READING AND MATHEMATICS

Technical Report
For 2006 FCAT Test Administrations

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

Under subcontract to and in cooperation with
Harcourt Assessment, Inc.
San Antonio, TX

San Antonio, TX

January 2007

Table of Contents

(Appendices may be retrieved through the Florida Department of Education, Office of Assessment.)

List of Tables

List of Figures

INTRODUCTION AND OVERVIEW

This report presents technical information on the measurement characteristics of the Reading and Mathematics assessments included in the Florida Comprehensive Assessment Test® (FCAT) for Spring 2006. These characteristics provide an indication of the current quality of FCAT assessments in these two content areas.

Although this report is technical in nature, it is written for an audience familiar with basic testing concepts. Summary data is provided in the main body of the report, while more detailed data are found in the Appendices. More detail on the FCAT and information about test construction, scoring, and reporting are provided in the *FCAT Handbook—A Resource for Educators* (http://fcat.fldoe.org/handbk/fcathandbook.asp).

Description of the FCAT

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

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

FCAT items have various roles. In 2006, there were core items, anchor items, and field-test items. Core items are designed to assess on-grade SSS for each grade and are items for which students receive their scores. Core items are released to the public in some administration years

[1] Grades/subjects that include performance tasks are sometimes referred to as "PT Grades."

as determined by FDOE. In addition to core items on the FCAT, each test also includes anchor items and/or field-test items. Anchor items are items used repeatedly on the test in order to link scores from year to year and are not released to the public. Field-test items do not count toward students' test scores, but they are being administered to determine their usability as core items on future administrations of the FCAT. To accommodate items on the 2006 FCAT, 30 separate test forms were constructed for each grade/subject combination. All forms within a grade/subject contained the same core items plus six to eight anchor or field-test items. Core and anchor items were included on Forms 27–30 and were taken by an early-return calibration sample of students. Forms 1–26 consisted of core and field-test items. By having numerous forms for anchor and field-test items, a relatively large number of the items were dispersed among subsets of students. Student responses to anchor and field-test items did not contribute to their scores.

On the 2006 FCAT, the number of core items varied for mathematics tests by grade, as seen in Table 1. For FCAT Reading, the number of core items was identical for all grades.

Table 1. Number of Core Items by Subject and Grade

	Mathematics		Reading	
Grade	Number of Core Items	Total Points	Number of Core Items	Total Points
3	40	40	45	45
4	39	39	45	51
5	50	60	45	45
6	44	44	45	45
7	44	44	45	45
8	50	60	45	51
9	44	44	45	45
10	50	60	45	51

Score reports consist of reading and mathematics scale scores plus subscores on performance-category assignments. Performance-category assignments are based on standard-setting procedures that divide the reading and mathematics scales into distinct levels of performance (FDOE, 1998, November 6, 2001).

FCAT Reading tests report subscores in four reporting categories (also referred to as clusters):

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

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

- Number Sense, Concepts, and Operations
- Measurement

- Geometry and Spatial Sense
- Algebraic Thinking
- Data Analysis and Probability

Report Content

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

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

The FCAT is a continuous assessment system. While the essential structure and focus of the FCAT tests remain fairly fixed over time and student achievement results maintain a level of comparability across testing years, specific questions on a test administered in any given year may vary. In addition to the variability of test questions administered on the "core" portion of the test (i.e., the portion of the test that actually contributes to students' reported scores), students will also answer some items on the test that do not count toward their ultimate scores. Instead, these items will be used for equating (anchor items[2]) or field testing. Field-test items provide necessary data for the development of future tests.

This report refers to *core* and *anchor* items. Before 2004, FCAT core and anchor items comprised the total set of items used to scale and equate. However, to address the release of test items to the public, FDOE decided to remove anchors from the set of items used to determine student scale scores. In doing so, anchor items can still be used for equating but will not be released to the public (since students do not receive scores for them); thus, the equating process is not compromised.

Removing the anchors from the core set changed the way data are summarized in this report. To begin, core and anchor-item statistics are presented separately in the Appendices. Secondly, summary statistics presented in the main body of the report are for core items only. Summary

---

[2] Anchor items were separated from the core set of items beginning in 2004.

statistics for anchor items appear in Tables 1b–1g, Appendix A (Reading), and Appendix B (Mathematics).

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

ITEM PREPARATION AND TEST ASSEMBLY

The FDOE staff and several committees review the passages on which the FCAT Reading items are based. Item reviews[3] are conducted following reading passage reviews. Reading items must go through a three-phase development process before they are included on the FCAT. During the first phase, education professionals familiar with both the style and intent of each FCAT benchmark draft the items. Draft items received by the FDOE contractor are subjected to critical content and editorial reviews. These items are then forwarded to the content staff at the Test Development Center (TDC) in Tallahassee, where they receive an additional review. Any item submitted typically has 1 of 3 fates: (a) it is accepted with no (or minor) edits, (b) it is rejected as inappropriate for the FCAT, or (c) it is returned to the contractor with comments requesting changes in style or focus, so the item can be returned to the review process. Ongoing dialogue on the "accept with revisions" items between the contractor and TDC staff assures that both the contractor and the TDC staff deem all items appropriate.

In the second phase of item development, FCAT items go through a rigorous review process before they can be field tested. The procedures used for item review for the 2006 FCAT 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, May 2001).

In phase three, items are field tested during the regular FCAT administration. The items are quantitatively evaluated and placed in the item bank for possible use as core items in subsequent FCAT assessments.

Harcourt and TDC staffs build forms through a multistep process (FDOE, 2004). This process is guided by (a) content considerations required by the test blueprints for each content area and grade and (b) the statistical characteristics tied to each item. Typically, Harcourt content and psychometric staffs propose draft forms for each grade and subject for TDC review. These draft forms are assembled according to the content guidelines documented for each test as well as statistical guidelines documenting how well the proposed tests (i.e., whole tests as well as reportable strands/clusters) match the characteristics of previously administered versions of the FCAT.

---

[3] Item reviews are conducted by the following parties: (a) the FDOE for content, sensitivity/bias, match to benchmark, and FCAT style; (b) community sensitivity committees; (c) bias committees, with representatives from diverse backgrounds; and (d) grade-level content committees, with professional representatives from schools, school districts, and universities.

CONSTRUCTED-RESPONSE SCORING PROCEDURES

For some grade/content combinations, students must provide handwritten responses to performance task questions. These responses are then scored by individual human scorers, rather than by machines. All procedures related to scoring constructed-response items, also called performance task items, are guided by a set of *Handscoring Specifications*. The procedures include rangefinding, hiring, staffing, training, scoring, and reporting constructed-response scores. Because the *Handscoring Specifications* contain secure information about FCAT content, they are not available to the public. For additional information about handscoring procedures, consult the *FCAT Handbook—A Resource for Educators* (http://fcat.fldoe.org/handbk/fcathandbook.asp).

Short- and extended-response performance task items are handscored by professional scorers. To be selected and eligible to score the FCAT, candidates must have at least a bachelor's degree in a field related to the subject they will be scoring. For reading, examples of subject-related fields are Education, English Literature, Journalism, and Communications. For mathematics, examples of subject-related fields are Education, Mathematics, Engineering, Accounting, and Finance. Depending on the subject, applicants may be required to also take a subject-area exam or write an essay.

Educator Involvement

The anchor papers and item-specific criteria for the performance task items are developed initially by Florida educators serving on Rangefinder Committees. After performance task items are selected for use as operational items, Rangefinder Review Committees review the scoring guides and training materials originally established by the Rangefinder Committees. The role of the Rangefinder Review Committee is chiefly to clarify scoring criteria, not to modify the scoring standards initially set by the Rangefinder Committee. Each committee is comprised of Florida educators, including teachers from the targeted grade levels and subject areas, school and district curriculum specialists, and university faculty from the discipline areas.

Scorer Training

Training of scorers is accomplished through the use of FDOE-approved training materials determined during the "Rangefinder Review" sessions held with state educators and members of the FDOE.

Potential scorers are given an overview of the project along with FDOE expectations and guidelines. To ground them in the rules of scoring, they are shown several sets of training papers. Scorers are then given "qualification sets" to ensure that a minimum agreement percentage is met. Items are scored in groups of two or more [this process is known as the "rater item block" (RIB) format], and the scorer must qualify on all items within the RIB in order to score the RIB. Only after the successful completion of the qualifying process are scorers allowed to assess actual student responses. To ensure consistency between training sessions (i.e., if an item or group of items are used in training with more than one group of scorers at separate times), papers are presented in the same order with the same comments. This is done so that each group of scorers will complete training with the same rules and information.

At the end of training, candidates must pass a qualifying examination. The examination requires them to score sets of sample essays or students responses for which scores have been established by Florida educators. To pass the examination, candidates must match the pre-established scores.

