Total Raw Score (45)	Word & Phrases (7)	Main Ideas (18)	Relationship (10)	Research Ref. (10)
Scale Score	0.964	0.772	0.881	0.824	0.741
Total Raw Score		0.785	0.913	0.849	0.788
Word & Phrases			0.636	0.593	0.533
Main Ideas				0.683	0.607
Relationships					0.566

Table 77. **Grade 10 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N=4,743**

	Total Raw Score (45)	Word & Phrases (9)	Main Ideas (14)	Relation-ship (11)	Research Ref. (11)
Scale Score	0.977	0.789	0.817	0.845	0.841
Total Raw Score		0.798	0.845	0.857	0.866
Word & Phrases			0.606	0.595	0.589
Main Ideas				0.625	0.612
Relationships					0.656

Tables for Mathematics

Table 78. **Grade 3 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N=4,687**

	Total Raw Score (40)	Number (12)	Measure-ment (8)	Geometry (7)	Algebra (6)	Data (7)
Scale Score	0.969	0.871	0.773	0.666	0.712	0.782
Total Raw Score		0.890	0.782	0.720	0.739	0.804
Number			0.606	0.534	0.605	0.623
Measurement				0.461	0.473	0.560
Geometry					0.412	0.508
Algebra						0.510

Table 79. **Grade 4 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N=4,606**

	Total Raw Score (40)	Number (11)	Measure-ment (8)	Geometry (7)	Algebra (7)	Data (7)
Scale Score	0.958	0.867	0.775	0.672	0.727	0.731
Total Raw Score		0.882	0.812	0.696	0.780	0.776
Number			0.636	0.515	0.615	0.600
Measurement				0.477	0.535	0.537
Geometry					0.440	0.456
Algebra						0.514

Table 80. Grade 5 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N=4,479

	Total Raw Score (50)	Number (12)	Measure-ment (11)	Geometry (9)	Algebra (10)	Data (8)
Scale Score	0.956	0.836	0.857	0.792	0.831	0.745
Total Raw Score		0.867	0.886	0.830	0.859	0.796
Number			0.730	0.633	0.692	0.591
Measurement				0.664	0.716	0.632
Geometry					0.642	0.580
Algebra						0.600

Table 81. Grade 6 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N=4,444

	Total Raw Score (44)	Number (9)	Measure-ment (9)	Geometry (9)	Algebra (8)	Data (9)
Scale Score	0.951	0.750	0.818	0.727	0.725	0.708
Total Raw Score		0.804	0.844	0.752	0.764	0.760
Number			0.589	0.486	0.539	0.525
Measurement				0.566	0.548	0.550
Geometry					0.459	0.454
Algebra						0.491

Table 82. Grade 7 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N=4,392

	Total Raw Score (44)	Number (9)	Measure-ment (9)	Geometry (8)	Algebra (9)	Data (9)
Scale Score	0.959	0.774	0.791	0.702	0.827	0.800
Total Raw Score		0.824	0.834	0.757	0.816	0.835
Number			0.605	0.524	0.583	0.608
Measurement				0.553	0.597	0.632
Geometry					0.526	0.534
Algebra						0.604

Table 83. **Grade 8 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N=4,484**

	Total Raw Score (50)	Number (11)	Measure-ment (11)	Geometry (8)	Algebra (11)	Data (9)
Scale Score	0.963	0.844	0.863	0.833	0.838	0.829
Total Raw Score		0.880	0.893	0.867	0.868	0.860
Number			0.730	0.693	0.713	0.706
Measurement				0.723	0.724	0.708
Geometry					0.685	0.676
Algebra						0.683

Table 84. **Grade 9 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N=4,527**

	Total Raw Score (44)	Number (8)	Measure-ment (7)	Geometry (11)	Algebra (10)	Data (8)
Scale Score	0.947	0.726	0.709	0.844	0.806	0.769
Total Raw Score		0.793	0.787	0.886	0.839	0.773
Number			0.560	0.614	0.576	0.523
Measurement				0.648	0.568	0.505
Geometry					0.657	0.610
Algebra						0.580

