FLORIDA COMPREHENSIVE ASSESSMENT TEST (FCAT) FOR READING AND MATHEMATICS

Technical Report For Test Administrations of FCAT 2002

Produced Jointly by
Human Resources Research Organization
(HumRRO)
Alexandria, Virginia
Under subcontract to and in concert with
Harcourt Educational Measurement
San Antonio, TX

Harcourt
Educational Measurement
A Harcourt Assessment Company

San Antonio, TX

September 2002

TABLE OF CONTENTS

INTRODUCTION AND OVERVIEW

This report presents information on the measurement characteristics of the reading and mathematics assessments that were included in the Florida Comprehensive Assessment Test (FCAT) for spring 2002. These characteristics provide an indication of the current quality of FCAT assessments in these two content areas.

The report is technical in nature; however, an attempt has been made to make it accessible to an audience with a basic understanding of testing concepts. Summary data are provided in the report itself. Detailed data may be found in appendices published under a separate cover.

Description of FCAT

As part of a student assessment and school accountability program of the Florida Department of Education (FDOE), the FCAT assessments are designed to measure student mastery and achievement of reading and mathematics content as described by the *Sunshine State Standards* (FDOE, 1996). The administration of the 1998 and 1999 FCAT included tests in reading for Grades 4, 8, and 10, and in mathematics for Grades 5, 8, and 10. In the spring of 2000, students in Grades 3, 5, 6, 7, and 9 took a field-test version of the reading assessment; and students in Grades 3, 4, 6, 7, and 9 took a field-test version of the mathematics assessment. The ten new grade and subject combinations for reading and mathematics became part of FCAT in 2001. The 2002 FCAT includes both reading and mathematics tests for Grades 3-10.

The 2002 FCAT Mathematics Test included 40 core items for Grades 3 and 4; 44 items for Grades 6, 7, and 9; and 50 items for Grades 5, 8, and 10. The 2002 FCAT Reading Test included 40 core items for Grade 3; 41 items for Grade 4; 43 items for Grades 5, 6, and 7; and 44 items for Grades 8, 9, and 10.

The tests varied somewhat in item format. All 16 tests included multiple-choice (MC) items. In addition, mathematics tests in Grades 5 and higher included gridded-response (GR) items that required students to determine numerical answers by filling in corresponding bubbles in grids on an answer sheet. Both MC and GR items were machine-scored.

The FCAT administered since 1998 (Grades 4, 8, and 10 for reading and Grades 5, 8, and 10 for mathematics) also included two types of performance tasks (sometimes called constructed-response items) that required students to write responses to open-ended questions. The two types of performance tasks (PT) differed in the length of the response and in the number of points possible. While a correct MC or GR answer was assigned 1 point, student responses to short-response (SR) items may be assigned 0, 1, or 2 points, depending on the accuracy and thoroughness of the response. Likewise, student responses to extended-response (ER) items may be assigned 0, 1, 2, 3, or 4 points. These items were hand-scored by trained raters using a process that is described later in the report. In this report, grades that included PT items are referred to as the PT Grades.

In the spring 2002 administrations, FCAT reading and mathematics assessments also included field-test items and vertical-scaling items. To accommodate these items, 30 separate test forms were constructed for each grade and subject combination. All forms within a grade and subject contained the same core items, plus six to eight extra items. Field-test items were dispersed among 24 forms in order to collect data for a relatively large number of items while only requiring any one student to complete a small number of items. For the remaining six forms, items from adjacent grades were used to construct a vertical scale linking each of the tested grades. A complete description of the vertical scaling process and results can be found in a separate report (FDOE, 2001, November; FDOE, 2002, September).

Score reports consisted of overall reading and mathematics scale scores, plus performance category assignments and a number of subscores for components of the tests. Performance category assignments were based on standard setting procedures that divided both the reading and mathematics scales into five distinct levels (FDOE, 1998, 2001, November 6,). The FCAT reading tests reported subscores in four reporting categories (also referred to as reading clusters): (a) Words and Phrases in Context; (b) Main Idea, Plot, and Purpose; (c) Comparisons and Cause/Effect; and (d) Reference and Research. FCAT mathematics tests provided subscores in five reporting categories (also referred to as mathematics strands): (a) Number Sense, Concepts and Operations; (b) Measurement; (c) Geometry and Spatial Sense; (d) Algebraic Thinking; and (e) Data Analysis and Probability.

Report Contents

Test validity and reliability analyses are key in establishing the quality of an achievement test such as the FCAT. These two issues are intertwined since measurement error, 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. It includes data at several levels of analysis, including item-level statistics, test- or student-level statistics, and state-level statistics. This report includes background information on item response theory (IRT) scoring methods (Lord & Novick, 1968) used to process FCAT data. This report also contains summary statistics to represent 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 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. However, 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 that actually contributes to reported student scores), every student will answer some field-test questions that do not count toward his or her/his score. These field-test items provide data for the development of future tests. This report refers only to core-test items. A supplemental report (FDOE, 2002, January) 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 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 FCAT assessments, test items go though a three-phase developmental process. The first phase involves drafting items to match FCAT style and *Sunshine State Standards* (SSS) benchmarks.

Education professionals familiar with both the FCAT style and intent of each of the FCAT benchmarks initially draft items. Drafted items received by the contractor are subjected to a critical content and editorial review and forwarded to content staff at the Test Development Center (TDC) in Tallahassee where they receive an additional review. Items are typically reviewed and accepted with no or minor edits, rejected as being inappropriate for FCAT, or returned to the contractor with comments regarding necessary changes in style or focus before the items can be moved further through the review process. A dialogue between the contractor and TDC staff on the “accept with revisions” items assures that both the contractor and 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 2002 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, 2001, May). Reviews were conducted by the following groups: (1) the Florida Department of Education committees 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 areas cited above, reviewed mathematics and reading items as well as reading passages on which the FCAT reading items are based. Similar procedures for passage and item reviews were followed in previous years for core items in FCAT tests.

After this review process, items are included as field-test items during regular FCAT administrations. These 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 staff and staff from the TDC build forms through a process involving many steps. Typically, Harcourt content and psychometric staff propose draft forms by grade and subject for TDC staff review. These draft forms have been 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 and clusters) match the characteristics of previously-administered versions of FCAT.

CONSTRUCTED-RESPONSE SCORING PROCEDURES

Scorer Training

As previously noted for some grade and content combinations, open-ended questions require students to provide handwritten responses. These responses are judged individually by human scorers rather than by machines. Scorers are trained with FDOE-approved assessment materials that include scores agreed upon during the rangefinder review sessions held with state educators. Potential scorers are given an overview of the project and informed of FDOE expectations and guidelines. Scorers are shown several sets of training papers to ground them in the scoring rules and are tested on “qualification sets” to ensure quality control. Items are scored in groups of two or more (which is known as the rater item block or RIB format), and the scorer must qualify on all items before scoring all items within the RIB. Only after successful completion of the qualifying process are scorers allowed to assess actual student responses. In the event that an item, or group of items, is presented to more than one group of scorers at separate times, training papers are distributed in the same order with the same comments to ensure consistency between training sessions. This is done so that each group of scorers will complete training with 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, where scoring rationales are discussed. Only 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 look back at actual student responses that have been scored by members of their team (which consists of no more than 12 scorers and one Team Leader). This process helps ensure that scorers are assigning correct scores to student responses. At the beginning of the project, Team Leaders spend their time doing read-behinds for each scorer several times a day; this tends to identify the strengths of individual scorers. Team Leaders may ask scorers to review and re-score papers that have been incorrectly scored to help the scorer (who has failed to adhere to the standards) to understand how his or her scoring is in error. Throughout the project, the read-behind process continues to ensure test scoring accuracy.

Control of Scorer Drift

There are many methods implemented for control of scorer drift. One category of methods involves training and supervisory feedback to scorers during the scoring session. A second category of methods involves the review of statistical information about scorer agreement. The statistical methods are used to inform the leaders of the scoring session and the team leaders about group and individual needs for feedback or intervention.

One of the training methods implemented daily during scoring is for each scorer to review the rangefinder and training papers, previously used in training. Typically the first 15 minutes (or longer, if needed) of each day is spent reviewing these papers. This method helps to keep all scorers and team leaders grounded in the rules and guidelines for scoring.

Another process, called read-behinds, is where the team leader reads papers scored by each scorer on his/her team. This process provides the opportunity for one-on-one feedback when scores may be drifting away from the established criteria.

A third training and feedback process involves large or small group training sessions conducted periodically during the weeks of scoring. These sessions, referred to as calibration sessions, involve the presentation and scoring of unique papers. Group discussions of the calibration papers help with the control of scorer drift by reinforcing the established criteria for scoring.

The availability of statistical information about each scorer and the entire group provides valuable information to the scoring directors and team leaders about the quality of the scoring. This information is also very helpful in controlling scorer drift. Both reliability and validity data are available in a series of reports accessible at any point during the scoring session. Using these reports, scoring directors and team leaders can check which papers are being scored incorrectly, which readers are assigning incorrect scores, and whether readers are scoring too high or too low. Appropriate adjustments in the form of training, group or individual feedback, or intervention can be made based on this information.

The validity reports are based on the scores readers assign to a series of pre-scored papers, called validity papers. These validity papers are randomly embedded into the scoring stream and the reader's scores are compared to the "true" score for each paper. Reports of the score a reader assigns to each validity paper are available for review each day and are useful in determining whether the scoring criteria are being applied correctly or if some scorer drift is occurring.

Reliability reports can also be used to control scorer drift. The reliability reports indicate the degree to which a single reader is in agreement with other readers. Therefore, they indicate whether a reader is drifting from the established standards the group is using in as much as he/she is consistently high or consistently low. Team leaders can use this information to identify individual readers who need to be more closely monitored and who may need additional training or intervention.

2002 FCAT STATISTICS

This section of the report presents psychometric analyses of the 2002 FCAT core assessments. Because of the requirements for rapid turnaround in score reporting, traditional item analyses and IRT analyses for the initial reporting period were conducted using a special calibration sample of students. A set of schools was chosen specifically for this purpose and those schools returned their student responses on an early timeline. The general strategy was to select schools that would provide a sample of students representative of the state’s regions, ethnic diversity, and achievement scores in past years. Only standard curriculum students were used in the analyses; exceptional student education (ESE) students and students in the limited English proficiency (LEP) program for two or fewer years were excluded. In addition, students in the calibration sample had to meet criteria indicating they had attempted the test.[1] More details about the selection of this sample appear in *Plan for Selecting the Calibration Sample for the 2002 FCAT Administration* (FDOE, 2001, October).

Because of the importance of the calibration sample, this section (although it is out of chronological order) begins with a comparison of the calibration sample to the state’s total distribution of students. It is recognized that this comparison could only be made after all of the analyses were completed. However, the comparison is presented here to establish the credibility of the remaining analyses.

[1] Test scores were computed only for students who met criteria showing that they attempted to take the test. The criteria are that a student has at least six non-blank answers in each of two sessions, with the exception of Grade 4 reading and Grade 5 mathematics which required at least four non-blank responses in each of the three sessions.

Calibration Sample Review

The tables on the following pages compare each grade and subject calibration sample with other statewide sets of students. One set of comparison students, labeled “total population,” includes all students with FCAT records for March 2002. Some students who took the test, however, did not receive FCAT scores because they did not answer enough questions, that is, they did not meet 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 sample. Note also that, because of missing ethnicity and gender information, the numbers of students across the respective categories does not match the totals listed.

In addition to the gender and ethnicity distributions, test scores for the calibration sample are compared to those for the total population and the standard curriculum population. Test score means for these groups are also disaggregated by ethnicity and gender. For this comparison, students who did not meet attemptedness criteria are not included. Three sets of tables of statistics are presented for each grade and subject on the following pages. Tables 1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 31, 34, 37, 40, 43, and 46 show ethnicity distributions. These tables indicate that ethnicity representations of the calibration sample are a reasonable approximation of the state distributions, and this match tends to be better for the standard curriculum distributions.

Tables 2, 5, 8, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44 and 47 show gender distributions; these results for standard curriculum students are also similar to those for the total population.

Tables 3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33, 36, 39, 42, 45 and 48 present FCAT score means and standard deviations. Score means are lower and standard deviations are higher for all 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 sample are also reflected in the calibration sample.

This pattern of results supports the representativeness of the calibration sample. If analyses were conducted on the full set of standard curriculum students, slight differences in the results might be observed; however, such differences should have no practical impact.