Handscoring

FCAT scoring of performance tasks is *holistic*, as opposed to *analytic*,[4] meaning that a single rating is given for the response as a whole. For FCAT Reading and Mathematics, scorers assign scores of 0, 1, or 2 for short-response performance task items. For extended-response performance task items, scorers use a scale of 0, 1, 2, 3, or 4.

Those qualified as professional scorers work in teams of 10-15 members, with each team having a Team Leader. Each team specializes in a set of two to three performance task items, or "rater item blocks" (RIBs). A Scoring Director and an Assistant Scoring Director supervise all of the teams assigned to a RIB. Prior to the scoring sessions, all student responses are scanned electronically. At the scoring centers, scorers work individually at computer workstations to read the scanned student responses assigned to them on their computer monitors.

Each student response is independently read and scored by at least two professional scorers. For short-response performance task items, if the scorers' two scores are not identical, a third scorer reviews the response to resolve the difference. For extended-response performance task items, the two scores assigned are averaged for a final score. A third scorer is used if the two scores assigned are nonadjacent. This third scoring, called resolution scoring, is performed by a Team Leader.

Year-to-Year Calibration

In order to ensure that an item scored in a previous administration will be scored the same way in a current administration, all previous training materials are sent to the "Rangefinder Review" session and scoring rationales are discussed. Minimal changes are made to the training and validity sets, and the same scoring notes are used. Scores on individual papers cannot be changed.

Backreading

Backreading 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 teams (teams consist of no more than 12 scorers and one team leader). This process helps ensure that the scorers are assigning valid scores to the student responses. At the beginning of the project, team

[4] An analytic score is based on a combination of separate ratings for specified traits of the response.

leaders are asked to spend their time performing backreading for everyone several times a day to identify the strength of individual scorers. Team leaders ask scorers to review papers that have been incorrectly scored in order to assess their skills and help scorers who fail to adhere to scoring standards. To ensure accuracy throughout the project, backreading is implemented for all scorers.

Control of Scorer Drift

There are many methods implemented to control scorer drift. One daily process is to have team members spend 10–15 minutes, or longer if needed, reviewing rangefinder and horizontal training sets for each item in the RIB that they are scoring. Rangefinder sets consist of two to four student responses (selected by the rangefinder review committees) for each score point and are used to illustrate how the holistic rubric is applied. Horizontal training sets are constructed of 30–80 student responses, divided into three or four sets that fit within the scoring criteria. These sets allow scorers to practice applying the rubric while internalizing all nuances of the holistic rubric. Before scoring begins, a “start of shift” refresher of the scoring material occurs by silent or team reviews, followed by an opportunity to ask scoring questions. The scoring directors/assistant scoring directors, along with the team leaders, lead discussions to reorient scorers and re-anchor them in the common scoring criteria. Discussions to address simple clarifications may occur within teams, or larger issues may be addressed to the entire group by the scoring director. As needed, a pre-scored set of calibration papers, also referred to as discussion papers, is used for calibrating and identifying any unforeseen issues that may arise from particular unanticipated types of responses. The selection of material to review may vary daily. Scorers are encouraged to refer to rubrics and rangefinders often to assist them in assigning accurate scores. This helps to keep all scorers and team leaders grounded in the rules and guidelines laid out in training.

Another process available to control scorer drift is the use of calibration sets. Calibration is a form of training that leads to a greater level of accuracy and consensus within the scoring pool (i.e., scorers and their team leaders). Calibration sets are selected responses that illustrate specific issues for large or small group discussions.

Embedded in the flow of student responses that scorers score at their work stations are responses for which scores have already been established by the FCAT Rangefinder and Rangefinder Review Committees. As a monitoring tool, a validity report shows how frequently a scorer agrees with the “true score” given to pre-selected and expert-scored validity responses. By accessing validity reports, the scoring director can see which validity papers are being missed, which scorers are missing validity papers, and which scorers are scoring the papers too high or too low.

Reliability reports show how often two scorers give the same score when scoring the same response. These reports also show if scorers deviate from the standard in a way that is consistently high or low. The scoring director can then use specific information from these reports to reground scorers in the relevant training materials and scoring guidelines.

As mentioned above, backreading helps reduce scorer drift by alerting scorers to their mistakes. All of the validity and reliability reports, along with calibration sets, are quality control measures that help prevent scorer drift. Retraining is conducted for scorers whose scores are consistently inaccurate or fall below acceptable standards. If retraining is unsuccessful, scorers are dismissed from the program.

2006 FCAT STATISTICS

This section of the report presents psychometric analyses of the 2006 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 students' responses on an early timeline. The general selection strategy was to pick schools that provide a sample of students that are representative of the state's regions, ethnic diversity, and achievement scores in past years. Only standard curriculum students were used in the analyses; exceptional 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.[5] More details about the selection of this sample appear in *Plan for Selecting the Calibration Sample for the 2006 FCAT Administration* (FDOE, November, 2006).

This section begins with a description of the calibration sampling procedure and presents a comparison of the calibration samples to the state's total distributions of students. It is recognized that this presentation is out of chronological order and was, in fact, conducted after all of the analyses were completed; however, the comparison is presented first to establish the credibility of the remaining analyses.

Calibration Sample

The Florida Sampling Plan is designed to select a representative sample of schools in order to provide a timely analysis of the results of the test administration. The schools are selected to model the overall demographic and academic characteristics of the state.

In order to accomplish this goal in a timely fashion, enrollment and scoring information from the previous administration are analyzed. The analysis establishes a target range of characteristics the sample schools need to meet in order to provide a good model that reflects the attributes of the state and geographic regions (North, Central, and South).

The use of historic information and the process involved is based on the following assumption: within a geographic region and across the state, only minor variations of demographic characteristics or academic performance occur in any given year, and any variation that may have occurred in a school selected for the sample would not be so extreme that a fair analysis could not be performed.

[5] Test scores are only computed for students who meet the "attemptedness" criteria. The criteria specifies that a student have at least 6 nonblank answers in each of 2 sessions.

Characteristics

In order to provide an adequate sample size, the selected schools should be able to provide between 8,000 and 8,800 students in total, with at least 8,000 students for each content area. Every grade in the selected schools has to participate in the sample and must have a minimum enrollment of 20 students per grade. Also, schools that participated in the previous year's sample selection were not selected this year.

The sample must meet the following characteristics for each grade and content area:

a. The sample should maintain the same geographic region distribution, plus or minus 200 students.
b. The number of schools selected should maintain the same geographic region distribution, plus or minus three schools.
c. The sample must include, at each grade level, a school from each of the largest six districts in the state.
d. The percentage of the four major ethnic groups (White, African-American, Hispanic and Other, which includes Asian, American Indian, and Multiracial students) should maintain the same ratio as the state and within each geographic region (North, Central, and South), plus or minus 5 percent.
e. The standard deviation unit (computed by dividing the absolute value of the difference between the sample mean and the state mean by the standard deviation of the state) must be 0.2 or less.
f. The standard deviation ratio (computed by dividing the standard deviation of the sample by the standard deviation of the state) must be between 0.9 and 1.1.

Evaluation of Representativeness

Tables 2 through 49 on the following pages compare each grade/subject calibration sample with statewide sets of students. One set of comparison students, labeled "Total," includes all students with FCAT records for 2006.[6] Some of these students, however, did not receive FCAT scores because they failed the attemptedness criteria. A second set of students includes all standard curriculum students, including those who did not receive test scores because of failing the attemptedness criteria. These two sets of students provide a basis for comparing the gender and ethnicity distributions of the calibration samples. Also, note that the number of students across the respective categories do not sum to the total listed because of missing ethnicity and gender information (some students do not provide this information).

In addition to the gender and ethnicity distributions, test scores for the calibration samples are compared to test scores for the total population that received scores and for the total standard curriculum population that received test scores. Test score means for these groups are disaggregated by ethnicity and gender.

---

[6] Exceptions are students who fell into the following categories: home schooled (home_sch), districts (dist) 69 or 70, and special school codes (SPCSHC) 10 or 11.

The first table on each of the following pages examines ethnicity distributions. These tables show that ethnicity representations for the "calibration sample" are reasonable approximations of the state ethnicity distributions; however, the ethnicity distributions of "standard curriculum students" tend to match the overall student population distributions a little more closely than the calibration sample. The second table on each page examines gender distributions that indicate similar results for gender as for the ethnicity distributions. The last table on each page presents FCAT score means and standard deviations for different sampling groups. As expected, score means are lower and standard deviations are higher for the total population of students than for standard curriculum students only. Score means for the calibration sample closely match those for the full set of standard curriculum students. Gender distributions for standard curriculum students are also replicated in the calibration samples.