Table 85. **Grade 10 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parenthesis) N=4,630**

	Total Raw Score (50)	Number (10)	Measure-ment (9)	Geometry (10)	Algebra (13)	Data (8)
Scale Score	0.935	0.827	0.772	0.839	0.824	0.776
Total Raw Score		0.871	0.849	0.895	0.881	0.823
Number			0.696	0.715	0.716	0.658
Measurement				0.727	0.682	0.603
Geometry					0.714	0.661
Algebra						0.666

Student Classification Accuracy and Consistency

Based on their FCAT scale scores, students are classified into one of five performance levels. While it is always important to know the reliability of student scores in any examination, the ability to assess the reliability of the decisions based on these scores is of even greater importance. Evaluation of the reliability of classification decisions is performed through estimation of the probabilities of correct and consistent classification of students. Procedures were used from Livingston and Lewis (1995), and Lee, Hanson, and Brennan (2000) to derive measures of the accuracy and consistency of the classifications. A brief description of the procedures that were used and the results derived from them are presented in this section.

Accuracy of Classification

According to Livingston and Lewis (1995, p. 180), the accuracy of a classification is ". . . the extent to which the actual classifications of the test takers . . . agree with those that would be made on the basis of their true score, if their true scores could somehow be known." Accuracy estimates are calculated from cross-tabulations between "classifications based on an observable variable (scores on . . . a test) and classifications based on an unobservable variable (the test takers's true scores)." True score is also referred to as a hypothetical mean of scores from all possible forms of the test if they could be somehow obtained (Young and Yoon, 1998). Since these true scores are not available, Livingston and Lewis provide a method to estimate the true score distribution of a test and create the cross-tabulation of the true score and observed score classifications. The example of the 5x5 cross-tabulation of the true score vs. observed score classifications for the FCAT Grade 3 Mathematics is given in Table 86. This example is provided to aid in interpreting the overall indices of accuracy found in Table 89. It shows the proportions of students who were classified into each performance category by the actual observed scores and by estimated true scores.

Table 86. **FCAT 2003 Grade 3 Mathematics True Scores vs. Observed Scores Cross-Tabulation (Accuracy Table)**

True Score	Observed Score					
	LEVEL 1	**LEVEL 2**	**LEVEL 3**	**LEVEL 4**	**LEVEL 5**	**Total**
LEVEL 1	0.18	0.03	0.00	0.00	0.00	0.21
LEVEL 2	0.04	0.10	0.04	0.00	0.00	0.17
LEVEL 3	0.00	0.05	0.19	0.05	0.00	0.29
LEVEL 4	0.00	0.00	0.06	0.19	0.06	0.31
LEVEL 5	0.00	0.00	0.00	0.01	0.02	0.03
Total	0.22	0.17	0.29	0.25	0.08	1.00

Note: Columns and row totals are computed from non-rounded values. Shaded cells are used for computing overall accuracy index (explained in further sections).

Consistency of Classification

Consistency is “. . . the agreement between classifications based on two non-overlapping, equally difficult forms of the test” (Livingston and Lewis, 1995, p. 180). Consistency is estimated using actual response data from a test and the test’s reliability in order to statistically model two parallel forms of the test and compare the classifications on those alternate forms. The example of 5x5 cross-tabulation between a form taken and an alternate form for the FCAT Grade 3 Mathematics is given in Table 87. This example is provided to aid in interpreting the overall indices of consistency found in Table 89. The table shows the proportions of students who were classified into each performance category by the actual test and by another (hypothetical) parallel test form.

Note that the consistency table is symmetrical, i.e., the same values are observed for Level 1 – Level 2 or Level 2 – Level 1 because the comparisons are based on the same scores. However, the accuracy table is non-symmetrical because it compares classifications based on two different types of scores. Also note that agreement rates are lower in the consistency table because both classifications contain measurement errors, whereas in the accuracy table true score classification is assumed to be errorless.