Grade 3 Reading Calibration Sample

Table 1. Grade 3 Reading: Number and Percent by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration Sample	65 1.40%	978 21.05%	1102 23.72%	14 0.30%	103 2.22%	2361 50.83%	4645
Total std curriculum	2900 1.79%	39366 24.26%	31220 19.24%	491 0.30%	4110 2.53%	83518 51.47%	162258
Total population	3362 1.75%	47076 24.47%	40438 21.02%	561 0.29%	4646 2.41%	95187 49.48%	192385

Table 2. Grade 3 Reading: Number and Percent by Gender

	Male	Female	Total
Calibration sample	2233 48.07%	2405 51.78%	4645
Total std curriculum	79522 49.01%	82403 50.79%	162258
Total population	98448 51.17%	93363 48.53%	192385

Table 3. Grade 3 Reading: Score Distributions

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	301.42	59.23	4645	303.42	60.12	161792	293.01	65.98	188546
Male	299.32	59.48	2233	302.15	61.61	79266	289.37	68.12	96357
Female	303.37	58.97	2405	304.74	58.58	82206	296.96	63.35	91779
African American	274.43	55.15	978	274.15	56.42	39206	265.06	60.91	45880
Hispanic	287.79	58.00	1102	290.69	58.39	31122	276.98	64.78	39519
White	318.48	55.78	2361	321.122	56.14	83344	312.40	62.53	93897

Grade 3 Mathematics Calibration Sample

Table 4. Grade 3 Mathematics: Number and Percent by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	65 1.40%	978 21.07%	1102 23.74%	14 0.30%	102 2.20%	2358 50.81%	4641
Total std curriculum	2900 1.79%	39366 24.26%	31220 19.24%	491 0.30%	4110 2.53%	83518 51.47%	162258
Total population	3362 1.75%	47076 24.47%	40438 21.02%	561 0.29%	4646 2.41%	95187 49.48%	192385

Table 5. Grade 3 Mathematics: Number and Percent by Gender

	Male	Female	Total
Calibration sample	2230 48.05%	2404 51.80%	4641
Total std curriculum	79522 49.01%	82403 50.79%	162258
Total population	98448 51.17%	93363 48.53%	192385

Table 6. Grade 3 Mathematics: Score Distributions

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	311.18	58.89	4641	311.27	61.37	161824	301.79	66.83	188765
Male	314.06	60.51	2230	315.09	62.87	79283	303.01	69.45	96515
Female	308.47	57.25	2404	307.70	59.59	82218	300.65	63.86	91836
African American	280.34	58.82	978	277.64	59.99	39218	268.79	64.09	45935
Hispanic	301.75	60.22	1102	303.48	60.62	31137	291.02	66.92	39601
White	327.54	52.77	2358	328.90	54.91	83346	321.21	60.67	93964

Grade 4 Reading Calibration Sample

Table 7. Grade 4 Reading: Number and Percent by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	65 1.35%	1124 23.36%	1044 21.70%	17 0.35%	89 1.85%	2466 51.26%	4811
Total std curriculum	3046 1.89%	38945 24.13%	30254 18.74%	482 0.30%	3755 2.33%	84384 52.27%	161424
Total population	3506 1.79%	48032 24.50%	40079 20.44%	563 0.29%	4288 2.19%	98640 50.31%	196079

Table 8. Grade 4 Reading: Number and Percent by Gender

	Male	Female	Total
Calibration sample	2324 48.31%	2480 51.55%	4811
Total std curriculum	78388 48.56%	82660 51.21%	161424
Total population	100016 51.01%	95423 48.67%	196079

Table 9. Grade 4 Reading: Score Distributions

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	311.33	53.86	4811	312.61	53.13	161199	299.47	63.25	192046
Male	306.87	55.74	2324	309.63	53.08	78269	293.62	64.96	97838
Female	315.60	51.73	2480	315.54	52.97	82566	305.73	60.73	93703
African American	284.73	52.62	1124	287.01	51.63	38868	274.34	60.26	46769
Hispanic	301.52	55.20	1044	303.35	54.01	30223	284.55	66.58	39296
White	326.75	48.34	2466	326.75	48.27	84288	316.30	57.87	97040

Grade 4 Mathematics Calibration Sample

Table 10. Grade 4 Mathematics: Number and Percent by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	66 1.42%	1088 23.37%	958 20.58%	17 0.37%	87 1.87%	2429 52.18%	4655
Total std curriculum	3063 1.89%	39117 24.14%	30325 18.72%	482 0.30%	3766 2.32%	84649 52.25%	162015
Total population	3507 1.79%	48032 24.50%	40080 20.44%	563 0.29%	4289 2.19%	98640 50.31%	196082

Table 11. Grade 4 Mathematics: Number and Percent by Gender

	Male	Female	Total
Calibration sample	2249 48.31%	2405 51.66%	4655
Total std curriculum	78757 48.61%	82883 51.16%	162015
Total population	100018 51.01%	95424 48.67%	196082

Table 12. Grade 4 Mathematics: Score Distributions

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	302.31	55.45	4655	304.47	56.31	161605	293.57	63.43	192549
Male	303.83	56.91	2249	308.14	56.95	78531	294.42	65.52	98168
Female	300.90	54.03	2405	301.13	55.42	82714	292.86	61.09	93912
African American	272.75	54.43	1088	273.24	55.25	38986	262.33	61.58	46903
Hispanic	292.93	56.43	958	297.45	55.73	30256	283.11	64.02	39401
White	318.60	48.89	2429	320.20	50.22	84479	311.43	57.17	97268

Grade 5 Reading Calibration Sample

Table 13. Grade 5 Reading: Number and Percent by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	86 1.88%	1027 22.47%	982 21.49%	19 0.42%	69 1.51%	2381 52.10%	4570
Total std curriculum	3197 2.00%	37467 23.45%	30075 18.83%	456 0.29%	3013 1.89%	84894 53.14%	159743
Total population	3592 1.83%	46980 23.90%	40148 20.43%	569 0.29%	3424 1.74%	100757 51.27%	196537

Table 14. Grade 5 Reading: Number and Percent by Gender

	Male	Female	Total
Calibration sample	2139 46.81%	2427 53.11%	4570
Total std curriculum	76711 48.02%	82717 51.78%	159743
Total population	99787 50.77%	96131 48.91%	196537

Table 15. Grade 5 Reading: Score Distributions

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	296.82	55.50	4570	297.61	54.18	159361	284.53	62.79	192876
Male	293.80	57.54	2139	295.03	55.92	76505	278.99	65.52	97814
Female	299.63	53.25	2427	300.13	52.32	82558	290.40	59.21	94665
African American	266.75	54.18	1027	268.72	51.82	37336	256.11	58.98	45849
Hispanic	286.38	51.37	982	286.96	52.95	29997	269.25	63.32	39414
White	313.07	51.26	2381	313.32	49.49	84746	302.61	57.97	99315

Grade 5 Mathematics Calibration Sample

Table 16. Grade 5 Mathematics: Number and Percent by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	87 1.83%	1073 22.62%	1063 22.41%	16 0.34%	71 1.50%	2426 51.15%	4743
Total std curriculum	3181 2.00%	37388 23.47%	29968 18.81%	456 0.29%	3002 1.88%	84695 53.17%	159295
Total population	3592 1.83%	46980 23.90%	40148 20.43%	569 0.29%	3424 1.74%	100757 51.27%	196537

Table 17. Grade 5 Mathematics: Number and Percent by Gender

	Male	Female	Total
Calibration sample	2217 46.74%	2520 53.13%	4743
Total std curriculum	76404 47.96%	82533 51.81%	159295
Total population	99787 50.77%	96131 48.91%	196537

Table 18. Grade 5 Mathematics: Score Distributions

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	329.47	49.47	4743	329.83	48.51	159096	318.04	58.09	192740
Male	332.05	49.29	2217	332.23	49.81	76313	317.37	61.03	97701
Female	327.36	48.52	2520	327.75	47.07	82446	318.89	54.77	94587
African American	299.41	50.06	1073	302.71	48.48	37334	289.51	58.82	45866
Hispanic	326.18	45.45	1063	325.60	47.75	29930	310.36	58.92	39402
White	343.23	43.61	2426	342.21	43.39	84618	332.95	51.70	99239

Grade 6 Reading Calibration Sample

Table 19. Grade 6 Reading: Number and Percent by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	132 2.73%	1369 28.28%	830 17.15%	11 0.23%	56 1.16%	2429 50.18%	4841
Total std curriculum	3290 2.01%	38840 23.75%	30964 18.93%	441 0.27%	2103 1.29%	87456 53.48%	163529
Total population	3673 1.87%	47344 24.11%	40115 20.43%	522 0.27%	2381 1.21%	101741 51.82%	196330

Table 20. Grade 6 Reading: Number and Percent by Gender

	Male	Female	Total
Calibration sample	2398 49.54%	2440 50.40%	4841
Total std curriculum	79240 48.46%	84078 51.41%	163529
Total population	100434 51.16%	95628 48.71%	196330

Table 21. Grade 6 Reading: Score Distributions

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	301.37	58.02	4841	304.57	57.30	162528	291.33	66.29	194624
Male	295.67	60.81	2398	301.66	60.21	78661	285.35	69.81	99332
Female	306.93	54.93	2440	307.40	54.22	83667	297.69	61.72	95035
African American	280.45	56.45	1369	277.41	54.44	38487	264.29	62.36	46733
Hispanic	285.15	61.25	830	291.23	57.98	30768	272.58	69.50	39750
White	317.39	52.33	2429	320.39	52.36	87056	310.07	60.20	101083

Grade 6 Mathematics Calibration Sample

Table 22. Grade 6 Mathematics: Number and Percent by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	132 2.73%	1370 28.29%	830 17.14%	11 0.23%	56 1.16%	2430 50.18%	4843
Total std curriculum	3292 2.01%	38840 23.75%	30964 18.93%	441 0.27%	2103 1.29%	87456 53.48%	163531
Total population	3675 1.87%	47344 24.11%	40115 20.43%	522 0.27%	2381 1.21%	101741 51.82%	196332

Table 23. Grade 6 Mathematics: Number and Percent by Gender

	Male	Female	Total
Calibration sample	2399 49.54%	2441 50.40%	4843
Total std curriculum	79240 48.46%	84080 51.42%	163531
Total population	100434 51.16%	95630 48.71%	196332

Table 24. Grade 6 Mathematics: Score Distributions

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	307.22	59.22	4843	310.32	56.89	162455	297.84	65.97	194443
Male	305.42	63.35	2399	311.41	59.30	78596	295.73	69.61	99162
Female	308.98	54.94	2441	309.40	54.43	83659	300.16	61.78	95027
African American	281.82	61.93	1370	279.04	57.60	38463	265.44	66.15	46665
Hispanic	293.23	59.81	830	301.40	55.30	30764	286.00	65.72	39746
White	323.96	50.54	2430	326.05	50.19	87002	315.95	58.82	100977

Grade 7 Reading Calibration Sample

Table 25. Grade 7 Reading: Number and Percent by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	120 2.29%	1450 27.66%	737 14.06%	21 0.40%	78 1.49%	2824 53.86%	5243
Total std curriculum	3434 2.11%	38721 23.75%	30218 18.53%	470 0.29%	1877 1.15%	87857 53.88%	163066
Total population	3845 1.97%	46967 24.08%	39188 20.09%	549 0.28%	2133 1.09%	101743 52.17%	195039

Table 26. Grade 7 Reading: Number and Percent by Gender

	Male	Female	Total
Calibration sample	2596 49.51%	2643 50.41%	5243
Total std curriculum	79218 48.58%	83614 51.28%	163066
Total population	99981 51.26%	94765 48.59%	195039

Table 27. Grade 7 Reading: Score Distributions

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	303.37	56.37	5243	305.36	57.95	161755	294.31	64.11	192946
Male	300.15	58.63	2596	304.34	59.81	78473	290.64	66.73	98671
Female	306.58	53.82	2643	306.43	56.06	83065	298.28	60.96	94007
African American	273.57	54.87	1450	273.34	56.48	38297	262.85	61.25	46284
Hispanic	294.43	54.78	737	294.57	56.62	29929	279.61	64.03	38708
White	319.16	50.43	2824	322.10	52.01	87324	313.21	58.13	100902

Grade 7 Mathematics Calibration Sample

Table 28. Grade 7 Mathematics: Number and Percent by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	120 2.29%	1456 27.73%	740 14.10%	21 0.40%	78 1.49%	2821 53.73%	5250
Total std curriculum	3434 2.11%	3871 23.75%	30218 18.53%	470 0.29%	1877 1.15%	87857 53.88%	163066
Total population	3845 1.97%	46967 24.08%	39188 20.09%	549 0.28%	2133 1.09%	101743 52.17%	195039

Table 29. Grade 7 Mathematics: Number and Percent by Gender

	Male	Female	Total
Calibration sample	2599 49.50%	2646 50.40%	5250
Total std curriculum	79218 48.58%	83614 51.28%	163066
Total population	99981 51.26%	94765 48.59%	195039

Table 30. Grade 7 Mathematics: Score Distributions

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	302.74	57.59	5250	304.33	59.05	161594	292.22	67.53	192714
Male	302.28	60.24	2599	305.41	61.79	78371	289.92	71.52	98523
Female	303.33	54.71	2646	303.44	56.23	83003	294.76	62.91	93917
African American	274.99	56.06	1456	272.07	59.77	38216	258.73	68.14	46156
Hispanic	294.11	56.85	740	294.62	57.64	29949	280.31	66.30	38750
White	317.02	52.22	2821	320.47	52.22	87218	310.54	60.55	100747

Grade 8 Reading Calibration Sample

Table 31. Grade 8 Reading: Number and Percent by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	112 2.43%	1258 27.29%	738 16.01%	13 0.28%	63 1.37%	2422 52.54%	4610
Total std curriculum	3449 2.18%	36669 23.22%	29110 18.44%	436 0.28%	1716 1.09%	86110 54.54%	157889
Total population	3826 2.03%	44227 23.50%	37552 19.95%	514 0.27%	1937 1.03%	99634 59.94%	188216

Table 32. Grade 8 Reading: Number and Percent by Gender

	Male	Female	Total
Calibration sample	2262 49.07%	2347 50.91%	4610
Total std curriculum	75831 48.03%	81723 51.76%	157889
Total population	95381 50.68%	92407 49.10%	188216