Table 2. Grade 3 Reading Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	233 (2.71%)	2,041 (23.77%)	2,168 (25.24%)	24 (0.28%)	331 (3.85%)	3,767 (43.86%)	8,588
Standard curriculum students	4,071 (2.33%)	40,318 (23.11%)	40,241 (23.06%)	537 (0.31%)	6,927 (3.97%)	81,872 (46.92%)	174,489
All students	4,617 (2.26%)	47,501 (23.26%)	50,482 (24.72%)	618 (0.30%)	7,800 (3.82%)	92,620 (45.35%)	204,238

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

Table 3. Grade 3 Reading Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	4,339 (50.52%)	4,234 (49.30%)	8,588
Standard curriculum students	88,183 (50.54%)	85,852 (49.20%)	174,489
All students	98,492 (48.22%)	105,204 (51.51%)	204,238

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

Table 4. Grade 3 Reading Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	323.75	50.13	8,588	323.07	49.80	174,489	313.49	56.80	204,238
Female	324.58	49.32	4,339	324.64	48.71	88,183	317.60	54.29	98,492
Male	323.03	50.89	4,234	321.60	50.74	85,852	309.79	58.71	105,204
African American	299.42	46.16	2,041	299.69	46.40	40,318	291.04	52.67	47,501
Hispanic	317.75	48.46	2,168	314.64	48.51	40,241	301.92	57.67	50,482
White	338.71	47.31	3,767	337.60	46.72	81,872	329.86	53.07	92,620

Table 5. Grade 3 Mathematics Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	236 (2.74%)	2,053 (23.83%)	2,177 (25.27%)	23 (0.27%)	333 (3.87%)	3,772 (43.87%)	8,615
Standard curriculum students	4,062 (2.33%)	40,047 (22.93%)	40,090 (22.96%)	533 (0.31%)	6,881 (3.94%)	81,453 (46.65%)	174,622
All students	4,599 (2.25%)	47,150 (23.06%)	50,202 (24.56%)	613 (0.30%)	7,746 (3.79%)	92,100 (45.05%)	204,429

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

Table 6. Grade 3 Mathematics Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	4,355 (50.55%)	4,247 (49.30%)	8,615
Standard curriculum students	88,251 (50.54%)	85,935 (49.21%)	174,622
All students	98,552 (48.21%)	105,343 (51.53%)	204,429

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

Table 7. Grade 3 Mathematics Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	333.16	60.10	8,615	332.57	59.81	174,622	323.56	64.93	204,429
Female	328.81	59.22	4,355	328.55	58.42	88,251	321.74	62.60	98,552
Male	337.85	60.44	4,247	336.86	60.84	85,935	325.43	66.93	105,343
African American	299.04	56.94	2,053	302.25	56.49	40,047	294.12	60.82	47,150
Hispanic	332.28	58.22	2,177	326.43	58.09	40,090	314.92	64.45	50,202
White	350.26	54.77	3,772	349.21	55.48	81,453	341.95	60.37	92,100

Table 8. Grade 4 Reading Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	207 (2.73%)	1,771 (23.32%)	1,861 (24.51%)	21 (0.21%)	285 (3.75%)	3,428 (45.14%)	7,594
Standard curriculum students	3,899 (2.41%)	35,213 (21.73%)	36,499 (22.52%)	498 (0.31%)	6,069 (3.74%)	79,495 (49.05%)	162,084
All students	4,401 (2.29%)	42,331 (21.99%)	46,908 (24.37%)	573 (0.30%)	6,883 (3.58%)	90,900 (47.23%)	192,480

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

Table 9. Grade 4 Reading Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	3,929 (51.74%)	3,642 (47.96%)	7,594
Standard curriculum students	83,668 (51.62%)	78,037 (48.15%)	162,084
All students	94,703 (49.20%)	97,306 (50.55%)	192,480

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

Table 10. Grade 4 Reading Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	324.15	45.44	7,594	323.94	44.56	162,084	313.67	53.48	192,480
Female	327.06	44.82	3,929	327.80	43.71	83,668	320.30	50.30	94,703
Male	324.16	45.93	3,642	319.96	44.98	78,037	307.38	55.57	97,306
African American	303.16	44.49	1,771	303.10	42.48	35,213	293.31	50.99	42,331
Hispanic	320.26	43.35	1,861	318.61	42.99	36,499	303.95	55.60	46,908
White	335.49	42.98	3,428	334.38	42.40	79,495	326.54	49.64	90,900

Table 11. Grade 4 Mathematics Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	208 (2.71%)	1,796 (23.43%)	1,862 (24.30%)	22 (0.29%)	282 (3.86%)	3,474 (45.33%)	7,664
Standard curriculum students	3,894 (2.40%)	34,970 (21.55%)	36,332 (22.39%)	491 (0.30%)	6,040 (3.72%)	79,117 (48.75%)	162,290
All students	4,381 (2.27%)	41,997 (21.80%)	46,577 (24.18%)	567 (0.29%)	6,844 (3.55%)	90,381 (46.92%)	192,635

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

Table 12. Grade 4 Mathematics Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	3,953 (51.58%)	3,688 (48.12%)	7,664
Standard curriculum students	83,738 (51.60%)	78,140 (48.15%)	162,290
All students	94,747 (49.18%)	97,393 (50.56%)	192,635

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

Table 13. Grade 4 Mathematics Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	326.08	54.88	7,664	327.76	53.76	162,290	317.84	60.79	192,635
Female	320.97	53.66	3,953	323.60	52.36	83,738	316.03	57.97	94,747
Male	331.83	55.51	3,688	332.43	54.74	78,140	319.80	63.28	97,393
African American	299.20	51.20	1,796	303.05	51.19	34,970	293.19	58.03	41,997
Hispanic	324.27	52.61	1,862	324.16	52.13	36,332	310.32	62.29	46,577
White	338.85	51.84	3,474	338.98	50.98	79,117	331.58	56.34	90,381

Table 14. Grade 5 Reading Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	207 (2.56%)	1,817 (22.46%)	1,985 (24.54%)	19 (0.23%)	248 (3.07%)	3,798 (46.95%)	8,090
Standard curriculum students	3,930 (2.38%)	36,492 (22.12%)	36,986 (22.42%)	504 (0.31%)	5,772 (3.50%)	80,875 (49.03%)	164,948
All students	4,436 (2.25%)	44,322 (22.49%)	47,274 (23.99%)	594 (0.30%)	6,538 (3.32%)	93,467 (47.43%)	197,054

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

Table 15. Grade 5 Reading Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	4,118 (50.90%)	3,955 (48.89%)	8,090
Standard curriculum students	84,797 (51.41%)	79,770 (48.36%)	164,948
All students	96,203 (48.82%)	100,404 (50.95%)	197,054

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

Table 16. Grade 5 Reading Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	316.33	49.89	8,090	315.47	51.97	164,948	304.50	59.69	197,054
Female	319.66	48.68	4,118	318.92	50.17	84,797	310.98	56.15	96,203
Male	312.99	50.90	3,955	311.96	53.48	79,770	298.43	62.21	100,404
African American	293.73	45.86	1,817	291.28	48.23	36,492	280.86	55.43	44,322
Hispanic	311.82	49.25	1,985	308.66	50.60	36,986	293.66	60.99	47,274
White	328.33	47.74	3,798	328.28	49.73	80,875	319.55	56.43	93,467

Table 17. Grade 5 Mathematics Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	209 (2.59%)	1,806 (22.39%)	1,974 (24.47%)	20 (0.25%)	245 (3.04%)	3,796 (47.06%)	8,066
Standard curriculum students	3,896 (2.36%)	36,007 (21.84%)	36,553 (22.17%)	497 (0.30%)	5,713 (3.47%)	80,375 (48.76%)	164,853
All students	4,396 (2.23%)	43,724 (22.18%)	46,694 (23.69%)	587 (0.30%)	6,471 (3.28%)	92,867 (47.11%)	197,111

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

Table 18. Grade 5 Mathematics Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	4,118 (51.05%)	3,934 (48.77%)	8,066
Standard curriculum students	84,775 (51.42%)	79,732 (48.37%)	164,853
All students	96,235 (48.82%)	100,445 (50.96%)	197,111

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

Table 19. Grade 5 Mathematics Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	336.87	43.53	8,066	337.57	43.04	164,853	328.77	50.82	197,111
Female	334.36	43.22	4,118	334.86	41.89	84,775	328.33	47.81	96,235
Male	339.60	43.65	3,934	340.57	43.94	79,732	329.31	53.45	100,445
African American	313.47	42.54	1,806	315.31	41.65	36,007	305.95	50.32	43,724
Hispanic	334.90	42.29	1,974	334.14	41.82	36,553	322.88	51.45	46,694
White	347.55	39.13	3,796	348.00	39.56	80,375	341.22	45.85	92,867

Table 20. Grade 6 Reading Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	129 (1.68%)	1,899 (24.67%)	1,536 (19.95%)	23 (0.30%)	195 (2.53%)	3,889 (50.51%)	7,699
Standard curriculum students	3,721 (2.31%)	35,345 (21.93%)	35,051 (21.75%)	512 (0.32%)	5,175 (3.21%)	80,869 (50.18%)	161,154
All students	4,105 (2.20%)	41,348 (22.12%)	42,900 (22.95%)	580 (0.31%)	5,750 (3.08%)	91,753 (49.08%)	186,948