Table 87. FCAT 2003 Grade 3 Mathematics True Scores vs. Observed Scores Cross-tabulation (Consistency Table)

Form Taken	Alternate Form					
	LEVEL 1	**LEVEL 2**	**LEVEL 3**	**LEVEL 4**	**LEVEL 5**	**Total**
LEVEL 1	0.17	0.04	0.01	0.00	0.00	0.22
LEVEL 2	0.04	0.08	0.05	0.00	0.00	0.17
LEVEL 3	0.01	0.05	0.15	0.07	0.01	0.29
LEVEL 4	0.00	0.00	0.07	0.13	0.04	0.25
LEVEL 5	0.00	0.00	0.01	0.04	0.03	0.08
Total	0.22	0.17	0.29	0.25	0.08	1.00

Note: Columns and row totals are computed from non-rounded values. Shaded cells are used for computing consistency index conditional on level (explained in further Phrases).

Accuracy and Consistency Indices

There are three types of accuracy and consistency indices that can be generated from the examples in Tables 86 and 87: *overall, conditional on level*, and *by cut point*. In order to facilitate their interpretation, a brief outline of computational procedures used to derive accuracy indices will be presented using the example of the FCAT Grade 3 Mathematics test.

The *overall accuracy* of performance level classifications is computed as a sum of the proportions on the diagonal of the joint distribution of true score and observed score levels, as indicated by shaded areas in Table 86. Actually, it is a proportion (or percentage) of correct classifications across all the levels. In the particular example, the overall accuracy index for the FCAT Grade 3 Mathematics test equals 0.68. It means that 68 percent of students are classified in the same performance categories based on their observed scores, as they would be classified based on their true scores if they could be known.

The *overall consistency* index is analogously computed as a sum of the diagonal cells in the consistency table. Using the data from Table 87, it can be determined that the overall consistency index for the FCAT Grade 3 Mathematics test equals 0.56. In other words, 56 percent of Grade 3 students would be classified in the same performance levels based on the alternate form, if they would have taken it. Another way to express *overall consistency* is to use Cohen's *kappa* (κ) coefficient (Cohen, 1960). Kappa is a measure of "... how much agreement exists beyond chance alone. . ." (Fleiss, 1973, p. 146), which means that it assesses the proportion of consistent classifications between two forms after removing the proportion of consistent classifications that would be expected by chance alone. Using the data from Table 87 for computation, Cohen's κ for the FCAT Grade 3 Mathematics test equals 0.43. Compared to the previously described overall consistency estimate, Cohen's κ has lower value because it is corrected for chance.

Consistency conditional on level is computed as the ratio between the proportion of correct classifications at the selected level (diagonal entry) and the proportion of all the students classified into that level (marginal entry). In Table 87, the row LEVEL 4 is outlined and corresponding cells are shaded. The ratio between 0.13 (proportion of correct classifications) and 0.25 (total proportion of students classified into the LEVEL 4) yields 0.52, which represents the index of consistency of classification for the FCAT Grade 3 Mathematics test that is conditional on LEVEL 4. It indicates that 52 percent of all the students whose performance is classified as LEVEL 4 would be classified in the same level based on the alternate form, if an alternate form were taken.

Accuracy conditional on level is analogously computed. The only difference is that both row and column marginal sums are the same in the consistency table, whereas, the sum that is based on true status is used as a total for computing accuracy conditional on level in the accuracy table. For example, in Table 88, the proportion of agreement between true score status and observed score status at LEVEL 1 is 0.18, whereas, the total proportion of students with true score status at this level is 0.21. The accuracy conditional on level is equal to the ratio between those two proportions which yields 0.86. It indicates that 86 percent of the students estimated to have true score status on LEVEL 1 are correctly classified into that category by their observed scores on the FCAT Grade 3 Mathematics test.