Table 33. Grade 8 Reading: Score Distributions

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	306.59	50.73	4610	306.96	53.39	156421	294.35	63.33	185969
Male	303.36	50.48	2262	303.90	55.05	75043	288.04	66.62	94038
Female	309.71	50.79	2347	309.96	51.49	81083	301.02	58.92	91550
African American	279.49	50.59	1258	277.42	52.63	36266	264.28	62.09	43571
Hispanic	297.34	48.34	738	296.12	54.43	28810	278.26	66.28	37080
White	322.35	44.67	2422	322.34	46.84	85420	312.67	55.77	98612

Grade 8 Mathematics Calibration Sample

Table 34. Grade 8 Mathematics: Number and Percent by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	111 2.39%	1265 27.27%	739 15.93%	13 0.28%	65 1.40%	2442 52.64%	4639
Total std curriculum	3449 2.18%	36669 23.22%	29110 18.44%	436 0.28%	1716 1.09%	86110 54.54%	157889
Total population	3826 1.03%	44227 23.50%	37552 19.95%	514 0.27%	1937 1.03%	99634 52.94%	188216

Table 35. Grade 8 Mathematics: Number and Percent by Gender

	Male	Female	Total
Calibration sample	2279 49.13%	2359 50.85%	4639
Total std curriculum	75831 48.03%	81723 51.76%	157889
Total population	95381 50.68%	92407 49.10%	188216

Table 36. Grade 8 Mathematics: Score Distributions

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	313.76	45.61	4639	315.76	48.36	156350	305.08	58.13	185835
Male	315.56	45.48	2279	317.38	50.13	75030	303.64	61.71	93964
Female	312.04	45.68	2359	314.40	46.49	81027	306.72	54.04	91493
African American	285.02	46.87	1265	284.73	50.85	36239	272.15	61.66	43485
Hispanic	309.36	38.18	739	306.47	46.79	28794	293.50	57.68	37090
White	328.56	39.34	2442	330.98	39.90	85388	322.67	48.67	98553

Grade 9 Reading Calibration Sample

Table 37. **Grade 9 Reading: Number and Percent by Ethnicity**

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	105 1.97%	1199 22.52%	1080 20.29%	12 0.23%	53 1.00%	2835 53.25%	5324
Total std curriculum	3867 2.13%	44260 24.42%	33962 18.74%	511 0.28%	1898 1.05%	95824 52.86%	181267
Total population	4280 2.00%	53213 24.90%	42994 20.12%	613 0.29%	2132 1.00%	109390 51.18%	213732

Table 38. **Grade 9 Reading: Number and Percent by Gender**

	Male	Female	Total
Calibration sample	2458 46.17%	2860 53.72%	5324
Total std curriculum	89434 49.34%	91353 50.40%	181267
Total population	110819 51.85%	102346 47.89%	213732

Table 39. **Grade 9 Reading: Score Distributions**

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	299.28	54.79	5324	296.27	56.25	177109	286.83	61.22	208242
Male	293.84	55.62	2458	292.52	57.54	87138	281.27	62.76	107542
Female	304.02	53.64	2860	300.06	54.64	89574	292.93	58.89	100226
African American	276.92	52.01	1199	268.33	53.34	42895	258.73	58.04	51382
Hispanic	282.22	54.91	1080	283.16	55.61	33032	271.04	60.82	41746
White	314.57	50.08	2835	312.95	51.24	94199	305.67	55.75	107276

Grade 9 Mathematics Calibration Sample

Table 40. Grade 9 Mathematics: Number and Percent by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	105 1.98%	1198 22.55%	1079 20.31%	12 0.23%	52 0.98%	2828 53.23%	5313
Total std curriculum	3867 2.13%	44260 24.42%	33962 18.74%	511 0.28%	1898 1.05%	95824 52.86%	181267
Total population	4280 2.00%	53213 24.90%	42994 20.12%	613 0.29%	2132 1.00%	109390 51.18%	213732

Table 41. Grade 9 Mathematics: Number and Percent by Gender

	Male	Female	Total
Calibration sample	2452 46.15%	2855 53.74%	5313
Total std curriculum	89434 49.34%	91353 50.40%	181267
Total population	110819 51.85%	102346 47.89%	213732

Table 42. Grade 9 Mathematics: Score Distributions

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	296.85	53.38	5313	294.61	56.29	176536	285.60	62.17	207305
Male	298.02	55.16	2452	296.76	58.10	86753	285.35	64.95	106866
Female	295.99	51.66	2855	292.71	54.26	89384	286.05	58.93	99961
African American	268.92	53.58	1198	261.96	55.61	42698	252.25	61.13	51007
Hispanic	282.41	53.15	1079	281.41	55.50	32934	271.47	61.38	41601
White	313.39	45.66	2828	313.00	48.00	93946	305.75	53.93	106891

Grade 10 Reading Calibration Sample

Table 43. Grade 10 Reading: Number and Percent by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	92 2.06%	998 22.35%	893 20.00%	12 0.27%	50 1.12%	2385 53.42%	4465
Total std curriculum	3317 2.49%	28060 21.06%	22992 17.26%	341 0.26%	1240 0.93%	76212 57.20%	133230
Total population	3744 2.34%	34756 21.68%	30442 18.99%	390 0.24%	1532 0.96%	87667 54.68%	160327

Table 44. Grade 10 Reading: Number and Percent by Gender

	Male	Female	Total
Calibration sample	1973 44.19%	2472 55.36%	4465
Total std curriculum	62185 46.67%	70217 52.70%	133230
Total population	78335 48.86%	80528 50.23%	160327

Table 45. Grade 10 Reading: Score Distributions

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	311.91	47.56	4465	310.72	47.90	132874	302.32	55.08	151702
Male	310.95	48.12	1973	310.68	48.90	62009	300.08	57.71	73762
Female	312.99	46.74	2472	311.12	46.71	70091	304.92	51.99	76963
African American	291.18	47.37	998	280.92	48.55	27950	271.65	55.70	32327
Hispanic	296.88	41.20	893	298.62	48.67	22923	285.76	57.58	28440
White	325.86	41.49	2385	325.10	40.57	76102	319.37	46.38	84325

Grade 10 Mathematics Calibration Sample

Table 46. **Grade 10 Mathematics: Number and Percent by Ethnicity**

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total
Calibration sample	92 2.07%	988 22.26%	877 19.76%	12 0.27%	49 1.10%	2378 53.57%	4439
Total std curriculum	3319 2.47%	28317 21.07%	23330 17.36%	342 0.25%	1259 0.94%	76667 57.05%	134392
Total population	3744 2.34%	34756 21.68%	30442 18.99%	390 0.24%	1532 0.96%	87667 54.68%	160327

Table 47. **Grade 10 Mathematics: Number and Percent by Gender**

	Male	Female	Total
Calibration sample	1957 44.09%	2457 55.35%	4439
Total std curriculum	62696 46.65%	70784 52.67%	134392
Total population	78335 48.86%	80528 50.23%	160327

Table 48. **Grade 10 Mathematics: Score Distributions**

	Calibration Sample			All Scored Standard Curriculum Students			All Scored Students		
	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N
All	323.97	39.66	4439	325.09	41.89	132733	318.49	47.89	151331
Male	327.30	40.18	1957	328.81	42.63	61805	320.18	50.02	73402
Female	321.72	38.67	2457	322.16	40.77	70118	317.25	45.50	76926
African American	303.18	38.69	988	296.50	43.75	27910	288.63	50.59	32229
Hispanic	313.27	38.12	877	315.51	40.21	22880	307.16	46.62	28344
White	335.76	35.30	2378	337.93	34.91	76009	333.11	40.14	84102

FCAT 2002 Item Analysis

This section contains traditional item analysis statistics: difficulty (*p*-values) and item-total correlations (discrimination indices). For each of the items on the 16 tests, item difficulties, item-total test correlations, and correlations between the item and reporting categories within each of the subject areas were computed.

Item Difficulty Summary

The following tables summarize the item analysis results by presenting the minimum, 25th-percentile, 50th-percentile, 75th-percentile, and maximum values for the data across all items.

For the dichotomously scored MC and GR items, *p*-values are simply the mean points scored across all students. The *p*-value corresponds 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 the total points possible. The resulting value is a comparable statistic to the *p*-value.

Tables 49 and 50, respectively, illustrate the distribution of *p*-values for all reading and mathematics items. Test *p*-values should show that the items vary in difficulty, but they should not be too high (above 0.90) or too low (near chance, 0.25, for multiple-choice items, or less than 0.10 for the other item types). Tables 49 and 50 reveal that there were no *p*-values less than 0.10, but 5 were greater than 0.90. These items were monitored during IRT processing to ensure appropriate item functioning. More generally, the item *p*-values were dispersed across a sufficient range to establish satisfactory measurement reliability over a wide range of achievement.

Table 49. Proportional* *p*-Value Summary Data for All Reading Items

Grade	Number of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	40	0.204	0.523	0.642	0.705	0.816
4	41	0.289	0.551	0.679	0.777	0.919
5	43	0.174	0.484	0.639	0.759	0.846
6	43	0.245	0.475	0.613	0.717	0.953
7	43	0.295	0.547	0.630	0.712	0.901
8	44	0.408	0.538	0.716	0.804	0.897
9	44	0.289	0.512	0.623	0.714	0.936
10	44	0.215	0.586	0.655	0.775	0.897

*Mean score divided by total score possible

Table 50. Proportional* *p*-Value Summary Data for All Mathematics Items

Grade	Number of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	40	0.273	0.513	0.583	0.706	0.879
4	40	0.192	0.478	0.584	0.714	0.898
5	50	0.322	0.492	0.581	0.680	0.902
6	44	0.221	0.355	0.513	0.625	0.751
7	44	0.136	0.384	0.449	0.575	0.819
8	50	0.251	0.416	0.546	0.651	0.879
9	44	0.151	0.337	0.492	0.624	0.846
10	50	0.264	0.396	0.545	0.624	0.875

*Mean score divided by total score possible

Pearson Item-Total Correlations

Table 51 (on page 27) provides Pearson correlations between individual FCAT reading item scores and Total Reading raw scores. Table 52 (on page 28) provides similar summary data for FCAT mathematics.

Total scores are the sums for all item scores. These values are shown by grade and reporting category. Values are given for selected items across the entire distribution of item correlations to show the range in these values. The maximum and minimum values show the highest and lowest coefficients and the other values represent key point in the distribution. Reporting category scores for these correlations are based on sums of the appropriate points per item—that is, the sum of items according to the reporting categories represented. Distributions for the item-reporting category correlations include only values for items from the matching reporting categories. For the MC and GR items, these values are equivalent to point-biserial correlations between the dichotomously scored item (right or wrong) and total score.

The most important criteria for these correlation statistics are that they not be: (1) negative, (2) near-zero. Items with negative correlations should not be used in IRT processing. Tables 51 and 52 show no negative correlations observed.

Table 51. Summary Data for Reading Item Total Correlations for All Items

Grade	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	Total	40	0.193	0.410	0.466	0.505	0.590
	Word & Text	6	0.551	0.561	0.586	0.621	0.665
	Main Idea	16	0.257	0.436	0.487	0.529	0.559
	Relationships	15	0.339	0.426	0.482	0.531	0.556
	Research Ref.	3	0.684	0.692	0.700	0.707	0.714
4	Total	41	0.256	0.392	0.407	0.460	0.594
	Word & Text	5	0.564	0.578	0.586	0.597	0.642
	Main Idea	18	0.310	0.406	0.430	0.489	0.645
	Relationships	13	0.337	0.414	0.457	0.501	0.634
	Research Ref.	5	0.464	0.518	0.552	0.582	0.625
5	Total	43	0.088	0.320	0.394	0.434	0.516
	Word & Text	7	0.393	0.506	0.529	0.564	0.607
	Main Idea	18	0.121	0.383	0.417	0.469	0.534
	Relationships	11	0.332	0.394	0.459	0.505	0.540
	Research Ref.	7	0.397	0.431	0.455	0.530	0.557
6	Total	43	0.189	0.329	0.420	0.464	0.529
	Word & Text	8	0.428	0.441	0.485	0.521	0.566
	Main Idea	19	0.245	0.341	0.396	0.485	0.529
	Relationships	10	0.394	o.475	0.517	0.534	0.569
	Research Ref.	6	0.378	0.501	0.562	0.584	0.593
7	Total	43	0.060	0.397	0.436	0.487	0.553
	Word & Text	6	0.378	0.432	0.500	0.583	0.592
	Main Idea	16	0.325	0.445	0.476	0.505	0.564
	Relationships	14	0.379	0.457	0.492	0.516	0.537
	Research Ref.	7	0.501	0.513	0.541	0.594	0.626
8	Total	44	0.142	0.334	0.391	0.432	0.609
	Word & Text	8	0.410	0.447	0.490	0.544	0.545
	Main Idea	18	0.213	0.354	0.408	0.436	0.676
	Relationships	9	0.378	0.439	0.501	0.523	0.590
	Research Ref.	9	0.365	0.458	0.478	0.497	0.504
9	Total	44	0.181	0.338	0.383	0.452	0.536
	Word & Text	7	0.385	0.453	0.477	0.516	0.533
	Main Idea	19	0.304	0.377	0.408	0.472	0.551
	Relationships	9	0.391	0.481	0.496	0.546	0.602
	Research Ref.	9	0.358	0.368	0.446	0.478	0.569
10	Total	44	0.086	0.342	0.387	0.455	0.547
	Word & Text	10	0.312	0.460	0.482	0.497	0.566
	Main Idea	13	0.238	0.400	0.418	0.444	0.515
	Relationships	13	0.380	0.437	0.480	0.519	0.604
	Research Ref.	8	0.380	0.420	0.446	0.502	0.676