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

Table 21. Grade 6 Reading Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	3,891 (50.54%)	3,782 (49.12%)	7,699
Standard curriculum students	81,770 (50.74%)	78,900 (48.96%)	161,154
All students	90,757 (48.55%)	95,650 (51.16%)	186,948

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

Table 22. Grade 6 Reading Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	315.26	49.94	7,699	319.87	49.54	161,154	310.89	55.83	186,948
Female	318.34	48.63	3,891	322.46	48.07	81,770	316.01	52.71	90,757
Male	312.27	51.02	3,782	317.41	50.75	78,900	306.21	58.16	95,650
African American	293.87	46.27	1,899	296.36	46.88	35,345	287.72	52.52	41,348
Hispanic	306.44	53.55	1,536	313.51	50.20	35,051	300.89	58.59	42,900
White	328.09	46.08	3,889	331.88	46.21	80,869	324.66	51.71	91,753

Table 23. Grade 6 Mathematics Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	129 (1.67%)	1,906 (24.73%)	1,536 (19.93%)	23 (0.30%)	194 (2.52%)	3,892 (50.50%)	7,707
Standard curriculum students	3,720 (2.31%)	35,316 (21.93%)	35,015 (21.74%)	511 (0.32%)	5,175 (3.21%)	80,815 (50.19%)	161,028
All students	4,105 (2.20%)	41,323 (22.12%)	42,865 (22.95%)	578 (0.31%)	5,754 (3.08%)	91,660 (49.07%)	186,792

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

Table 24. Grade 6 Mathematics Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	3,890 (50.50%)	3,790 (49.18%)	7,707
Standard curriculum students	81,717 (50.75%)	78,830 (48.95%)	161,028
All students	90,711 (48.56%)	95,544 (51.15%)	186,792

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

Table 25. Grade 6 Mathematics Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	317.66	56.81	7,707	322.29	55.75	161,028	311.68	64.54	186,792
Female	317.40	53.49	3,892	321.29	53.71	81,717	313.60	60.37	90,711
Male	318.32	59.66	3,790	323.61	57.56	78,830	310.09	68.09	95,544
African American	288.55	56.05	1,906	293.85	55.08	35,316	282.51	63.91	41,323
Hispanic	312.07	58.31	1,536	315.93	55.10	35,015	301.88	66.16	42,865
White	332.77	50.32	3,892	335.83	50.68	80,815	327.34	58.46	91,660

Table 26. Grade 7 Reading Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	141 (1.70%)	2,209 (26.58%)	1,550 (18.65%)	21 (0.25%)	210 (2.53%)	4,159 (50.04%)	8,312
Standard curriculum students	3,890 (2.26%)	39,530 (23.00%)	37,861 (22.03%)	538 (0.31%)	4,854 (2.82%)	84,664 (49.26%)	171,857
All students	4,313 (2.13%)	47,295 (23.36%)	47,258 (23.34%)	609 (0.30%)	5,500 (2.72%)	96,872 (47.85%)	202,438

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

Table 27. Grade 7 Reading Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	4,258 (51.23%)	4,031 (48.50%)	8,312
Standard curriculum students	87,840 (51.11%)	83,501 (48.59%)	171,857
All students	98,559 (48.69%)	103,271 (51.01%)	202,438

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

Table 28. Grade 7 Reading Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	315.28	48.96	8,312	320.15	48.94	171,857	310.24	55.63	202,438
Female	318.48	48.55	4,258	323.07	48.07	87,840	315.81	53.15	98,559
Male	312.09	49.08	4,031	317.31	49.56	83,501	305.14	57.32	103,271
African American	294.59	46.10	2,209	299.42	45.56	39,530	289.42	52.18	47,295
Hispanic	306.20	48.69	1,550	311.79	48.98	37,861	298.14	57.59	47,258
White	328.67	46.13	4,159	332.52	46.31	84,664	324.98	51.66	96,872

Table 29. Grade 7 Mathematics Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	141 (1.69%)	2,221 (26.67%)	1,552 (18.64%)	21 (0.25%)	209 (2.51%)	4,160 (49.95%)	8,328
Standard curriculum students	3,889 (2.26%)	39,542 (23.02%)	37,851 (22.03%)	537 (0.31%)	4,850 (2.82%)	84,591 (49.24%)	171,783
All students	4,312 (2.13%)	47,285 (23.37%)	47,251 (23.36%)	605 (0.30%)	5,490 (2.71%)	96,770 (47.83%)	202,303

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

Table 30. Grade 7 Mathematics Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	4,268 (51.25%)	4,035 (48.45%)	8,328
Standard curriculum students	87,799 (51.11%)	83,466 (48.59%)	171,783
All students	98,522 (48.70%)	103,177 (51.00%)	202,303

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

Table 31. Grade 7 Mathematics Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	312.31	51.53	8,328	316.44	51.07	171,783	306.56	58.81	202,303
Female	311.27	50.00	4,268	314.67	49.60	87,799	307.40	55.62	98,522
Male	313.56	53.00	8,328	318.57	52.29	83,466	306.01	61.55	103,177
African American	286.37	50.42	2,221	291.83	50.78	39,542	280.91	59.17	47,285
Hispanic	306.39	50.97	1,552	310.09	49.78	37,851	298.23	57.96	47,251
White	326.82	46.38	4,160	329.26	46.43	84,591	321.32	53.62	96,770

Table 32. Grade 8 Reading Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	134 (1.73%)	2,021 (26.04%)	1,506 (19.40%)	21 (0.27%)	182 (2.35%)	3,878 (49.97%)	7,761
Standard curriculum students	3,945 (2.32%)	38,308 (22.53%)	37,184 (21.87%)	525 (0.31%)	4,508 (2.65%)	85,126 (50.06%)	170,052
All students	4,395 (2.19%)	46,013 (22.96%)	46,271 (23.09%)	594 (0.30%)	5,047 (2.52%)	97,581 (48.69%)	200,421

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

Table 33. Grade 8 Reading Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	3,996 (51.49%)	3,752 (48.34%)	7,761
Standard curriculum students	87,609 (51.52%)	81,993 (48.22%)	170,052
All students	98,289 (49.04%)	101,596 (50.69%)	200,421

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

Table 34. Grade 8 Reading Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	306.79	44.88	7,761	309.38	45.74	170,052	299.09	54.23	200,421
Female	309.98	44.64	3,996	313.04	45.02	87,609	305.68	51.17	98,289
Male	303.55	44.87	3,752	305.72	45.99	81,993	292.92	56.19	101,596
African American	288.06	42.32	2,021	288.25	45.30	38,308	276.75	54.17	46,013
Hispanic	296.56	47.09	1,506	302.27	46.98	37,184	288.58	57.16	46,271
White	319.91	40.86	3,878	320.96	41.12	85,126	313.21	48.13	97,581

Table 35. Grade 8 Mathematics Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	136 (1.75%)	2,019 (25.99%)	1,491 (19.19%)	21 (0.27%)	184 (2.37%)	3,883 (49.99%)	7,768
Standard curriculum students	3,933 (2.31%)	37,772 (22.19%)	36,774 (21.60%)	513 (0.30%)	4,456 (2.62%)	84,407 (49.58%)	170,244
All students	4,370 (2.18%)	45,237 (22.56%)	45,685 (22.79%)	578 (0.29%)	4,990 (2.49%)	96,606 (48.19%)	200,482

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

Table 36. Grade 8 Mathematics Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	3,986 (51.31%)	3,753 (48.31%)	7,768
Standard curriculum students	87,647 (51.48%)	82,061 (48.20%)	170,244
All students	98,325 (49.04%)	101,525 (50.64%)	200,482

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

Table 37. Grade 8 Mathematics Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	320.83	43.00	7,768	324.05	43.03	170,244	314.45	51.97	200,482
Female	320.69	41.63	3,986	323.18	41.54	87,647	316.24	48.39	98,325
Male	321.45	43.97	3,753	325.28	44.31	82,061	312.97	55.03	101,525
African American	298.87	42.10	2,019	300.87	43.11	37,772	289.56	53.35	45,237
Hispanic	316.57	42.80	1,491	318.66	41.77	36,774	307.35	51.34	45,685
White	333.31	37.53	3,883	336.02	37.66	84,407	328.51	45.63	96,606

Table 38. Grade 9 Reading Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	185 (2.24%)	2,180 (26.35%)	1,765 (21.33%)	32 (0.39%)	189 (2.28%)	3,853 (46.57%)	8,273
Standard curriculum students	4,275 (2.36%)	41,060 (22.64%)	39,336 (21.69%)	524 (0.29%)	4,033 (2.22%)	91,350 (50.36%)	181,388
All students	4,692 (2.20%)	49,266 (23.14%)	48,636 (22.84%)	614 (0.29%)	4,434 (2.08%)	104,361 (49.02%)	212,904