Perhaps the most important indices for accountability systems are those for the accuracy and consistency of classification decisions made at specific cut points. To evaluate decisions at specific cut points, the joint distribution of all the performance levels are collapsed into a dichotomized distribution around that specific cut point. For example, the dichotomization at the cut point that separates LEVEL 1 through LEVEL 3 (combined) from LEVEL 4 and LEVEL 5 (combined) for the FCAT Grade 3 Mathematics test is depicted in Table 88. The proportion of correct classifications below that particular cut point is equal to the sum of the cells in the upper left shaded area (0.63), and the proportion of correct classifications above that particular cut point is equal to sum of the cells in the lower right shaded area (0.28).

Table 88. FCAT 2003 Grade 3 Mathematics true scores vs. observed scores cross-tabulation (Accuracy Table)

True Score	Observed Score					
	LEVEL 1	**LEVEL 2**	**LEVEL 3**	**LEVEL 4**	**LEVEL 5**	**Total**
LEVEL 1	0.18	0.03	0.00	0.00	0.00	0.21
LEVEL 2	0.04	0.10	0.04	0.00	0.00	0.17
LEVEL 3	0.00	0.05	0.19	0.05	0.00	0.29
LEVEL 4	0.00	0.00	0.06	0.19	0.06	0.31
LEVEL 5	0.00	0.00	0.00	0.01	0.02	0.03
Total	0.22	0.17	0.29	0.25	0.08	1.00

Note: Columns and row totals are computed from non-rounded values. Shaded cells are used for computing accuracy at a specific cut point.

The *accuracy index at cut point* is computed as the sum of the proportions of correct classifications around a selected cut point. In our example from Table 88, the sum of two shaded areas equals 0.91, which means that 91 percent of students were correctly classified either above or below the particular cut point. The sum of the proportions in the upper right non-shaded area (0.05) indicates false positives (i.e., there are 5 percent of students classified above the cut point by their observed scores, but falling below the cut point by their true scores), and the sum of the lower left non-shaded area (0.06) is the proportion of false negatives (i.e., there are 6 percent of students with observed levels below cut point whose true levels are above the cut point).

The *consistency at cut point* is obtained in an analogous way. For example, if data are taken from Table 87 and the distribution is dichotomized at the cut point between ‘LEVEL 1’ and all other levels combined, it can be determined that the proportion of correct classifications around that cut point equals 0.91. This means that 91 percent of students would be classified by an alternate form (if they would have taken it) in the same two categories (LEVEL 1 vs. LEVEL 2 through LEVEL 5 combined) as they were classified by the actual form taken.

Accuracy and Consistency Results for FCAT 2003

Detailed tables with accuracy and consistency cross-tabulations, dichotomized cross-tabulations, overall indices, indices conditional on level, and indices by cut point are presented in Appendix D. In this section, summary tables for all grades and subject areas are presented showing overall accuracy and consistency indices, accuracy indices at specific level, and accuracy and consistency indices at cut points.

Table 89. Estimates of Accuracy and Consistency of Performance-Level Classification by Grade and Subject

Grade	Subject	Accuracy	Consistency	Kappa (κ)
3	Reading	0.731	0.640	0.522
	Mathematics	0.667	0.563	0.434
4	Reading	0.721	0.621	0.485
	Mathematics	0.685	0.581	0.441
5	Reading	0.715	0.614	0.489
	Mathematics	0.716	0.613	0.484
6	Reading	0.679	0.584	0.455
	Mathematics	0.614	0.527	0.366
7	Reading	0.698	0.604	0.476
	Mathematics	0.650	0.554	0.406
8	Reading	0.697	0.599	0.461
	Mathematics	0.691	0.587	0.457
9	Reading	0.679	0.593	0.432
	Mathematics	0.678	0.572	0.438
10	Reading	0.617	0.537	0.383
	Mathematics	0.723	0.615	0.475

The overall indices of accuracy and consistency of classification for FCAT 2001 tests are presented in Table 89. It can be seen from the above table that overall accuracy indices are in the range between 0.61 and 0.73, overall consistency indices range between 0.53 and 0.64, and κ coefficients fall in the range between 0.37 and 0.52.