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

Grade	Reporting Category	No. of Items	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	Total	40	0.193	0.373	0.404	0.477	0.572
	Number	12	0.337	0.454	0.521	0.551	0.616
	Measurement	8	0.341	0.435	0.481	0.524	0.565
	Geometry	7	0.422	0.483	0.492	0.521	0.539
	Algebra	6	0.405	0.541	0.573	0.614	0.659
	Data	7	0.410	0.523	0.542	0.625	0.646
4	Total	40	0.231	0.347	0.419	0.491	0.580
	Number	11	0.452	0.486	0.526	0.579	0.634
	Measurement	8	0.353	0.444	0.492	0.557	0.575
	Geometry	7	0.398	0.439	0.495	0.527	0.532
	Algebra	7	0.446	0.473	0.551	0.558	0.577
	Data	7	0.478	0.517	0.554	0.576	0.597
5	Total	50	0.210	0.348	0.436	0.545	0.648
	Number	12	0.309	0.465	0.532	0.569	0.615
	Measurement	11	0.396	0.459	0.495	0.593	0.614
	Geometry	9	0.327	0.383	0.393	0.449	0.801
	Algebra	10	0.336	0.372	0.524	0.622	0.663
	Data	8	0.353	0.490	0.519	0.600	0.778
6	Total	44	0.128	0.305	0.410	0.468	0.557
	Number	9	0.359	0.396	0.492	0.508	0.592
	Measurement	9	0.381	0.455	0.504	0.528	0.592
	Geometry	9	0.420	0.442	0.469	0.517	0.542
	Algebra	8	0.294	0.422	0.530	0.569	0.607
	Data	9	0.401	0.477	0.504	0.513	0.569
7	Total	44	0.177	0.295	0.386	0.485	0.599
	Number	9	0.336	0.376	0.523	0.543	0.570
	Measurement	9	0.361	0.477	0.526	0.562	0.643
	Geometry	8	0.372	0.429	0.463	0.560	0.621
	Algebra	9	0.377	0.462	0.466	0.520	0.596
	Data	9	0.347	0.397	0.421	0.523	0.586
8	Total	50	0.211	0.380	0.449	0.533	0.733
	Number	11	0.326	0.386	0.464	0.543	0.587
	Measurement	11	0.389	0.504	0.573	0.580	0.602
	Geometry	8	0.410	0.425	0.453	0.597	0.822
	Algebra	11	0.394	0.472	0.520	0.586	0.741
	Data	9	0.316	0.435	0.477	0.524	0.838
9	Total	44	0.244	0.391	0.433	0.506	0.662
	Number	8	0.436	0.488	0.534	0.573	0.595
	Measurement	7	0.444	0.563	0.590	0.652	0.664
	Geometry	11	0.411	0.500	0.550	0.577	0.696
	Algebra	10	0.455	0.476	0.490	0.504	0.607
	Data	8	0.415	0.485	0.521	0.555	0.596
10	Total	50	0.174	0.334	0.402	0.537	0.756
	Number	10	0.386	0.428	0.507	0.617	0.672
	Measurement	9	0.462	0.479	0.496	0.646	0.651
	Geometry	10	0.340	0.413	0.496	0.619	0.841
	Algebra	13	0.277	0.346	0.435	0.500	0.705
	Data	8	0.365	0.368	0.468	0.541	0.824

Biserial Item-Total Correlations

Biserial or point-biserial correlations can be produced for all dichotomously scored FCAT items, that is, for multiple-choice and gridded response items. At least one limitation of point-biserial correlations items is that they are either very easy or very difficult. By contrast, biserial correlations adjust for item distributions. The computed values, however, may exceed +1 or -1. The biserial correlations presented in Tables 53 and 54 (on pages 30 and 31) reveal neither negative correlations nor values exceeding 1.

Table 53. **Summary Data for Biserial Correlations for All Reading Multiple-Choice Items with Performance Tasks in Reporting Category Scores**

	Reporting	No. of		25th		75th	
Grade	Category	Items	Minimum	Percentile	Median	Percentile	Maximum
3	Total	40	0.249	0.523	0.617	0.659	0.800
	Word & Text	6	0.691	0.713	0.784	0.820	0.902
	Main Idea	16	0.333	0.576	0.629	0.690	0.709
	Relationships	15	0.430	0.544	0.637	0.691	0.734
	Research Ref.	3	0.937	0.941	0.945	0.962	0.979
4	Total	37	0.332	0.508	0.546	0.618	0.750
	Word & Text	5	0.754	0.776	0.813	0.834	0.836
	Main Idea	16	0.395	0.556	0.586	0.663	0.792
	Relationships	11	0.438	0.547	0.612	0.632	0.714
	Research Ref.	5	0.645	0.724	0.733	0.745	0.786
5	Total	43	0.131	0.409	0.528	0.591	0.694
	Word & Text	7	0.533	0.662	0.680	0.758	0.807
	Main Idea	18	0.179	0.491	0.576	0.632	0.737
	Relationships	11	0.417	0.511	0.617	0.657	0.713
	Research Ref.	7	0.498	0.567	0.630	0.670	0.709
6	Total	43	0.239	0.446	0.559	0.602	0.703
	Word & Text	8	0.536	0.559	0.623	0.669	0.720
	Main Idea	19	0.309	0.497	0.615	0.654	0.719
	Relationships	10	0.596	0.605	0.667	0.686	0.753
	Research Ref.	6	0.517	0.643	0.721	0.759	0.771
7	Total	43	0.076	0.502	0.571	0.658	0.738
	Word & Text	6	0.478	0.630	0.689	0.756	0.786
	Main Idea	16	0.412	0.562	0.607	0.681	0.760
	Relationships	14	0.480	0.582	0.627	0.708	0.742
	Research Ref.	7	0.627	0.661	0.740	0.800	0.819
8	Total	40	0.178	0.431	0.532	0.600	0.778
	Word & Text	8	0.591	0.643	0.714	0.745	0.754
	Main Idea	16	0.267	0.428	0.542	0.610	0.782
	Relationships	8	0.482	0.600	0.675	0.732	0.747
	Research Ref.	8	0.458	0.602	0.629	0.652	0.689
9	Total	44	0.227	0.432	0.508	0.602	0.711
	Word & Text	7	0.518	0.575	0.634	0.671	0.681
	Main Idea	19	0.381	0.510	0.572	0.640	0.732
	Relationships	9	0.512	0.633	0.696	0.706	0.762
	Research Ref.	9	0.449	0.472	0.560	0.618	0.733
10	Total	39	0.109	0.447	0.512	0.595	0.707
	Word & Text	10	0.443	0.592	0.618	0.655	0.764
	Main Idea	12	0.302	0.509	0.560	0.625	0.701
	Relationships	11	0.503	0.625	0.663	0.709	0.733
	Research Ref.	6	0.488	0.534	0.554	0.620	0.672

Table 54. **Summary Data for Biserial Correlations for Mathematics Multiple-Choice Items with Performance Tasks in Reporting Category Scores**

	Reporting	No. of		25th		75th	
Grade	**Category**	**Items**	**Minimum**	**Percentile**	**Median**	**Percentile**	**Maximum**
3	Total	40	0.269	0.476	0.541	0.626	0.731
	Number	12	0.337	0.454	0.521	0.551	0.616
	Measurement	8	0.474	0.571	0.645	0.672	0.718
	Geometry	7	0.604	0.637	0.659	0.670	0.683
	Algebra	6	0.655	0.681	0.741	0.772	0.841
	Data	7	0.546	0.706	0.726	0.816	0.832
4	Total	40	0.305	0.438	0.552	0.641	0.768
	Number	11	0.566	0.663	0.713	0.816	0.855
	Measurement	8	0.466	0.561	0.628	0.705	0.724
	Geometry	7	0.566	0.612	0.622	0.709	0.752
	Algebra	7	0.559	0.599	0.693	0.715	0.729
	Data	7	0.618	0.673	0.702	0.750	0.759
5	Total	44	0.263	0.466	0.550	0.648	0.797
	Number	11	0.388	0.608	0.659	0.720	0.793
	Measurement	11	0.576	0.617	0.660	0.754	0.783
	Geometry	7	0.484	0.498	0.550	0.595	0.598
	Algebra	9	0.493	0.557	0.641	0.687	0.823
	Data	6	0.454	0.614	0.626	0.690	0.724
6	Total	44	168	0.398	0.535	0.616	0.779
	Number	9	0.450	0.516	0.638	0.681	0.745
	Measurement	9	0.478	0.598	0.648	0.690	0.828
	Geometry	9	0.530	0.554	0.638	0.664	0.687
	Algebra	8	0.384	0.546	0.688	0.734	0.761
	Data	9	0.506	0.598	0.634	0.704	0.731
7	Total	44	0.225	0.384	0.503	0.616	0.765
	Number	9	0.427	0.520	0.677	0.696	0.715
	Measurement	9	0.463	0.602	0.665	0.757	0.840
	Geometry	8	0.468	0.543	0.597	0.703	0.786
	Algebra	9	0.496	0.589	0.630	0.735	0.753
	Data	9	0.441	0.532	0.596	0.665	0.737
8	Total	44	0.265	0.459	0.576	0.675	0.816
	Number	10	0.410	0.503	0.581	0.703	0.777
	Measurement	10	0.507	0.616	0.703	0.747	0.765
	Geometry	6	0.516	0.530	0.560	0.581	0.690
	Algebra	10	0.499	0.634	0.680	0.758	0.787
	Data	8	0.504	0.530	0.604	0.655	0.690
9	Total	44	0.319	0.506	0.575	0.642	0.901
	Number	8	0.560	0.621	0.709	0.728	0.746
	Measurement	7	0.581	0.722	0.741	0.916	0.978
	Geometry	11	0.531	0.667	0.694	0.790	0.947
	Algebra	10	0.572	0.601	0.622	0.637	0.761
	Data	8	0.615	0.644	0.689	0.765	0.780
10	Total	44	0.261	0.433	0.508	0.589	0.824
	Number	9	0.484	0.569	0.657	0.750	0.785
	Measurement	8	0.582	0.600	0.623	0.732	0.842
	Geometry	8	0.435	0.526	0.594	0.638	0.865
	Algebra	12	0.416	0.447	0.547	0.611	0.742
	Data	7	0.460	0.523	0.575	0.643	0.722

IRT Scaling

IRT Framework

FCAT scoring is built on item response theory (IRT). In essence, IRT assumes that item responses by students are the result of underlying achievement levels possessed by those students. IRT algorithms search for item parameters that capture a nonlinear relationship between achievement and the likelihood of each student correctly answering each item. Items that fit the IRT model will exhibit a pattern of lower probabilities of correct responses from low-ability students and 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 difficulty, so the position of the point of inflection is higher or lower (to the right or to the left of the center zero point) along the achievement index. For example, the point of inflection of the curve for the sample item is centered at the mean achievement. Efficient tests are composed of items with characteristics similar to that depicted. By varying item difficulties a test developer is able to discriminate achievement levels along the entire index. 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.

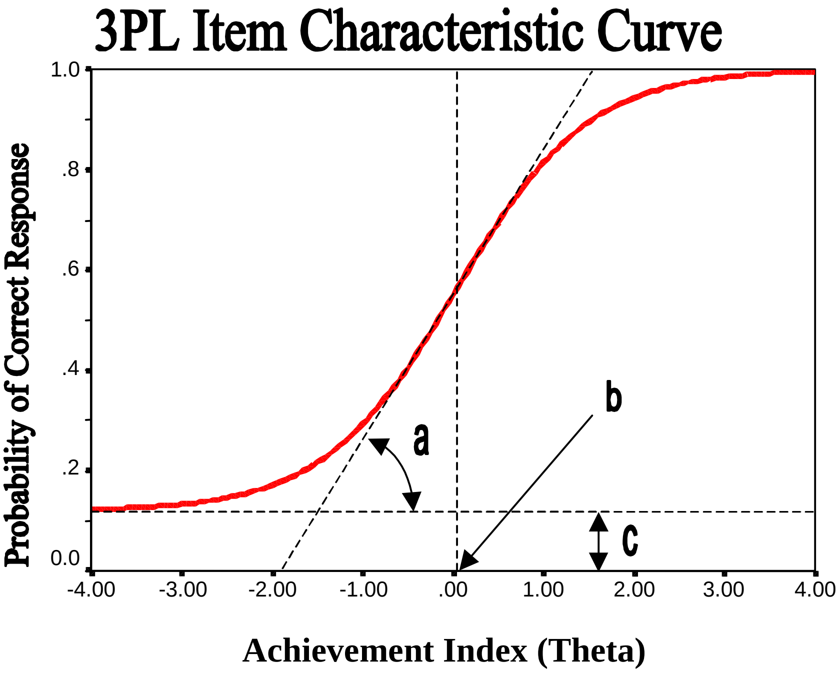

Figure 1. Item characteristic curve based on the three-parameter logistic trace line (A=1, B=0, C=.16).

While IRT modeling of performance tasks (PT) is conceptually similar, these 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 achieve on that task. 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. As shown earlier, 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 2002 Test Construction Specifications* (FDOE, 2001) presents the technical details of these models. Multilog (Thissen, 1991) was used for the IRT analyses.