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

Table 39. Grade 9 Reading Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	4,427 (53.51%)	3,782 (45.71%)	8,273
Standard curriculum students	93,247 (51.41%)	87,405 (48.19%)	181,388
All students	104,298 (48.99%)	107,755 (50.61%)	212,904

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

Table 40. Grade 9 Reading Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	315.35	46.66	8,273	314.90	47.78	181,388	306.03	54.01	212,904
Female	316.90	45.13	4,427	317.19	46.52	93,247	310.55	51.56	104,298
Male	314.23	47.89	3,782	312.77	48.79	87,405	301.93	55.83	107,755
African American	295.68	44.89	2,180	291.10	45.34	41,060	281.69	51.85	49,266
Hispanic	308.18	45.49	1,765	304.89	48.69	39,336	293.43	56.11	48,636
White	329.52	43.01	3,853	329.21	42.84	91,350	322.41	48.00	104,361

Table 41. Grade 9 Mathematics Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	186 (2.25%)	2,180 (26.37%)	1,762 (21.37%)	33 (0.40%)	187 (2.26%)	3,852 (46.59%)	8,267
Standard curriculum students	4,270 (2.36%)	40,957 (22.63%)	39,244 (21.69%)	531 (0.29%)	4,020 (2.22%)	91,148 (50.37%)	180,961
All students	4,687 (2.21%)	49,113 (23.13%)	48,523 (22.85%)	625 (0.29%)	4,424 (2.08%)	104,105 (49.02%)	212,359

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

Table 42. Grade 9 Mathematics Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	4,425 (53.53%)	3,780 (45.72%)	8,267
Standard curriculum students	93,059 (51.42%)	87,187 (48.18%)	180,961
All students	104,066 (49.00%)	107,470 (50.61%)	212,359

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

Table 43. Grade 9 Mathematics Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	309.49	47.73	8,267	310.75	48.11	180,961	302.07	54.22	212,359
Female	306.30	46.09	4,425	307.83	47.42	93,059	301.22	52.54	104,066
Male	313.90	49.01	3,780	314.19	48.43	87,187	303.16	55.67	107,470
African American	287.10	47.54	2,180	285.22	47.22	40,957	275.30	54.09	49,113
Hispanic	306.77	45.41	1,762	302.98	47.33	39,244	292.81	53.89	48,523
White	322.42	43.32	3,852	324.25	42.77	91,148	317.37	48.09	104,105

Table 44. Grade 10 Reading Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	180 (2.50%)	1,784 (24.78%)	1,655 (22.99%)	34 (0.47%)	156 (2.17%)	3,373 (46.86%)	7,198
Standard curriculum students	4,215 (2.61%)	34,567 (21.36%)	34,721 (21.46%)	468 (0.29%)	2,883 (1.78%)	84,318 (52.11%)	161,794
All students	4,592 (2.47%)	40,424 (21.78%)	41,733 (22.49%)	533 (0.29%)	3,191 (1.72%)	94,379 (50.86%)	185,568

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

Table 45. Grade 10 Reading Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	3,796 (52.74%)	3,387 (47.05%)	7,198
Standard curriculum students	85,100 (52.60%)	76,036 (47.00%)	161,794
All students	94,003 (50.66%)	90,830 (48.95%)	185,568

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

Table 46. Grade 10 Reading Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	310.57	51.95	7,198	307.61	52.94	161,794	298.39	59.32	185,568
Female	312.81	51.19	3,796	309.89	52.27	85,100	303.00	56.95	94,003
Male	308.34	52.52	3,387	305.42	53.34	76,036	293.92	61.19	90,830
African American	288.70	49.48	1,784	280.21	49.29	34,567	269.92	56.10	40,424
Hispanic	305.22	49.93	1,655	298.28	52.71	34,721	286.97	59.37	41,733
White	323.36	49.79	3,373	321.86	48.80	84,318	314.57	54.68	94,379

Table 47. Grade 10 Mathematics Frequency Distributions for Different Student Groups by Ethnicity

	Asian	African American	Hispanic	American Indian	Multi-racial	White	Total[a]
Calibration sample	181 (2.53%)	1,770 (24.69%)	1,652 (23.05%)	32 (0.45%)	154 (2.15%)	3,361 (46.89%)	7,168
Standard curriculum students	4,153 (2.58%)	33,530 (20.81%)	33,928 (21.05%)	458 (0.28%)	2,802 (1.74%)	82,765 (51.36%)	161,156
All students	4,518 (2.45%)	39,099 (21.17%)	40,613 (21.99%)	522 (0.28%)	3,097 (1.68%)	92,479 (50.07%)	184,707

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

Table 48. Grade 10 Mathematics Frequency Distributions for Different Student Groups by Gender

	Female	Male	Total[a]
Calibration sample	3,780 (52.73%)	3,372 (47.04%)	7,168
Standard curriculum students	84,789 (52.61%)	75,681 (46.96%)	161,156
All students	93,670 (50.71%)	90,272 (48.87%)	184,707

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

Table 49. Grade 10 Mathematics Mean Scale Scores for Different Student Groups

	Calibration Sample			Standard Curriculum Students			All Students		
	M	SD	N	M	SD	N	M	SD	N
All	331.11	36.59	7,168	330.83	38.15	161,156	324.04	45.43	184,707
Female	328.75	35.48	3,780	328.65	36.93	84,789	323.44	42.90	93,670
Male	333.98	37.14	3,372	333.59	39.01	75,681	324.92	47.76	90,272
African American	313.59	37.99	1,770	310.90	38.58	33,530	302.29	48.02	39,099
Hispanic	327.94	36.02	1,652	324.66	37.89	33,928	317.17	45.07	40,613
White	340.74	31.33	3,361	341.50	32.35	82,765	336.30	38.72	92,479

2006 FCAT Item Analysis

This section contains classical item analysis statistics for difficulty and item-total correlations. For each of the items on the 16 tests (2 subjects × 8 grades), item difficulties (*p*-values), item-total correlations, and correlations between the item and reporting categories within each of the subject areas were computed. Item-specific results are presented in Appendices A (Reading) and B (Mathematics). Tables 50–55 summarize the item analysis results by presenting the minimum, 25th percentile, 50th percentile, 75th percentile, and maximum values for each grade/subject test (across all core items).

Item Difficulty Summary

For MC and GR (1 point) items, *p*-values are simply the mean points across all students. For these items, the *p*-value also corresponds to the proportion of students who answered the item correctly. To facilitate comparisons among all item types, item difficulties for the PT items are computed as the mean points achieved divided by total possible points.

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

Table 50. Proportional[1] *p*-value Summary Data for All Reading Items

Grade	Number of Items	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
3	45	0.440	0.638	0.681	0.795	0.954
4	45	0.357	0.564	0.733	0.814	0.902
5	45	0.386	0.598	0.715	0.827	0.944
6	45	0.346	0.545	0.629	0.763	0.907
7	45	0.372	0.612	0.692	0.778	0.911
8	45	0.322	0.566	0.674	0.755	0.876
9	45	0.408	0.531	0.634	0.713	0.887
10	45	0.327	0.572	0.676	0.811	0.912

[1]Mean score divided by total possible score.

Table 51. Proportional[1] *p*-value Summary Data for All Mathematics Items

Grade	Number of Items	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
3	40	0.413	0.539	0.661	0.749	0.924
4	39	0.354	0.539	0.636	0.790	0.949
5	50	0.248	0.450	0.538	0.626	0.930
6	44	0.216	0.440	0.547	0.692	0.898
7	44	0.229	0.438	0.547	0.638	0.863
8	50	0.129	0.394	0.529	0.662	0.882
9	44	0.200	0.385	0.539	0.630	0.881
10	50	0.159	0.352	0.507	0.647	0.929

[1]Mean score divided by total possible score.

Pearson Item-Total Correlations

Tables 52 and 53 show the distribution of item-total raw score correlations and correlations between items and reporting category total scores. These are computed as Pearson correlations[7]. The total raw score is the sum of all item points. The reporting category score is the sum of points from items in that category (called clusters in reading and strands in mathematics). Distributions for the item-reporting category include only correlations of items from that category. Item-by-category correlations are presented in Appendices A and B and include statistics for all item types (MC, GR, SR, and ER).

The most important criterion for the correlation statistics is that they are not negative nor are they near zero. Items with negative correlations should not be used in IRT processing. As seen in Tables 52 and 53, no negative nor near zero correlations were observed.

[7] For the MC and GR items, these correlations are equivalent to point-biserial correlations between the dichotomous variable (right and wrong) and the total score.