In addition to overall ratings of decision accuracy, the levels of agreement at each performance level are also of interest. Table 90 displays the probability of students being classified as being in a particular performance level, given that their “true status” was the same category. It can be seen that in most tests the accuracy indices at the lowest performance level (LEVEL 1) are substantially higher than at other levels. Also, the accuracy at the highest performance level is typically elevated, but this is not so evident in the current data. The higher accuracy at extreme levels is due to the fact that extreme performance levels usually cover a wider range of the measured construct than the intermediate levels, and misclassification can occur in only one direction. It should be noted that in the current data the percentage of students whose observed scores are classified in the highest performance level are relatively low (it is below 5 percent in all the tests: see Appendix D), which makes indices conditional at that level less reliable. In several instances, the percentage of students whose estimated true scores fall in LEVEL 5 is equal to zero which makes the estimation of the accuracy at that level impossible; however, it is possible to estimate accuracy of decisions at the cut point between LEVEL 4 and LEVEL 5, and, moreover, this estimate can be high (see Table 91).

Table 90. Estimated Probability of Being Classified at a Proficiency Level given that the "True Status" is that Level by Grade and Subject

Grade	Subject	Level 1	Level 2	Level 3	Level 4	Level 5
3	Reading	0.878	0.481	0.654	0.789	0.643
	Mathematics	0.863	0.559	0.649	0.615	0.617
4	Reading	0.893	0.480	0.595	0.753	*
	Mathematics	0.870	0.595	0.622	0.643	*
5	Reading	0.876	0.538	0.638	0.731	*
	Mathematics	0.908	0.622	0.578	0.711	*
6	Reading	0.877	0.527	0.556	0.666	0.593
	Mathematics	0.884	0.449	0.512	0.512	*
7	Reading	0.904	0.501	0.635	0.651	0.525
	Mathematics	0.895	0.519	0.564	0.528	*
8	Reading	0.899	0.595	0.641	0.628	*
	Mathematics	0.922	0.597	0.662	0.586	*
9	Reading	0.888	0.557	0.517	0.477	0.551
	Mathematics	0.893	0.578	0.618	0.611	*
10	Reading	0.899	0.619	0.434	0.326	*
	Mathematics	0.914	0.581	0.558	0.745	*

* No accuracy estimates were calculated at 'LEVEL 5' because the number of estimated true scores at this level is zero.

The most important decisions about student scores often involve dichotomous choices. For example, the stakes are usually highest regarding decisions made at the pass-fail cut point, which makes it desirable to know the accuracy and consistency of dichotomous decisions made around that specific cut point. Another example is if a college awards credits to advanced and proficient students who achieve LEVEL 5 and LEVEL 4, but not to those in LEVEL 1 through LEVEL 3, the focus of interest would be in accuracy and consistency of dichotomous decisions below, versus at and above the 'LEVEL 4' threshold. Reporting in a "percent at-or-above cut" (PAC) metric requires a judgment about whether the student score is below or at-or-above a particular cut point. Table 91 presents the accuracy and consistency information for these dichotomous categorizations.

Table 91. Accuracy and consistency of dichotomous categorizations by grade and subject