Gridded-response items received a hybrid treatment. Item parameters were initially computed using a two-parameter logistic model and later converted to the 2PPC model for subsequent processing and maintenance in the item data bank.[2]

IRT item parameters provide the means for assigning achievement scores to individual students. Because the item parameters represent response probabilities, each student's score is assigned as the level most likely to have created that student's observed responses.[3] Use of the sophisticated IRT model is advantageous for continuous testing programs, such as the FCAT, where one must create a stable achievement scoring and reporting system, given the reality that items included on the tests change from one year to the next. In addition, IRT modeling can increase test score reliability.

IRT Results

The 3PL IRT model produces three parameters, A, B, and C, as shown in Figure 1. Distributions of these item parameters are presented in Tables 55 and 56 (on pages 35 and 36) for MC items. The parameters are in the IRT traditional metric,[4] and the achievement index can be interpreted as a standard scale with a score mean of 0 and a 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 the achievement level. The A parameter is similar to the individual item-total score correlation. For reference, the A for the sample curve in Figure 1 is 1.1.

Tables 55 and 56 show that the A parameters are centered at the median range from 0.71 to 0.90 for reading and 0.73 to 1.0 for mathematics. The results show that reading A parameters are slightly lower than mathematics A parameters. This suggests that the mathematics items are more homogeneous than are the reading items.

[2] The 2PL "b" parameter is multiplied by the "a" parameter.

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

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

The B parameter, representing item difficulty, indicates where the item slope is centered along the achievement index. The B parameter is conceptually similar to the *p*-value. For reference, the B in Figure 1 is set at 0.5, indicating that the curve is centered about one-half standard deviation above the population mean. B parameters should be spread across a wide range of achievement to measure accurate student performance at all levels of ability. Because of the way the curve flattens at the ends, an item centered in the middle of the achievement index 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 55 and 56 show that for all grades the B parameters are spread across the scale.

The 3PL C parameter represents the effects of examinees not knowing the answer (guessing) and still getting the item correct. This effect is also called the “pseudo-guessing” parameter. Notice in Figure 1 that the C is at the minimum achievement index score of -3.0. Students knowing nothing about the content of an MC item with four possible responses have a 1 in 4 chance of responding correctly. Typically, C values should be around 0.2. Higher values may signal poorly functioning distractors or some unusual curricular effects. Tables 55 and 56 show that C parameters tend to fall within expected ranges. There are, however, a few items with high C parameters.

For the 2002 test, Grade 10 reading item 17 was not included in any of these tables. This item was removed from the test by FDOE staff and not included in scoring, scaling, or equating for content reasons. Student scores were computed for the Grade 10 reading test without item 17.

Table 55. 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	0.490	0.760	0.905	1.035	1.590
(40)	B	-1.170	-0.550	-0.225	0.335	2.340
	C	0.040	0.120	0.165	0.245	0.430
4	A	0.430	0.670	0.760	0.890	1.220
(37)	B	-1.950	-1.330	-0.580	0.170	1.650
	C	0.070	0.100	0.150	0.220	0.390
5	A	0.110	0.590	0.730	0.910	1.730
(43)	B	-1.890	-0.880	-0.080	0.820	2.280
	C	0.050	0.120	0.180	0.240	0.370
6	A	0.330	0.660	0.830	1.060	1.850
(43)	B	-3.060	-0.730	0.040	0.640	2.250
	C	0.040	0.110	0.180	0.230	0.380
7	A	0.520	0.740	0.930	1.140	1.560
(43)	B	-1.360	-0.510	-0.100	0.430	2.520
	C	0.040	0.140	0.200	0.270	0.570
8	A	0.110	0.475	0.710	0.835	1.260
(40)	B	-2.640	-1.405	-0.715	0.240	1.970
	C	0.050	0.130	0.170	0.215	0.420
9	A	0.190	0.570	0.730	1.055	1.390
(44)	B	-1.920	-0.560	-0.040	0.535	1.940
	C	0.070	0.160	0.180	0.220	0.610
10	A	0.300	0.540	0.730	0.870	1.210
(39)	B	-2.860	-1.360	-0.560	-0.040	2.820
	C	0.060	0.110	0.180	0.240	0.590

Table 56. 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	0.220	0.570	0.815	0.955	1.800
(40)	B	-2.640	-0.810	-0.010	0.380	1.850
	C	0.030	0.080	0.135	0.215	0.270
4	A	0.310	0.590	0.850	0.975	1.700
(40)	B	-1.610	-0.520	-0.025	0.625	1.790
	C	0.040	0.095	0.155	0.245	0.520
5	A	0.430	0.690	0.800	1.040	1.340
(33)	B	-2.040	-0.510	0.070	0.550	1.540
	C	0.060	0.160	0.180	0.240	0.630
6	A	0.240	0.620	0.730	1.070	1.790
(33)	B	-1.100	-0.260	0.470	1.150	2.340
	C	0.070	0.160	0.200	0.230	0.400
7	A	0.210	0.660	0.880	1.100	1.510
(33)	B	-1.310	0.240	0.920	1.310	2.450
	C	0.080	0.150	0.210	0.260	0.430
8	A	0.540	0.800	0.975	1.190	1.600
(30)	B	-1.830	-0.150	0.500	0.940	1.430
	C	0.050	0.160	0.210	0.280	0.570
9	A	0.560	0.810	1.030	1.335	2.610
(28)	B	-1.460	-0.115	0.520	1.050	1.550
	C	0.040	0.145	0.205	0.245	0.520
10	A	0.300	0.595	0.750	1.110	1.600
(28)	B	-1.950	-0.265	0.315	0.735	1.610
	C	0.040	0.115	0.175	0.300	0.420

The parameters for the 2PPC model used to score GR and PT items are conceptually more difficult to translate graphically. Therefore, Table 57 (on page 37) presents only distributions of the A parameters for these items. The A parameters for GR and PT items tend to be higher than those for MC items. However, we are able to make a direct algebraic comparison to the 3 PL model. Because IRT processing tries to fit the same achievement construct to all items, this provides 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 test, and they are indicated as the minimum and maximum values.)

Table 57. The 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)	0.730		0.930		1.510
	ER (1)			0.900		
8	SR (3)	0.500		0.790		1.040
	ER (1)			0.720		
10	SR (4)	0.630	0.830	1.060	1.135	1.180
	ER (1)			0.600		
Mathematics						
5	GR (11)	0.970	1.050	1.430	1.650	2.140
	SR (4)	0.800	0.870	0.945	1.130	1.310
	ER (2)	0.640				0.840
6	GR (11)	1.170	1.190	1.370	1.420	2.050
7	GR (11)	0.690	1.080	1.460	1.540	1.920
8	GR (14)	0.850	1.170	1.505	1.730	2.380
	SR (4)	0.480	0.515	0.915	1.545	1.810
	ER (2)	0.890				1.120
9	GR (16)	0.760	1.075	1.455	2.030	2.820
10	GR (16)	0.430	0.875	1.240	1.670	2.410
	SR (4)	0.890	0.940	1.050	1.125	1.140
	ER (2)	1.160				1.200

Scale Conversion and Test Equating

IRT scaling produces item parameters for an achievement index called the Theta scale with a score mean of 0 and score standard deviation of 1. By converting the Theta scale, FCAT scores can be reported on a new scale from 100 to 500. A transformation is needed for the IRT item parameters in order for this process to produce the appropriate scores.

In addition to the need for student scores to be placed on a comprehensible and stable scale, there is also the need for these current scores to be comparable to scores from past years. Students from 2002 are expected to perform differently (presumably better) than students in previous years. To report scores in 2002 on the FCAT 100-to-500 scale and to make them comparable to scores from past years, the data output from the IRT model must be adjusted through an equating process. This process involved (1) repeating anchor items in the 2002 test that had been used in previous FCAT administrations and (2) applying the Stocking/Lord (1983) procedure. The anchor items and the Stocking/Lord procedure were used to equate 2002 test scores to the test scores originally reported in 1998 (or 2001). This procedure, with different anchor items, has been conducted every year since 1998 (or 2001).

With the completion of the 2002 scaling, the anchor items have two sets of item parameters: (1) new parameters on the mean = 0/standard deviation = 1 scale produced this year and (2) 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 index and then searches for a transformation multiplier and additive constant that can be combined to make the new parameters replicate the 1998 (or 2001) achievement index 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 produced by transformations of new parameters. Since the items are the same, a comparable index should 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. Only Item 17 from the Grade 10 Reading FCAT was dropped from scoring. In addition to Human Resources Research Organization (HumRRO) and Harcourt Educational Measurement (HEM), NCS/Pearson and the FDOE staff also participated in these reviews. In previous years, this procedure was conducted by separately examining each set of corresponding item parameters. This year, HumRRO introduced a computational procedure that produced a metric reflecting differences between the shapes of the item characteristic curves generated by the current year 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 the equating process.

Table 58 reports the number of anchor items used in equating and the transformation constants that were derived to replicate the base-year FCAT scale. The M2 values are called additive constants because they are simply added to the Theta scale score multiplied by the M1 multiplier to determine a score. For example, a Grade 3 Theta score of 1.0 is used as follows: multiply 51.544 times 1.0 and add 1.0 to find the final score of 363.566.

The M2 additive constant projects the change in average scores expected for standard curriculum students. Thus, while an average standard curriculum student would have been expected to score 300 for Grade 4 reading in 1998, the same student in 2002 would be expected to have a score of approximately 311.

Table 58. **Equating Multiplicative and Additive Constants**

Grade	Anchor Item Type and Number	M1 Multiplier	M2 — Additive Constant
Reading			
3	16 MC	50.869	301.673
4	14 MC, 1 SR	48.331	310.979
5	15 MC	48.065	297.026
6	10 MC	51.971	301.227
7	13 MC	48.056	303.642
8	13 MC	44.804	306.160
9	16 MC	47.788	299.295
10	14 MC, 1 SR	41.749	311.763
Mathematics			
3	15 MC	51.544	312.022
4	15 MC	47.809	303.038
5	8 MC, 4 GR	44.246	329.831
6	10 MC, 5 GR	50.399	309.429
7	11 MC, 4 GR	47.452	305.912
8	9 MC, 4 GR	39.410	315.387
9	7 MC, 5 GR	44.817	299.331
10	8 MC, 4 GR	35.830	321.482

IRT Fit Statistics

IRT scaling algorithms attempt to find item parameters (numerical characteristics) that create a match between observed and theoretical response patterns as defined by the selected IRT models. The Q1 statistic (Yen, 1981) may be used as an index for how well theoretical item curves match observed item responses. Q1 uses student achievement scores in combination with estimated item parameters to compute expected performance levels on each item. Differences between expected and observed item performance are then compared at selected intervals across the range of student achievement. Q1 is a ratio involving expected and observed item performance levels and may be interpreted as a chi-square statistic.

Q1 for each item type has a different number of degrees of freedom because of the different numbers of IRT parameters; however, Q1 is not directly comparable across item types. An adjustment (conversion to a z-score, ZQ) is made for different numbers of item parameters and sample sizes to create a more general statistic. The FCAT has a set standard for a minimum ZQ for an item to be labeled as having "acceptable" versus "poor" fit (FDOE, 1998).[5] Complete Q1 results are published in the appendices under a

[5] If ZQ > (1500•4)/sample size, then fit is rated as "poor."

separate cover. Tables 59 and 60 present the distributions of ZQs and Table 61 presents the numbers of poorly fitting items, by item type.

Table 59. **Z Transformation of Q1 Statistic, Summary Data— All Reading Items**

Grade	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	-1.379	-0.297	0.496	2.062	13.251
4	-1.285	0.028	0.613	1.690	20.200
5	-0.953	-0.172	0.719	2.076	6.947
6	-0.885	0.469	1.760	2.904	9.320
7	-0.888	-0.110	0.639	1.218	6.917
8	-1.077	0.596	1.208	2.342	8.383
9	-1.242	-0.208	0.585	2.081	12.006
10	-1.345	0.049	0.893	2.013	8.012

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

Grade	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	-1.004	-0.013	0.551	1.683	5.531
4	-0.815	-0.250	0.569	1.690	6.447
5	-0.797	0.061	0.808	1.630	10.110
6	-1.179	-0.010	0.625	1.399	4.011
7	-1.105	-0.124	0.703	2.085	5.809
8	-1.156	-0.279	1.295	2.297	12.426
9	-1.690	-0.239	0.879	2.330	7.863
10	-1.118	0.169	1.206	3.321	13.927

Table 61. **Number of Items with Low Q1 Statistics—All Items**

	Reading			Mathematics			
Grade	MC	SR	ER	MC	Gridded	SR	ER
3	1/40			0/40			
4	0/37	1/3	0/1	0/40			
5	0/43			0/33	0/11	0/4	0/2
6	0/43			0/33	0/11		
7	0/43			0/33	0/11		
8	0/40	0/3	0/1	0/30	0/14	1/4	0/2
9	0/44			0/28	0/16		
10	0/39	0/4	0/1	0/28	1/16	1/4	0/2

Note: Data shown are the number of items with “poor fit”/total number of items.

The low proportion of poorly fitting items is consistent with the previously reported patterns of strong point-biserials and strong A parameters. The set of items in each FCAT test is thought to converge on a common achievement construct.

Achievement Index Unidimensionality

By fitting all items simultaneously to the same achievement index, the IRT model is operating under the assumption that there is a single construct that underlies the performance of all items. Under this assumption, performance on the items should be related to achievement indices (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 construct is the local dependence assumption of unidimensional IRT, and it suggests a relatively straightforward test for this characteristic called the Q3 statistic (Yen, 1984).