Table 52. Item-Total Correlation Summary by Cluster: Reading Core Items

Grade	Reporting Category	No. of Items	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
3	Total	45	0.238	0.379	0.449	0.494	0.578
	Word & Text	7	0.497	0.520	0.564	0.579	0.588
	Main Idea	22	0.282	0.389	0.468	0.509	0.546
	Relationships	12	0.325	0.427	0.508	0.550	0.591
	Research Ref.	4	0.469	0.548	0.655	0.686	0.690
4	Total	45	0.209	0.347	0.390	0.441	0.660
	Word & Text	6	0.465	0.479	0.511	0.523	0.550
	Main Idea	19	0.247	0.386	0.428	0.480	0.699
	Relationships	15	0.333	0.411	0.439	0.476	0.574
	Research Ref.	5	0.495	0.552	0.570	0.617	0.679
5	Total	45	0.169	0.330	0.383	0.445	0.540
	Word & Text	7	0.396	0.403	0.523	0.569	0.571
	Main Idea	17	0.376	0.392	0.453	0.493	0.527
	Relationships	15	0.206	0.386	0.438	0.455	0.567
	Research Ref.	6	0.319	0.503	0.525	0.538	0.578
6	Total	45	0.295	0.390	0.421	0.457	0.535
	Word & Text	11	0.382	0.444	0.482	0.511	0.532
	Main Idea	15	0.352	0.422	0.450	0.478	0.535
	Relationships	11	0.382	0.454	0.515	0.551	0.563
	Research Ref.	8	0.453	0.499	0.518	0.525	0.574
7	Total	45	0.245	0.391	0.434	0.461	0.560
	Word & Text	7	0.503	0.503	0.525	0.554	0.587
	Main Idea	20	0.297	0.397	0.460	0.490	0.526
	Relationships	9	0.466	0.479	0.531	0.544	0.567
	Research Ref	9	0.466	0.486	0.493	0.512	0.521
8	Total	45	0.183	0.353	0.386	0.443	0.633
	Word & Text	6	0.466	0.512	0.535	0.591	0.597
	Main Idea	18	0.228	0.365	0.409	0.424	0.590
	Relationships	8	0.367	0.442	0.512	0.532	0.581
	Research Ref.	13	0.337	0.423	0.448	0.533	0.728
9	Total	45	0.274	0.394	0.427	0.466	0.533
	Word & Text	4	0.636	0.645	0.655	0.659	0.662
	Main Idea	20	0.313	0.432	0.442	0.483	0.544
	Relationships	10	0.414	0.450	0.484	0.510	0.552
	Research Ref.	11	0.455	0.468	0.499	0.536	0.549
10	Total	45	0.263	0.340	0.389	0.437	0.638
	Word & Text	6	0.476	0.488	0.524	0.570	0.576
	Main Idea	15	0.330	0.392	0.439	0.488	0.497
	Relationships	12	0.363	0.416	0.465	0.495	0.523
	Research Ref.	12	0.363	0.397	0.459	0.506	0.711

Table 53. Item-Total Correlation Summary by Strand: Mathematics Core Items

Grade	Reporting Category	No. of Items	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
3	Total	40	0.319	0.405	0.448	0.517	0.587
	Number Sense	12	0.401	0.451	0.513	0.570	0.592
	Measurement	8	0.387	0.508	0.557	0.624	0.656
	Geometry	7	0.457	0.482	0.504	0.531	0.536
	Algebra	6	0.497	0.525	0.580	0.626	0.635
	Data	7	0.475	0.495	0.613	0.645	0.648
4	Total	39	0.262	0.373	0.423	0.473	0.573
	Number Sense	10	0.303	0.444	0.492	0.569	0.596
	Measurement	8	0.387	0.460	0.533	0.570	0.591
	Geometry	7	0.385	0.429	0.514	0.567	0.574
	Algebra	7	0.451	0.490	0.574	0.578	0.585
	Data	7	0.482	0.511	0.548	0.570	0.598
5	Total	50	0.210	0.403	0.453	0.529	0.706
	Number Sense	12	0.450	0.486	0.561	0.583	0.670
	Measurement	11	0.322	0.466	0.547	0.583	0.650
	Geometry	9	0.288	0.453	0.492	0.550	0.780
	Algebra	10	0.454	0.491	0.528	0.553	0.656
	Data	8	0.426	0.466	0.507	0.617	0.789
6	Total	44	0.260	0.386	0.445	0.510	0.605
	Number Sense	9	0.436	0.459	0.468	0.515	0.576
	Measurement	9	0.526	0.552	0.591	0.635	0.642
	Geometry	9	0.442	0.496	0.512	0.562	0.622
	Algebra	8	0.419	0.463	0.529	0.571	0.607
	Data	9	0.329	0.499	0.538	0.556	0.593
7	Total	44	0.299	0.395	0.455	0.520	0.624
	Number Sense	9	0.402	0.474	0.503	0.533	0.593
	Measurement	9	0.528	0.566	0.600	0.652	0.672
	Geometry	8	0.457	0.476	0.494	0.524	0.596
	Algebra	9	0.444	0.501	0.585	0.592	0.625
	Data	9	0.489	0.521	0.529	0.554	0.602
8	Total	50	0.184	0.372	0.480	0.560	0.690
	Number Sense	12	0.352	0.434	0.515	0.557	0.647
	Measurement	11	0.360	0.496	0.613	0.670	0.731
	Geometry	8	0.362	0.488	0.531	0.587	0.770
	Algebra	10	0.360	0.490	0.498	0.537	0.763
	Data	9	0.276	0.456	0.557	0.590	0.774
9	Total	44	0.294	0.402	0.465	0.499	0.638
	Number Sense	8	0.430	0.468	0.502	0.554	0.571
	Measurement	7	0.395	0.491	0.643	0.675	0.706
	Geometry	11	0.364	0.483	0.543	0.601	0.641
	Algebra	10	0.490	0.529	0.537	0.547	0.572
	Data	8	0.455	0.489	0.528	0.549	0.572
10	Total	50	0.252	0.373	0.459	0.560	0.763
	Number Sense	11	0.337	0.412	0.464	0.565	0.610
	Measurement	9	0.450	0.492	0.573	0.598	0.759
	Geometry	10	0.388	0.464	0.583	0.690	0.836
	Algebra	12	0.400	0.510	0.542	0.593	0.716
	Data	8	0.382	0.442	0.491	0.542	0.815

Biserial Item-Total Correlations

The Pearson item-total or point-biserial correlations produced for dichotomous items shown in Tables 52 and 53 are restricted in possible range to the extent that the items are either very easy or very difficult. The biserial correlation may be understood as an estimate of the correlation that would have been obtained if the dichotomous item had actually been a normally distributed continuous measure (see Tables 54 and 55). It will generally be larger than the corresponding point biserial. In fact, if the total score on the test is not normally distributed, the biserial correlation can nonsensically exceed 1 (Cohen & Cohen, 1975). The PT items are not included in the calculation of the biserial correlation.

Table 54. Biserial Correlation Summary by Cluster: Reading Core Items

Grade	Reporting Category	No. of Items	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
3	Total	45	0.230	0.524	0.619	0.685	0.791
	Word & Text	7	0.720	0.727	0.800	0.854	0.889
	Main Idea	22	0.429	0.557	0.627	0.688	0.784
	Relationships	12	0.427	0.594	0.680	0.693	0.809
	Research Ref.	4	0.811	0.824	0.863	0.891	0.893
4	Total	41	0.319	0.447	0.554	0.597	0.761
	Word & Text	6	0.656	0.685	0.714	0.782	0.788
	Main Idea	18	0.376	0.523	0.576	0.646	0.741
	Relationships	13	0.432	0.528	0.631	0.648	0.807
	Research Ref.	4	0.693	0.706	0.723	0.751	0.775
5	Total	45	0.280	0.452	0.535	0.634	0.733
	Word & Text	7	0.513	0.656	0.725	0.732	0.739
	Main Idea	17	0.483	0.522	0.606	0.730	0.771
	Relationships	15	0.416	0.559	0.610	0.682	0.750
	Research Ref.	6	0.551	0.649	0.691	0.736	0.751
6	Total	45	0.387	0.503	0.553	0.611	0.758
	Word & Text	11	0.501	0.579	0.638	0.684	0.749
	Main Idea	15	0.502	0.554	0.585	0.646	0.759
	Relationships	11	0.563	0.662	0.696	0.727	0.791
	Research Ref.	8	0.585	0.640	0.662	0.670	0.721
7	Total	45	0.313	0.503	0.581	0.657	0.804
	Word & Text	7	0.647	0.650	0.695	0.771	0.842
	Main Idea	20	0.380	0.572	0.605	0.646	0.711
	Relationships	9	0.584	0.697	0.719	0.742	0.799
	Research Ref.	9	0.585	0.627	0.637	0.684	0.781
8	Total	41	0.281	0.445	0.500	0.552	0.701
	Word & Text	6	0.642	0.693	0.711	0.750	0.807
	Main Idea	17	0.350	0.473	0.532	0.567	0.663
	Relationships	8	0.591	0.628	0.653	0.750	0.774
	Research Ref.	10	0.438	0.530	0.555	0.589	0.677
9	Total	45	0.360	0.513	0.561	0.619	0.722
	Word & Text	4	0.822	0.823	0.827	0.832	0.835
	Main Idea	20	0.413	0.550	0.572	0.622	0.724
	Relationships	10	0.523	0.624	0.658	0.684	0.710
	Research Ref.	11	0.597	0.611	0.652	0.703	0.716
10	Total	41	0.363	0.448	0.518	0.584	0.715
	Word & Text	6	0.615	0.633	0.667	0.724	0.733
	Main Idea	14	0.490	0.571	0.612	0.645	0.745
	Relationships	11	0.509	0.554	0.614	0.643	0.725
	Research Ref.	10	0.508	0.580	0.590	0.645	0.746