Grade	Subject	Accuracy				Consistency			
		1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	Reading	0.931	0.916	0.910	0.965	0.903	0.882	0.873	0.950
	Mathematics	0.934	0.907	0.888	0.931	0.907	0.869	0.844	0.907
4	Reading	0.941	0.921	0.872	0.977	0.917	0.888	0.819	0.958
	Mathematics	0.934	0.901	0.858	0.985	0.907	0.861	0.807	0.973
5	Reading	0.933	0.912	0.894	0.969	0.906	0.876	0.852	0.951
	Mathematics	0.949	0.919	0.864	0.978	0.928	0.885	0.808	0.960
6	Reading	0.923	0.904	0.896	0.947	0.891	0.865	0.855	0.926
	Mathematics	0.926	0.883	0.800	0.974	0.895	0.830	0.754	0.951
7	Reading	0.927	0.911	0.901	0.952	0.897	0.874	0.861	0.933
	Mathematics	0.929	0.896	0.841	0.966	0.900	0.852	0.791	0.940
8	Reading	0.933	0.900	0.868	0.991	0.905	0.859	0.823	0.984
	Mathematics	0.953	0.930	0.861	0.943	0.933	0.900	0.807	0.907
9	Reading	0.904	0.892	0.908	0.959	0.864	0.848	0.872	0.940
	Mathematics	0.936	0.905	0.870	0.961	0.909	0.866	0.820	0.935
10	Reading	0.919	0.868	0.856	0.930	0.884	0.814	0.814	0.893
	Mathematics	0.956	0.931	0.879	0.952	0.937	0.901	0.825	0.918

The data in Table 91 reveals that the level of agreement in terms of both accuracy and consistency for these dichotomous categorizations is very high—above 80 percent in all but one case. The level of agreement for decision accuracy falls below 85 percent in only two instances. Although the rates of agreement for decision consistency are slightly lower, the rate of agreement falls below 80 percent only in two instances. In general, high rates of accuracy and consistency are available to support decisions about PACs.

The issue of dichotomous classifications has particular relevance in the case of high-stakes situations, such as that exemplified by the high school graduation standard associated with the Grade 10 test. Students hoping to receive a standard high school diploma are required, among other things, to achieve a score of 287 or better on the Grade 10 FCAT Reading test and a score of 295 or better on the Grade 10 FCAT Mathematics test. In principle, it is possible for three situations to be found:

1. Observed performance of students is accurately reflected in terms of the standard and in terms of their true level of ability. (Students whose ability is at or above the minimum acceptable standard achieve test scores at or above that standard. Students whose true ability is below the standard achieve scores below the standard.)

2. Students whose true ability is below the standard receive scores that are, in fact, above the standard (“False Positives”).

3. Students whose true ability is, in fact, above the standard, but whose observed scores indicate (inaccurately) that they have not met the standard. ("False Negatives" that will, inappropriately, be required to take the test again.)

False positive and false negative rates for all dichotomous classifications for FCAT tests are presented in Table 92. An examination of the FCAT results for the Grade 10 Reading and Mathematics tests, in terms of the high school standards, reveals the following:

- Grade 10 Reading has the fail-pass threshold that is the same as the threshold between performance LEVELS 1 and 2. The accuracy of fail-pass decisions for this test is equal to the accuracy of dichotomous categorization between LEVEL 1 and LEVELS 2, 3, 4, and 5 combined. It can be seen from Table 91 that 92 percent of the students are correctly classified into either the pass or fail category (situation 1) based on their observed performance in Grade 10 Reading.

- Because the threshold score for fail-pass decisions in Grade 10 Mathematics falls in the middle of performance LEVEL 2, a separate analysis to estimate the accuracy of fail-pass decisions for this test was performed. The analysis shows that 95 percent of students were classified correctly into either a pass or fail category (situation 1) based on their observed performance in Grade 10 Mathematics.

Table 92. Accuracy of Dichotomous Categorizations: False Positive and False Negative Rates