Computation of the Q3 statistic mimics that of the Q1 statistic; that is, expected student performance on each item is calculated by using item parameters and estimated achievement levels. Next, for each student and each item, the difference between expected and observed item performance is calculated. This 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 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) is held constant or explained. 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 40 items, there are 780 Q3 values. Q3 values should all be small. To summarize Q3 data, Tables 62 and 63 present the minimum, 5th percentile, median, 95th percentile, and maximum values for each FCAT grade/subject combination. To add perspective to the meaning of Q3 values, 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.

The data in Tables 62 and 63 indicate that, for all grades and subjects, at least 90 percent of the items are expectedly small with Q3 values between -.06 and .02. These data, coupled with the Q1 data in Tables 59, 60, and 61, indicate that the unidimensional IRT model provides a very reasonable solution for capturing the essence of student achievement as defined by the carefully selected sets of items for each grade and subject.

Table 62. Q3 Statistic, Summary Data—All Reading Items

		Q3 Distribution				
Grade	Average Correlation	Minimum	5th Percentile	Median	95th Percentile	Maximum
3	0.177	-0.117	-0.066	-0.023	0.025	0.148
4	0.153	-0.097	-0.062	-0.024	0.019	0.103
5	0.118	-0.105	-0.060	-0.021	0.019	0.175
6	0.135	-0.093	-0.059	-0.019	0.018	0.141
7	0.166	-0.110	-0.063	-0.020	0.017	0.133
8	0.124	-0.108	-0.061	-0.020	0.021	0.116
9	0.125	-0.118	-0.058	-0.019	0.018	0.152
10	0.132	-0.135	-0.063	-0.020	0.022	0.116

Table 63. Q3 Statistic, Summary Data—All Mathematics Items

		Q3 Distribution				
Grade	Average Correlation	Minimum	5th Percentile	Median	95th Percentile	Maximum
3	0.150	-0.124	-0.070	-0.022	0.022	0.170
4	0.148	-0.139	-0.062	-0.022	0.018	0.368
5	0.178	-0.109	-0.061	-0.017	0.024	0.230
6	0.134	-0.094	-0.058	-0.018	0.020	0.124
7	0.130	-0.104	-0.057	-0.018	0.017	0.125
8	0.188	-0.086	-0.055	-0.016	0.020	0.082
9	0.183	-0.111	-0.056	-0.019	0.015	0.088
10	0.173	-0.121	-0.058	-0.015	0.024	0.124

Item Bias Analyses

FCAT items receive intensive, qualitative reviews by panels of experts before being used in field tests, including reviews for possible gender and/or ethnicity bias (FDOE, 2002, May). In addition, items are re-examined for quantitative evidence of differential performance by various subgroups representing gender/race/ethnicity of examinees, whose achievement levels are assumed to be comparable. Thus, test scores of females are compared with the scores of males, scores of African Americans are compared with the scores of whites, and the scores of Hispanics are compared with those of white students.

The analyses for differential item functioning (DIF) were completed using two methods 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. At any given level of overall achievement, performance on each item should be similar for African Americans or Hispanics when compared with whites. The Mantel (1963) statistic, a version of the Mantel-Haenszel (1959) statistic that accommodates performance task items, is a chi-square procedure that tests the statistical significance (or probability level) of differences in item performance.

Another statistic, the standardized mean difference (SMD), looks at the size of the observed differences and is particularly useful with large sample sizes, such as those found in FCAT calibrations. A statistically significant difference – on examination by educators and policymakers – may not be deemed large enough to cause concern from a practical perspective. To assist with this analysis, an SMD rating system was put into place (FDOE, 1998), grouping each item into one of seven categories according to its demonstrated differential functioning for or against any of the identified comparison groups.

Tables 64 and 65 present the distributions 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 was not surprising.

Table 64. **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	39		1				
4	37	3			1		
5	40	3					
6	42	1					
7	41	2					
8	38	4		1	1		
9	41	2	1				
10	37	3	3			1	

Table 65. **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	40						
4	39	1					
5	41	5	2				
6	43	1					
7	41	2	1				
8	45	3					
9	44						
10	42	3	2	1			

Test Reliability and Standard Error of Measurement

The previous discussions have indicated that FCAT items on each test reflect the presence of a common achievement index. Additional investigations of reliability and conditional standard errors of measurement and reliability are presented in this section.

Test reliability refers to the consistency of measurement. This concept holds that a test score results from some theoretical level of achievement, plus measurement error. For a population of students, reliability is a ratio between variations in theoretical achievement and variations in observed test scores. The less that 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. 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 2 and 3 on the following pages, are methods for depicting test reliability. These 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 specific level of achievement will have observed test scores within 1 SEM of their theoretical scores. For example, from Figure 2, the Grade 4 reading SEM plots reveal that students whose theoretical achievement level is 200 will have a SEM of approximately 25. That means that approximately two-thirds of these students will have test scores between 175 and 225. The remaining one-third of these students will have test scores more than 25 points away from 200. As expected, the SEM is larger at the tails of the achievement index distribution and smaller in the center. Most students, however, score near the center of the achievement index. Cut-scores, used to determine student performance categories (Achievement Levels 1-5), are located near the center of these indices (see Tables 66 and 67).

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 makes reliability data consistent across grades and subjects and avoids confounding any differences in calibration samples. In addition, these estimates are consistent with the reporting of FCAT scores; they characterize test results for all students who receive scores.

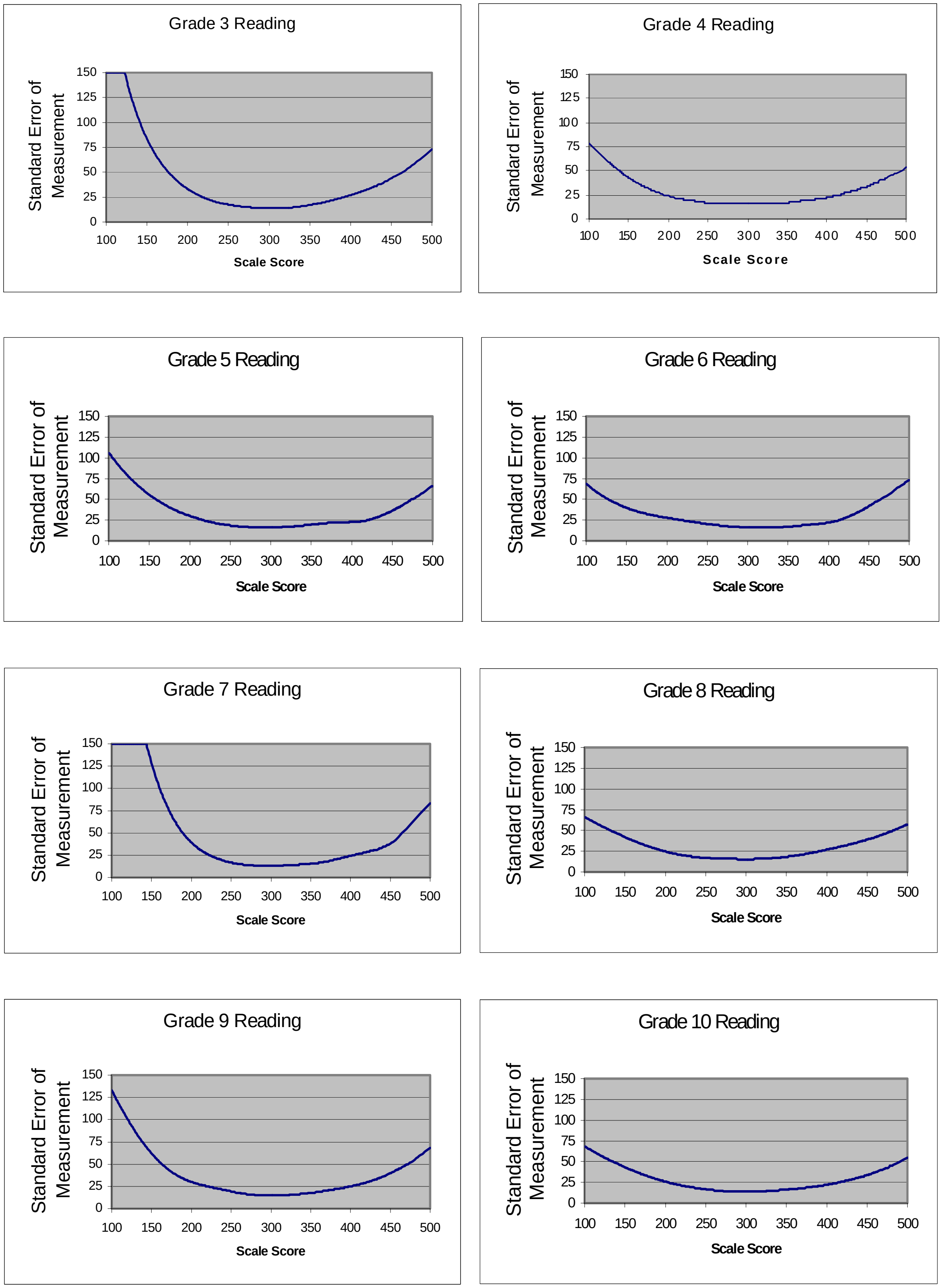

Figure 2. Standard error of measurement plots for FCAT Reading.

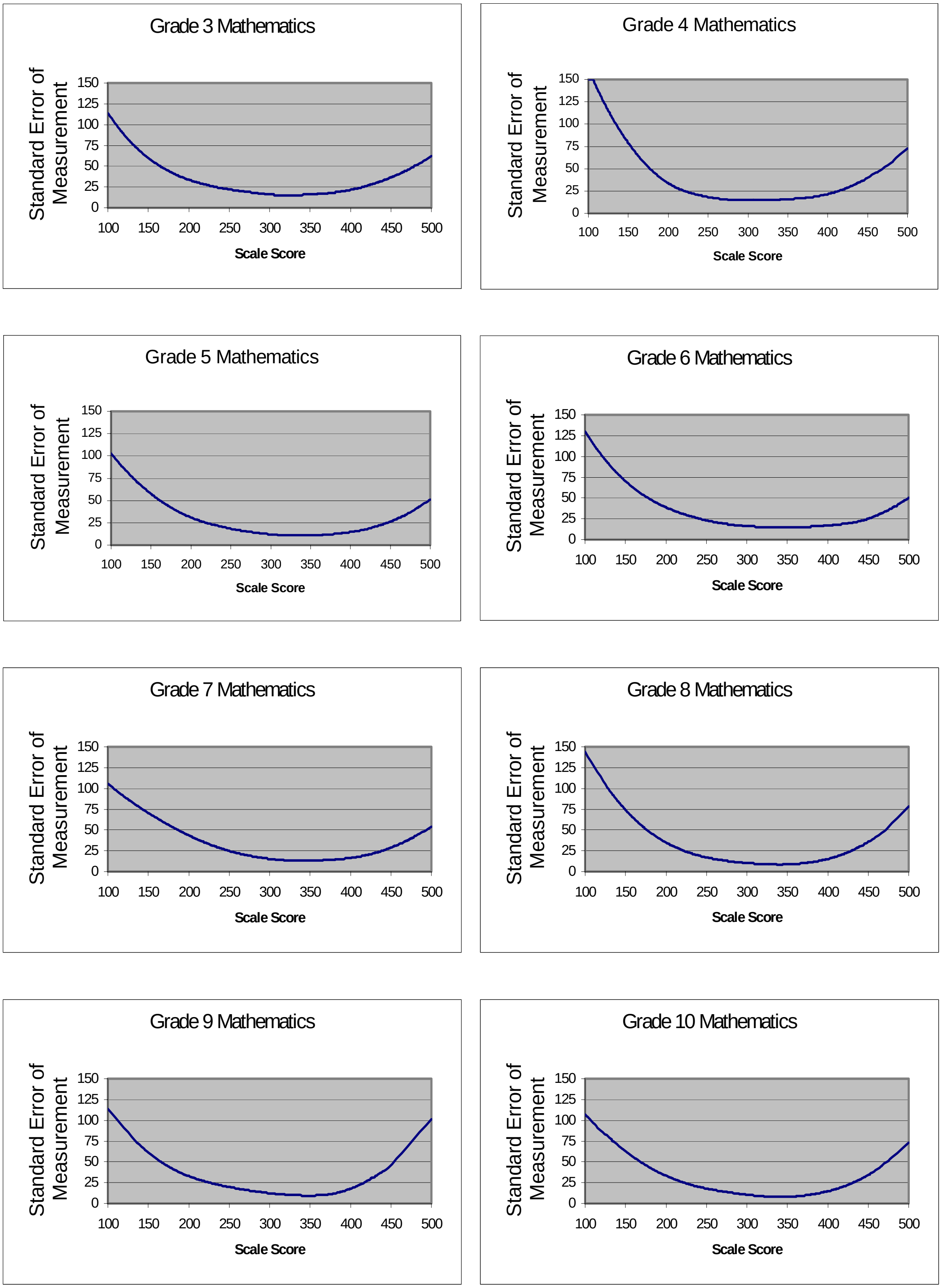

Figure 3. Standard error of measurement plots for FCAT Mathematics.

Table 66. Reading SEM at Cut-scores for the Achievement Level Categories 1-5 (Student scores at or above cut are in higher category.)