Table 55. Biserial Correlation Summary by Strand: Mathematics Core Items

Grade	Reporting Category	No. of Items	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
3	Total	40	0.400	0.539	0.629	0.685	0.761
	Number Sense	12	0.511	0.671	0.717	0.749	0.783
	Measurement	8	0.652	0.676	0.745	0.821	0.844
	Geometry	7	0.605	0.611	0.661	0.672	0.710
	Algebra	6	0.750	0.781	0.803	0.827	0.840
	Data	7	0.662	0.676	0.777	0.809	0.812
4	Total	39	0.433	0.522	0.555	0.614	0.759
	Number Sense	10	0.557	0.619	0.639	0.745	0.778
	Measurement	08	0.610	0.652	0.678	0.715	0.836
	Geometry	7	0.611	0.646	0.679	0.717	0.730
	Algebra	7	0.660	0.682	0.725	0.735	0.747
	Data	7	0.659	0.689	0.731	0.767	0.782
5	Total	44	0.381	0.526	0.571	0.650	0.781
	Number Sense	11	0.574	0.608	0.708	0.743	0.817
	Measurement	11	0.512	0.604	0.687	0.732	0.820
	Geometry	7	0.539	0.547	0.609	0.691	0.695
	Algebra	9	0.574	0.641	0.671	0.695	0.761
	Data	6	0.564	0.611	0.637	0.651	0.687
6	Total	44	0.383	0.495	0.577	0.648	0.846
	Number Sense	9	0.550	0.587	0.635	0.658	0.724
	Measurement	9	0.670	0.702	0.780	0.832	0.900
	Geometry	9	0.572	0.634	0.680	0.720	0.782
	Algebra	8	0.557	0.594	0.663	0.749	0.773
	Data	9	0.559	0.637	0.683	0.720	0.779
7	Total	44	0.416	0.516	0.592	0.659	0.849
	Number Sense	9	0.586	0.636	0.675	0.693	0.747
	Measurement	9	0.715	0.743	0.800	0.892	0.905
	Geometry	8	0.585	0.627	0.642	0.665	0.767
	Algebra	9	0.559	0.642	0.734	0.743	0.826
	Data	9	0.614	0.660	0.682	0.704	0.757
8	Total	44	0.301	0.504	0.570	0.664	0.891
	Number Sense	12	0.456	0.584	0.657	0.702	0.839
	Measurement	10	0.458	0.624	0.762	0.887	0.955
	Geometry	6	0.556	0.594	0.665	0.682	0.691
	Algebra	8	0.455	0.614	0.624	0.662	0.742
	Data	8	0.451	0.577	0.685	0.739	0.796
9	Total	44	0.374	0.536	0.599	0.650	0.838
	Number Sense	8	0.548	0.593	0.652	0.714	0.730
	Measurement	7	0.644	0.702	0.819	0.863	0.927
	Geometry	11	0.466	0.630	0.683	0.783	0.859
	Algebra	10	0.631	0.673	0.676	0.726	0.746
	Data	8	0.577	0.651	0.686	0.712	0.749
10	Total	44	0.357	0.477	0.587	0.669	0.840
	Number Sense	11	0.521	0.566	0.608	0.712	0.785
	Measurement	8	0.615	0.626	0.748	0.780	0.808
	Geometry	8	0.488	0.602	0.673	0.826	0.909
	Algebra	10	0.517	0.644	0.684	0.715	0.841
	Data	7	0.529	0.553	0.607	0.668	0.753

Item Response Theory Scaling

Measurement Models

The FCAT Reading, Mathematics, and Science assessments are analyzed and scores are reported using a combination of Item Response Theory (IRT) measurement models. IRT provides a seamless approach to a variety of test analyses, development, and reporting activities. IRT is facilitated by fitting, or calibrating, statistical models to student responses. Application of these statistical models results in the simultaneous scaling of item difficulty and student (population) achievement.

HumRRO and Harcourt are responsible for conducting all psychometric analyses. To ensure the accuracy and quality of reported scores, all activities that directly contribute to student scores are verified by the DOE and an independent vendor.

Models

Calibration is facilitated via the mixed-model capabilities of the software program MULTILOG (Thissen, 1991).

3 Parameter Logistic (3PL)

The 3PL model (Lord & Novick, 1968; Lord, 1980) is used to calibrate and analyze MC items. In this model, the probability that a student with an achievement estimate θ responds correctly to item i is

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

where a_i is the item discrimination, b_i is the item difficulty and c_i is the pseudo-guessing parameter.

2 Parameter Partial Credit (2PPC, Generalized Partial Credit)[8]

The 2PPC (Andrich, 1978; GPCM described in Muraki, 1997) is used to calibrate and analyze GR, SR, and ER items.[9] In this model, the probability that a student with an achievement estimate θ responds at the *k*-th level (i.e., category) of the *j*-th item is

$$P_i(\theta) = P\left(x_j = k - 1 \middle| \theta\right) = \frac{\exp Z_{jk}}{\sum_{i=1}^{m_j} \exp Z_{ji}}, k = 1 \ldots m_j,$$

---

[8] The 2PPC and GPCM are equivalent (Ercikan, Schwarz, Julian, Berket, Weber & Link, 1998).
[9] Gridded-response items are calibrated in MULTILOG using the 2 parameter logistic model. The resulting parameters are converted to the 2PPC model for all analysis and scoring activities.

where

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

FCAT uses a special case of the 2PPC that makes the following constraints:

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

and

$$C_{jk} = -\sum \gamma_{ji}\text{ , where } \gamma_{j0} = 0\,,$$

where A_j is a polynomial function of item discrimination applied to each score category and γ_{ji} is a pseudo-location parameter for each category.

The first constraint implies that higher item scores reflect higher ability levels and items can vary in their discriminations. Each item has m_j-1 independent, γ_{ji} parameters, and one a_j parameter. A total of m_j independent item parameters is estimated.

Item Response Theory Framework

FCAT scoring is built on IRT. In essence, IRT assumes that test-item responses by students are the result of underlying levels of achievement possessed by those students. IRT algorithms search for "item parameters" which capture a nonlinear relationship between achievement and the likelihood of correctly answering each item. Items that fit the IRT model will exhibit a pattern of lower probabilities of correct responses from low-ability students to higher probabilities of correct responses from high-ability students. This is reflected in an "item characteristic curve," or ICC, as depicted in Figure 1 for a multiple-choice item.

Items vary in difficulty such that the position of the point of inflection is higher or lower (i.e., to the right or to the left) along the achievement scale. For example, the point of inflection of the curve for the sample item in Figure 1 is centered at zero, which is the mean on the achievement index. An efficient test is composed of items with test characteristics similar to those depicted, but with varying difficulties ("B" parameter) that discriminate achievement along the entire achievement scale, which is typically called "theta." Item characteristic curves also differ in their lower asymptotes (related to how easy it is to get the item correct by guessing, "C" parameter) and the gradient of their slopes at the inflection point ("A" parameter).

While IRT modeling of PTs is conceptually similar, these items require a more complex mathematical treatment. In the end, however, IRT modeling of a PT captures the expected number of points that students should achieve on the performance task, depending on their achievement level. The resulting curves are similar to those shown in Figures 1 and 2, where the *y*-axis represents the probability of correct response.

The 3PL model (Lord & Novick, 1968) is used to process MC items, and the two-parameter partial credit (2PPC) model (Muraki, 1992) is used to process PT items. Figure 1 depicts an item characteristic curve using the 3PL model. For the PT items, student scores could fall into any of several different score categories (i.e., 0, 1, or 2 for SR 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. Figure 2 depicts the probabilities of a correct answer for an SR item. *FCAT 2006 Test Construction Specifications* (FDOE, 2005) presents the technical details of these models more fully.

Gridded-response items receive a hybrid treatment. Initially, item parameters are computed using a two-parameter logistic model (2PL). Then they are converted to the 2PPC for subsequent processing and maintenance in the item data bank.[10] Parameter estimation for FCAT in initial years used an IRT computer program that would not treat dichotomous (GR) items as 2PPC. In order to use the program, psychometricians used the two-parameter logistic model and then converted to the 2PPC metric to make the parameters more comparable with those calculated for PT items. Parameters are more easily interpreted and processed when all constructed-response items in the item bank are in the same metric.