Grade	**Subject**	**False Positives**				**False Negatives**			
		1 / 2+3+4+5	**1+2 / 3+4+5**	**1+2+3 / 4+5**	**1+2+3+4 / 5**	**1 / 2+3+4+5**	**1+2 / 3+4+5**	**1+2+3 / 4+5**	**1+2+3+4 / 5**
3	Reading	.032	.037	.052	.024	.037	.047	.039	.011
	Mathematics	.029	.042	.051	.058	.037	.051	.060	.010
4	Reading	.028	.031	.057	.023	.030	.047	.071	.000
	Mathematics	.032	.038	.064	.015	.033	.061	.078	.000
5	Reading	.033	.040	.056	.031	.034	.048	.050	.000
	Mathematics	.023	.038	.051	.022	.028	.043	.086	.000
6	Reading	.037	.041	.061	.046	.040	.055	.043	.007
	Mathematics	.035	.048	.106	.026	.039	.069	.094	.000
7	Reading	.029	.042	.048	.041	.044	.048	.051	.006
	Mathematics	.032	.040	.067	.034	.038	.064	.092	.000
8	Reading	.027	.046	.063	.009	.040	.054	.069	.000
	Mathematics	.019	.030	.062	.057	.028	.040	.078	.000
9	Reading	.048	.058	.056	.038	.049	.050	.037	.003
	Mathematics	.027	.045	.061	.039	.037	.050	.069	.000
10	Reading	.033	.049	.090	.070	.048	.083	.054	.000
	Mathematics	.018	.031	.050	.048	.026	.039	.072	.000

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-47.

Fleiss, J.L. (1973). *Statistical methods for rates and proportions*. New York: Wiley.

Florida Department of Education (1996). *Sunshine State Standards.* Retrieved September 20, 2002, from the Florida Department of Education web site: http://www.firn.edu/doe/curric/prek12/frame2.htm.

Florida Department of Education (1998). *Technical Report: Florida Comprehensive Assessment Test (FCAT): 1998.* Unpublished. Tallahassee, FL: Author.

Florida Department of Education (2000). *The FCAT 2001 Test Construction Specifications.* Unpublished. Tallahassee, FL: Author.

Florida Department of Education (2000, October). *Plan for Selecting the Calibration Sample for the 2001 FCAT Administration.* Unpublished. Tallahassee, FL: Author.

Florida Department of Education (2001, May). *Analysis of the FCAT Test Item Review Conducted by the Florida Department of Education and Harcourt Educational Measurement.* Unpublished. Tallahassee, FL: Author.

Florida Department of Education (2001, November 6). *Florida Comprehensive Assessment Test Achievement Level Setting Technical Report*. Unpublished. Tallahassee, FL: Author.

Florida Department of Education (2001, November). *Florida Comprehensive Assessment Test: Technical Report on Vertical Scaling for Reading and Mathematics*. Unpublished. Tallahassee, FL: Author.

Florida Department of Education (2002, January). *Florida Comprehensive Assessment Test Technical Report Field Test Supplement for Test Administration in Spring 2001*. Unpublished. Tallahassee, FL: Author.

Hoffman, R.G., Wise, L.L., Thacker, A.A., and Ford, L.A. (2002). *Technical Report on Vertical Scaling for Reading and Mathematics*. San Antonio, TX: Harcourt Educational Measurement.

Lee, W., Hanson, B. A., & Brennan, R. L. (2000, October). *Procedures for computing classification consistency and accuracy indices with multiple categories*. (ACT Research Report Series 2000-10). Iowa City, IO: ACT, Inc.

Livingston, S. A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32(2),* 179-197.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *Journal of American Statistical Association. 58,* 690-700.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Measurement, 7,* 159-176.

Rogosa, D. (2000). Statistical topics in educational assessment: individual scores, group summaries, and accountability systems. Presented to the March 14, 2000 CCSSO Technical Issues in Large Scale Assessment Workshop, San Diego, California.

Rogosa, D. (1994). Misclassification in student performance levels. In CTB/McGraw-Hill. (1994). 1994 CLAS Assessment Technical Report. Monterrey, CA: Author.

Stocking, M. L. & Lord, F. M., (1983). Developing a common metric in item response theory. *Applied Measurement*, 7, 201-210.

Thissen, D. (1991). Multilog™ User's Guide. Lincolnwood, IL: Scientific Software.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5,* 2, 245-262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 2,* 125-145.

Young, M. J. & Yoon, B. (1998, April). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment.* (CSE Technical Report 475). Center for the Study Evaluation, National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles, CA: University of California, Los Angeles.

Zwick, R., Donoghue, J. R.,+- & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement. 30(3),* 233-251.