Grade	Cut-scores	SEM
3	259	16
	284	14
	332	15
	394	26
4	275	16
	299	16
	339	17
	386	20
5	256	18
	286	16
	331	18
	384	22
6	265	19
	296	16
	339	16
	387	20
7	267	14
	300	13
	344	15
	389	22
8	271	16
	310	16
	350	18
	394	25
9	285	15
	322	15
	354	18
	382	22
10	287	14
	327	15
	355	17
	372	18
PASS (10 only)	300	14

Table 67. Mathematics SEM at Cut-scores for Achievement Level Categories 1-5 (Student scores at or above the cut are in higher category.)

Grade	Cut-scores	SEM
3	253	22
	294	16
	346	16
	398	21
4	260	17
	298	15
	347	16
	394	20
5	288	13
	326	11
	355	11
	395	14
6	283	18
	315	15
	354	15
	391	16
7	275	19
	306	15
	344	13
	379	14
8	280	12
	310	9
	347	9
	371	10
9	261	18
	296	13
	332	10
	367	10
10	287	12
	315	9
	340	8
	375	10
PASS (10 only)	300	10

Viewing both the reliability and SEM data is important. The marginal reliabilities (provided in Table 68) indicate that FCAT scores have reliabilities similar to those of other standardized and statewide tests. Individual test scores will vary more toward the upper and lower portions of the distribution. Rogosa (1994, 2000) examines 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 should also be interpreted using the SEM because of the random fluctuations that can occur in individual score reliability. The SEM curves indicate that individuals near the center of the distribution will have test scores that vary by chance fewer than 20 points (that is, plus or minus the lowest SEM). Tables 66 and 67 indicate that the cut-scores, used to determine achievement level categories, nearly all fall within that range, concluding that the FCAT is a reliable indicator of student achievement.

Table 68 also shows traditional reliability statistics based on Cronbach's alpha. These reliability estimates are based on raw scores only and have been reported for the total set of items and for all items that comprise each of the separate reporting categories. Lower reliabilities reflect the reality that fewer numbers of items are associated with each of the reporting categories. The numbers of items are in parentheses.

Table 68. **IRT Marginal Reliabilities and Cronbach's Alpha for Reading**

	IRT	**Cronbach's Alpha**					
Reading	*Marginal* r_{ii}	*Total*	*Word and Phrases*	*Main ideas*	*Comparisons*	*Reference Research*	
Grade 3	0.90	0.908	0.667 (6)	0.779 (16)	0.776 (15)	0.551 (3)	
4	0.89	0.903	0.613 (5)	0.807 (18)	0.745 (13)	0.471 (5)	
5	0.87	0.872	0.597 (7)	0.740 (18)	0.647 (11)	0.445 (7)	
6	0.89	0.887	0.568 (8)	0.775 (19)	0.692 (10)	0.498 (6)	
7	0.90	0.909	0.451 (6)	0.797 (16)	0.772 (14)	0.668 (7)	
8	0.87	0.892	0.647 (8)	0.754 (18)	0.663 (9)	0.600 (9)	
9	0.88	0.875	0.467 (7)	0.774 (19)	0.675 (9)	0.478 (9)	
10	0.89	0.885	0.632 (10)	0.682 (13)	0.729 (13)	0.515 (8)	
Mathematics	*IRT Marginal* r_{ii}	*Total*	*Number Sense, Concepts, Operations*	*Measure-ment*	*Geometry and Spatial Sense*	*Algebraic Thinking*	*Data Analysis/ Probability*
Grade 3	0.89	0.892	0.759 (12)	0.540 (8)	0.527 (7)	0.608 (6)	0.668 (7)
4	0.89	0.890	0.784 (11)	0.574 (8)	0.468 (7)	0.578 (7)	0.648 (7)
5	0.93	0.923	0.749 (12)	0.756 (11)	0.631 (9)	0.707 (10)	0.699 (8)
6	0.89	0.880	0.579 (9)	0.582 (9)	0.621 (9)	0.603 (8)	0.626 (9)
7	0.88	0.882	0.592 (9)	0.667 (9)	0.577 (8)	0.653 (9)	0.550 (9)
8	0.93	0.929	0.676 (11)	0.783 (11)	0.658 (8)	0.785 (11)	0.665 (9)
9	0.91	0.914	0.644 (8)	0.651 (7)	0.766 (11)	0.702 (10)	0.667 (8)
10	0.93	0.923	0.731 (10)	0.722 (9)	0.718 (10)	0.705 (13)	0.607 (8)

Intercorrelations Between Reporting Categories and Scale Scores

Tables 69 through 84 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 are affected by differences in scale reliabilities (see Table 68) which are related to the number of items in each category. For example, in Table 69, observed correlations with the Research and Reference reporting category at Grade 3 are lower than other correlations because there are only three items. To illustrate this effect, for Grade 3 reading only, "true" correlations have been estimated from the observed correlations and scale reliabilities using the correction due to attenuation. These estimated "true" correlations are in brackets. The estimates in Table 69 indicate that if the number of items on the test in each category were increased, higher observed reliabilities would result.

Tables for Reading

Table 69. Grade 3 Reading Reporting Category and Scale Score Intercorrelations (Number of items in parentheses) N = 4645

	Total Raw Score (40)	Words & Phrases (6)	Main Ideas (16)	Comparisons (15)	Ref. Research (3)
Scale Score	0.958	0.795	0.890	0.864	0.692
Total Raw Score		0.819	0.927	0.919	0.690
Word & Text			0.682 [0.87]	0.678 [0.87]	0.535 [0.79]
Main Ideas				0.758 [1.0]	0.591 [0.92]
Relationships					0.557 [0.82]

Note: Brackets contain correlations corrected for attenuation.

Table 70. Grade 4 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N= 4811

	Total Raw Score (41)	Words & Phrases (5)	Main Ideas (18)	Comparisons (13)	Ref. Research (5)
Scale Score	0.980	0.742	0.914	0.875	0.664
Total Raw Score		0.752	0.931	0.897	0.677
Word & Text			0.619	0.615	0.447
Main Ideas				0.735	0.543
Relationships					0.520

Table 71. **Grade 5 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N=4570**

	Total Raw Score (43)	Words & Phrases (7)	Main Ideas (18)	Comparisons (11)	Ref. Research (7)
Scale Score	0.967	0.745	0.886	0.813	0.663
Total Raw Score		0.764	0.903	0.845	0.715
Word & Text			0.580	0.558	0.459
Main Ideas				0.661	0.529
Relationships					0.497

Table 72. **Grade 6 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N=4841**

	Total Raw Score (43)	Words & Phrases (8)	Main Ideas (19)	Comparisons (10)	Ref. Research (6)
Scale Score	0.967	0.740	0.889	0.837	0.744
Total Raw Score		0.792	0.910	0.856	0.772
Word & Text			0.611	0.579	0.540
Main Ideas				0.678	0.609
Relationships					0.594

Table 73. **Grade 7 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N=5243**

	Total Raw Score (43)	Words & Phrases (6)	Main Ideas (16)	Comparisons (14)	Ref. Research (7)
Scale Score	0.953	0.646	0.884	0.865	0.788
Total Raw Score		0.693	0.925	0.906	0.824
Word & Text			0.561	0.540	0.497
Main Ideas				0.750	0.689
Relationships					0.677

Table 74. **Grade 8 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N=4610**

	Total Raw Score (44)	Words & Phrases (8)	Main Ideas (18)	Comparisons (9)	Ref. Research (9)
Scale Score	0.975	0.740	0.879	0.834	0.763
Total Raw Score		0.754	0.914	0.840	0.783
Word & Text			0.590	0.561	0.497
Main Ideas				0.666	0.597
Relationships					0.585

Table 75. **Grade 9 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N=5324**

	Total Raw Score (44)	Words & Phrases (7)	Main Ideas (19)	Comparisons (9)	Ref. Research (9)
Scale Score	0.966	0.702	0.900	0.824	0.704
Total Raw Score		0.740	0.921	0.842	0.752
Word & Text			0.590	0.548	0.441
Main Ideas				0.688	0.577
Relationships					0.531

Table 76. **Grade 10 Reading Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N=4465**

	Total Raw Score (41)	Words & Phrases (6)	Main Ideas (15)	Comparisons (9)	Ref. Research (11)
Scale Score	0.974	0.792	0.814	0.891	0.776
Total Raw Score		0.822	0.838	0.895	0.811
Word & Text			0.583	0.652	0.558
Main Ideas				0.668	0.571
Relationships					0.642

Tables for Mathematics

Table 77. Grade 3 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N=4641

	Total Raw Score (40)	Number (12)	Measure-ment (8)	Geometry (7)	Algebra (6)	Data (7)
Scale Score	0.967	0.865	0.740	0.667	0.748	0.801
Total Raw Score		0.889	0.777	0.713	0.765	0.812
Number			0.590	0.505	0.634	0.639
Measurement				0.475	0.502	0.555
Geometry					0.425	0.522
Algebra						0.530

Table 78. Grade 4 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N=4655

	Total Raw Score (40)	Number (11)	Measure-ment (8)	Geometry (7)	Algebra (7)	Data (7)
Scale Score	0.957	0.883	0.724	0.636	0.728	0.777
Total Raw Score		0.880	0.786	0.674	0.782	0.814
Number			0.603	0.482	0.607	0.656
Measurement				0.426	0.524	0.540
Geometry					0.426	0.478
Algebra						0.548

Table 79. Grade 5 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N=4743

	Total Raw Score (50)	Number (12)	Measure-ment (11)	Geometry (9)	Algebra (10)	Data (8)
Scale Score	0.969	0.855	0.855	0.793	0.831	0.837
Total Raw Score		0.884	0.871	0.831	0.855	0.862
Number			0.719	0.643	0.708	0.705
Measurement				0.658	0.695	0.691
Geometry					0.635	0.630
Algebra						0.679

Table 80. **Grade 6 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N=4843**

	Total Raw Score (44)	Number (9)	Measure-ment (9)	Geometry (9)	Algebra (8)	Data (9)
Scale Score	0.946	0.770	0.734	0.768	0.770	0.787
Total Raw Score		0.811	0.801	0.792	0.811	0.833
Number			0.545	0.536	0.601	0.607
Measurement				0.554	0.567	0.576
Geometry					0.540	0.564
Algebra						0.602

Table 81. **Grade 7 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N=5250**

	Total Raw Score (44)	Number (9)	Measure-ment (9)	Geometry (8)	Algebra (9)	Data (9)
Scale Score	0.922	0.717	0.750	0.701	0.780	0.734
Total Raw Score		0.788	0.834	0.787	0.810	0.773
Number			0.566	0.525	0.550	0.501
Measurement				0.582	0.592	0.557
Geometry					0.546	0.515
Algebra						0.536

Table 82. **Grade 8 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N=4639**

	Total Raw Score (50)	Number (11)	Measure-ment (11)	Geometry (8)	Algebra (11)	Data (9)
Scale Score	0.938	0.774	0.850	0.795	0.862	0.817
Total Raw Score		0.847	0.886	0.870	0.893	0.876
Number			0.685	0.662	0.700	0.685
Measurement				0.706	0.753	0.714
Geometry					0.710	0.707
Algebra						0.740

Table 83. **Grade 9 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N=5313**

	Total Raw Score (44)	Number (8)	Measure-ment (7)	Geometry (11)	Algebra (10)	Data (8)
Scale Score	0.921	0.777	0.692	0.784	0.798	0.803
Total Raw Score		0.825	0.808	0.889	0.864	0.779
Number			0.573	0.645	0.656	0.594
Measurement				0.717	0.599	0.522
Geometry					0.685	0.589
Algebra						0.619

Table 84. **Grade 10 Mathematics Reporting Category and Scale Score Intercorrelations. (Number of items in parentheses) N=4439**

	Total Raw Score (44)	Number (8)	Measure-ment (7)	Geometry (11)	Algebra (10)	Data (8)
Scale Score	0.953	0.838	0.813	0.852	0.815	0.803
Total Raw Score		0.873	0.863	0.896	0.870	0.845
Number			0.696	0.720	0.703	0.690
Measurement				0.729	0.691	0.666
Geometry					0.703	0.693
Algebra						0.669

Student Classification Accuracy and Consistency

Based on their FCAT scale scores, students are classified into one of five performance levels. Evaluation of the reliability of classification decisions involved estimation of the probabilities associated with correct and consistent placements by level. The procedures used were from Livingston and Lewis (1995) and Lee, Hanson, and Brennan (2000). A brief description of these procedures 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." Additionally, Livingston and Lewis indicate that 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' true scores)." 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 FCAT Grade 3 Reading is given in Table 85. It shows the proportions of students who were classified into each performance category by the actual observed scores and by estimated true scores.

Table 85. FCAT 2002 Reading Grade 3 "True" Scores Vs. Observed Scores Cross-Tabulation (Accuracy Table)

Estimate of True Score	**Observed Score**					
	LEVEL 1	**LEVEL 2**	**LEVEL 3**	**LEVEL 4**	**LEVEL 5**	**Total**
LEVEL 1	.258	.029	.004	.000	.000	.291
LEVEL 2	.031	.053	.033	.000	.000	.116
LEVEL 3	.006	.040	.177	.044	.000	.267
LEVEL 4	.000	.000	.051	.245	.028	.325
LEVEL 5	.000	.000	.000	.000	.000	.000
Total	.294	.122	.265	.290	.028	1.00

Note: Column and row totals are computed from non-rounded values. Shading is used to demonstrate the computation of the overall accuracy index (explained in further text).

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 FCAT Grade 3 Reading is given in Table 86. 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.

Table 86. FCAT 2002 Reading Grade 3 "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	.245	.035	.013	.000	.000	.294
LEVEL 2	.035	.042	.042	.003	.000	.122
LEVEL 3	.013	.042	.146	.064	.001	.265
LEVEL 4	.000	.003	.064	.202	.021	.290
LEVEL 5	.000	.000	.001	.021	.006	.028
Total	.294	.122	.265	.290	.028	1.00

Note: Column and row totals are computed from non-rounded values. Shading is used to demonstrate the computation of the consistency index conditional on level (explained in further text).

Accuracy and Consistency Indices

There are three types of accuracy and consistency indices that can be generated from these tables: *overall*, *conditional on level*, and *by cut-score*. In order to facilitate their interpretation, a brief outline of computational procedures used to derive accuracy indices will be presented below.

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 the shaded area in Table 85. Actually, overall accuracy is a proportion (or percentage) of correct classifications across all levels. In the example offered, the overall accuracy index for the FCAT Grade 3 reading test equals 0.733.

The *overall consistency* index is computed as the sum of the diagonal cells in the consistency table. Using the data from Table 86, it can be determined that the overall consistency index for the FCAT Grade 3 reading test equals 0.641. 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). *Kappa* assesses the proportion of consistent classifications between two different test forms after removing the proportion of consistent classifications that would be expected by chance alone. Using the data from Table 86, Cohen's κ for FCAT Grade 3 Reading equals 0.517. Compared to the previously described overall consistency estimate, Cohen's κ has a 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 86, the row LEVEL 4 is outlined and corresponding cells are shaded. The ratio between 0.20182 (non-rounded proportion of correct classifications) and 0.29002 (total non-rounded proportion of students classified into the LEVEL 4) yields 0.696, which represents the index of consistency of classification for FCAT Grade 3 Reading that is conditional on LEVEL 4.

Accuracy conditional on level is computed in a similar manner. The only difference is that in the consistency table both row and column marginal sums are the same; whereas, in the accuracy table, the sum that is based on estimated status is used as a total for computing accuracy conditional on level. For example, in Table 87, the non-rounded proportion of agreement between estimated score status and observed score status at LEVEL 1 is 0.25786; whereas, the total non-rounded proportion of students with true score status at this level is 0.29094. The accuracy conditional on level is equal to the ratio between these two proportions, which yields 0.886. This value indicates that 88.6 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 reading test.

To evaluate decisions at specific cut-scores the joint distribution of all the performance levels are collapsed into a dichotomized distribution around that specific cut-score. For example, the dichotomization at the cut-score that separates LEVEL 1 through LEVEL 3 (combined) from LEVEL 4 and LEVEL 5 (combined) for FCAT Reading Grade 3 is depicted in Table 87. The proportion of correct classifications below that particular cut-score is equal to the sum of the cells in the upper left shaded area (0.6299); and the proportion of correct classifications above this particular cut-score is equal to sum of the cells in the lower right shaded area (0.2736).

Table 87. FCAT 2002 Reading Grade 3 "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	0.258	0.029	0.004	0.000	0.000	0.291
LEVEL 2	0.031	0.053	0.033	0.000	0.000	0.116
LEVEL 3	0.006	0.040	0.177	0.044	0.000	0.267
LEVEL 4	0.000	0.000	0.051	0.245	0.028	0.325
LEVEL 5	0.000	0.000	0.000	0.000	0.000	0.000
Total	0.294	0.122	0.265	0.290	0.028	1.00

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

The *accuracy index* at cut-score is computed as the sum of the proportions of correct classifications around this selected cut-score. In the example from Table 87, the sum of both shaded areas equals 0.903, which means that 90.3 percent of students were correctly classified either above or below the particular cut-score. The sum of the proportions in the upper right non-shaded area (0.045) indicates false positives (i.e., 4.5 percent of students classified above the cut-score by their observed score, but falling below the cut-score with their “true” score). The sum of the lower left non-shaded area (0.052) is the proportion of false negatives (i.e., 5.2 percent of students with an observed level below the cut-score whose “true” level is above the cut-score).

The *consistency* at a specific cut-score is obtained in a similar way. For example, by dichotomizing the distribution in Table 86 at the cut-score between ‘LEVEL 1’ and all other levels combined, it can be determined that the proportion of correct classifications around that cut-score equals 0.902. This means that 90.2 percent of students would be placed by an alternate form (if they had taken one) in the same two categories (LEVEL 1 vs. LEVELS 2 through 5 combined) as they were classified by the actual form taken.

Accuracy and Consistency Results for FCAT 2002

In this section, summary tables for all grades and subject areas are presented to show overall accuracy and consistency indices, accuracy indices at specific levels, and accuracy and consistency indices at cut-scores.

The overall indices of accuracy and consistency of classification for FCAT 2002 tests are presented in Table 88.

Table 88. Estimates of Accuracy and Consistency of Performance-Level Classifications in FCAT 2002 Tests

Grade	Subject	Accuracy	Consistency	Kappa (κ)
3	Reading	.733	.641	.517
	Mathematics	.689	.586	.461
4	Reading	.706	.613	.476
	Mathematics	.704	.603	.469
5	Reading	.686	.590	.446
	Mathematics	.720	.619	.499
6	Reading	.685	.590	.450
	Mathematics	.646	.556	.403
7	Reading	.718	.620	.494
	Mathematics	.630	.538	.385
8	Reading	.670	.579	.434
	Mathematics	.668	.587	.434
9	Reading	.670	.589	.410
	Mathematics	.684	.584	.451
10	Reading	.658	.551	.390
	Mathematics	.734	.630	.497

Table 88 indicates overall accuracy indices range between 0.630 and 0.734; overall consistency indices range between 0.538 and 0.641, and κ coefficients fall in a range between 0.385 and 0.517. Compared to last year's values (*FCAT Technical Report 2001*), accuracy and consistency indices for most of the FCAT grade-level and subject area tests are at about the same level. Average accuracy across all grades and subject combinations in the year 2001 equaled 0.703, and this year it is 0.688. Average consistency last year was 0.607, and this year it equals 0.592. Average κ. value last year was 0.487, and this year it is 0.451.

In addition to overall ratings of decision accuracy, the levels of agreement at each performance level are also of interest. Table 89 displays the probabilities of students being classified in a particular performance level, given that their "true status" was the same category (accuracy conditional on level). 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. This is due to the fact that this performance level covers a wider range of the measured construct than the intermediate levels, and misclassification can occur only in one direction. However, the accuracy at the highest performance level could not be computed in most of the tests because there were no "true" scores classified into this category. It should be noted that the percentage of students whose observed scores are classified in this performance level is relatively low (below 5 percent in all instances except Mathematics Grade 3), which makes indices at that level unreliable and impossible to estimate. By contrast, it is possible to estimate the accuracy of decisions at the cutscore between LEVEL 4 and LEVEL 5, and this estimate can be high (see Table 90).

Table 89. Estimated Probability of Being Classified at a Proficiency Level Given that the "True" Status is that Level (Accuracy Conditional on Level)

Grade	Subject	Level 1	Level 2	Level 3	Level 4	Level 5
3	Reading	.886	.453	.663	.754	*
	Mathematics	.869	.589	.664	.637	.591
4	Reading	.909	.375	.630	.688	*
	Mathematics	.881	.569	.643	.684	*
5	Reading	.899	.422	.587	.701	*
	Mathematics	.911	.635	.552	.723	*
6	Reading	.903	.416	.589	.667	*
	Mathematics	.897	.499	.489	.573	*
7	Reading	.900	.537	.654	.680	*
	Mathematics	.910	.433	.493	.530	*
8	Reading	.892	.614	.559	.568	*
	Mathematics	.928	.561	.673	.476	*
9	Reading	.907	.524	.476	.433	*
	Mathematics	.906	.601	.597	.581	*
10	Reading	.910	.588	.525	*	*
	Mathematics	.909	.619	.568	.764	*

* No accuracy estimates were calculated at 'LEVEL 5' because the number of estimated "true" scores in this cell 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-score, which makes it desirable to know the accuracy and consistency of dichotomous decisions made around that specific score. Reporting in a percent at-or-above the cut (PAC) requires a judgment about whether the student score is below or at-or-above a particular cut-score. Table 90 presents the accuracy and consistency information for these dichotomous categorizations.

Table 90. Accuracy and Consistency of Dichotomous Categorizations

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	.931	.917	.903	.972	.902	.882	.865	.956
	Mathematics	.933	.908	.896	.948	.906	.870	.855	.928
4	Reading	.928	.914	.887	.961	.898	.878	.842	.938
	Mathematics	.930	.904	.879	.986	.901	.864	.834	.974
5	Reading	.918	.898	.870	.984	.885	.856	.819	.970
	Mathematics	.943	.917	.893	.962	.919	.883	.849	.938
6	Reading	.919	.903	.885	.962	.887	.864	.839	.939
	Mathematics	.919	.883	.843	.978	.885	.834	.794	.959
7	Reading	.930	.912	.899	.972	.901	.876	.858	.954
	Mathematics	.925	.893	.829	.953	.893	.846	.775	.919
8	Reading	.933	.891	.845	.992	.905	.844	.802	.984
	Mathematics	.956	.929	.805	.969	.937	.898	.763	.943
9	Reading	.897	.882	.900	.966	.856	.834	.864	.946
	Mathematics	.942	.913	.857	.964	.918	.877	.806	.937
10	Reading	.924	.859	.869	.958	.892	.800	.825	.926
	Mathematics	.955	.929	.883	.962	.937	.899	.833	.934

The data in Table 90 reveal that the level of agreement in terms of both accuracy and consistency for these dichotomous categorizations is above 80 percent in all but three cases (Grades 6, 7, and 8 Mathematics at cut-score 1+2+3 vs. 4+5). In relatively few instances does the level of agreement for decision accuracy fall below 90 percent. Although the rates of agreement for decision consistency are slightly lower, in no case does the rate of agreement fall below 76 percent. In general, high rates of accuracy and consistency support cut decisions.

The inference that high accuracy percentages support cut decisions is even further supported by data on the percentages of false positives and false negatives derived from the dichotomized “true” vs. “observed” status categorizations (see Table 91). On average, only 4.73 percent of students were classified in a lower or higher level than their “true” level across all grades and subjects. The range of false positives and false negatives is from 0.00 to 0.08, indicating that not more than 7 percent of students were classified differently from the level required to meet each cut-score standard.

Table 91. 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	.033	.037	.045	.028	.036	.046	.052	*
	Mathematics	.030	.041	.048	.047	.037	.051	.056	.005
4	Reading	.028	.042	.049	.039	.044	.044	.064	*
	Mathematics	.033	.045	.063	.014	.037	.052	.057	*
5	Reading	.030	.045	.065	.016	.052	.057	.065	*
	Mathematics	.024	.039	.054	.038	.033	.044	.053	*
6	Reading	.031	.048	.059	.038	.050	.049	.056	*
	Mathematics	.035	.042	.086	.022	.046	.075	.072	*
7	Reading	.031	.041	.047	.028	.039	.047	.054	*
	Mathematics	.029	.043	.083	.047	.046	.064	.088	*
8	Reading	.032	.039	.081	.008	.035	.070	.074	*
	Mathematics	.018	.031	.073	.031	.026	.040	.122	*
9	Reading	.040	.055	.063	.034	.063	.064	.038	*
	Mathematics	.028	.036	.056	.036	.030	.051	.087	*
10	Reading	.027	.058	.131	.042	.049	.083	*	*
	Mathematics	.019	.030	.050	.038	.025	.041	.067	*

* False negatives could not be estimated at 1+2+3+4 vs. 5 cutpoint because the number of estimated "true" scores in the cell was zero.

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 FCAT. Students hoping to receive a regular diploma are required, among other things, to achieve a score of at least 287 on the Grade 10 FCAT reading test and at least 295 on the Grade 10 FCAT mathematics test. In sum, three situations are possible:

1. Students whose observed performance is accurately reflected in terms of the standard and 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, but who achieve scores above the standard ("False Positives").

3. Students whose "true" ability is above the standard, but their attained scores indicate that they have not yet met the standard.

Examination of the FCAT results for the Grade 10 Reading and Mathematics, in terms of the high school standards, reveals the following:

The FCAT Grade 10 reading test has a fail-pass threshold between performance levels 1 and 2. The accuracy of fail-pass decisions for this test is equal to the accuracy of dichotomous categorizations between LEVEL 1 and LEVELS 2, 3, 4, and 5 combined. Table 90 indicates that 92 percent of the students are correctly classified into either the “pass” or “fail” category (situation 1) based on their performance on the Grade 10 Reading FCAT. Table 91 shows that 3 percent of students passed although their “true” ability is below the standard (situation 2); and 5 percent failed although their “true” ability is above the standard (situation 3).

A separate analysis was performed to estimate the accuracy of fail-pass decisions for the FCAT Grade 10 mathematics test: the threshold score for fail-pass decisions fell in the middle of performance LEVEL 2. The analysis shows that 93 percent of students were classified correctly into either a “pass” or “fail” category (situation 1) based on their performance, whereas three percent of students were “false positive” classifications (situation 2), and four percent of students were “false negative” classifications (situation 3).

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 http://www.firn.edu/doe/curric/prek12/frame2.htm

Florida Department of Education. (1998). <u>Technical Report: Florida Comprehensive Assessment Test</u> *(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.

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. Monterey, 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.