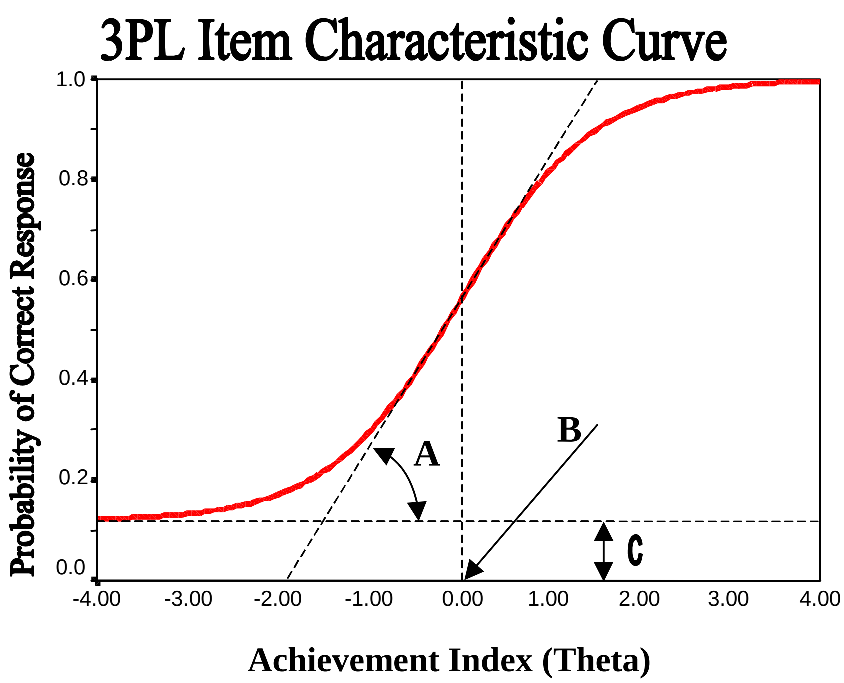

Figure 1. Item Characteristic Curve based on the three-parameter logistic trace line.

[10] Young, M.J. & Yoon, B. (1998, April). Estimating the consistency and accuracy of classifications in a standards-referenced assessment. CSE Technical Report 475. Los Angeles, CA: Center for the Study.

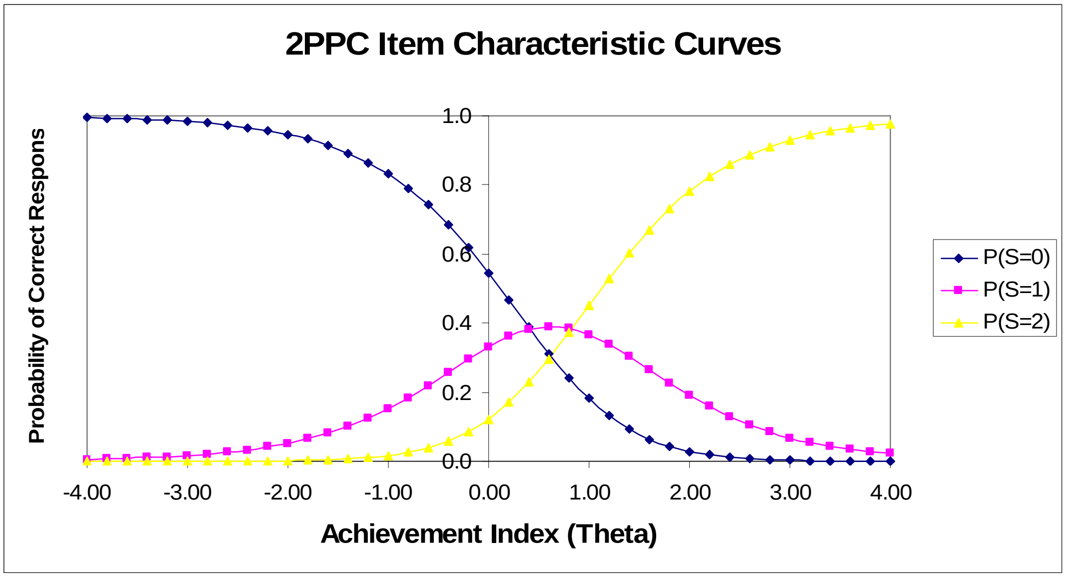

Figure 2. Probability of receiving a correct answer for a short-response item.

IRT item parameters provide the means for assigning achievement scores to individual students. Because the item parameters represent response probabilities, each student's achievement score is assigned as the level of achievement most likely to have created that student's observed responses.[11] Use of the sophisticated IRT model is advantageous for continuous testing programs such as the FCAT program, which must create a stable achievement scoring system, given the reality that items included on the tests change from one year to the next.

IRT Results

Distributions of the three 3PL item parameters are presented in Tables 56 and 57 for MC items. IRT parameters for every core and anchor item are presented in Appendix A (Reading) and Appendix B (Mathematics). The parameters are in the IRT traditional metric[12] and the achievement scale can be interpreted as a standard scale with a true score mean of 0 and standard deviation of 1. The "A" parameter indicates the slope of the curve. The steeper the slope (the larger the "A"), the more the item contributes to the estimation of achievement scores. "A" is similar to item-total correlation. For reference, the "A" for the sample curve in Figure 1 is 1.1. Items with lower slopes are useful, as long as there are enough items.

Tables 56 and 57 show that the "A" parameters are centered from 0.657 to 0.949 for reading and from 0.792 to 0.942 for mathematics. The "B" parameter indicates the difficulty of the items by indicating where the item slope at the point of inflection is centered along the achievement scale. "B" is conceptually similar to an item's *p*-value. For reference, the "B" in Figure 1 is set at 0, indicating that the curve is centered at the population mean. The "B" parameters should be spread

[11] Scores are calculated using maximum likelihood estimation.
[12] A, B, and C are reported, where $P(\theta) = C + (1-C)/(1+ \exp(-1.7A(\theta-B)))$ (Lord & Novick, 1968).

across a wide range of achievement to accurately measure students at all levels of ability; that is, because of the way the curve flattens on the ends, an item centered in the middle of the achievement scale functions well only for students in the center of the achievement distribution. Items with higher and lower "B" parameters help to measure achievement for students in the upper and lower ends of the achievement distribution. Most students score toward the center of the distribution (near the mean, 0), and Tables 56 and 57 show that the preponderance of items have "B" parameters that are within one standard deviation of the mean. Because item information[13] is the highest at the point of the item "B" parameter, the test is most reliable where the majority of the students score. Reliability is not as strong toward the ends of the distributions or for very high- or low-ability students; however, "B" parameters are well represented for the range at which the cutpoints for FCAT are set. Cutpoints are the points that separate the FCAT performance levels (1–5). This report contains a later discussion of classification accuracy and consistency at the cutpoints.

The 3PL "C" parameter, called the "pseudo-guessing" parameter, is a measure of the likelihood guessing was involved in obtaining a correct answer to the item; that is, it estimates the extent to which examinees are likely to not know the answer and still get the item correct. Notice in Figure 1 that the curve asymptotes at a lower value of about 0.2. For MC items with four possible responses, without knowing anything about the item content, the chances of responding correctly are about one in four. Typically, "C" values should be around 0.2. Well-designed items have distracters that are very attractive to those with limited skills and have no knowledge of the correct answer. For this reason, the "C" parameter is sometimes referred to as pseudo-chance and this aspect of test design results in low "C" values for these items. Higher values may signal poorly functioning distracters or some unusual curriculum emphasis in portions of the state. Tables 56 and 57 show that median "C" parameters tend to fall within the expected range.

---

[13] See the section entitled *Test Reliability, Standard Error of Measurement, and Information*, page 47, for a more detailed discussion of "information."

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

Grade (No. of MC Items)	Parameter	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	A	0.261	0.792	0.949	1.087	1.484
(45)	B	-2.556	-1.042	-0.480	0.004	0.649
	C	0.035	0.106	0.177	0.249	0.444
4	A	0.290	0.578	0.727	0.920	1.394
(41)	B	-3.344	-1.212	-0.516	-0.899	0.986
	C	0.073	0.176	0.216	0.279	0.422
5	A	0.210	0.612	0.746	0.967	1.333
(45)	B	-4.246	-1.363	-0.691	0.184	1.316
	C	0.059	0.117	0.213	0.295	0.479
6	A	0.391	0.660	0.802	0.968	1.456
(45)	B	-2.100	-0.911	-0.272	0.237	1.116
	C	0.047	0.117	0.165	0.217	0.356
7	A	0.456	0.671	0.791	0.983	1.303
(45)	B	-2.266	-1.081	-0.562	0.024	1.681
	C	0.037	0.116	0.172	0.229	0.363
8	A	0.258	0.524	0.727	0.831	1.177
(41)	B	-3.512	-1.175	-0.480	0.190	0.929
	C	0.047	0.126	0.181	0.241	0.357
9	A	0.329	0.654	0.767	0.928	1.272
(45)	B	-1.995	-,677	-0.323	0.364	0.843
	C	0.034	0.087	0.164	0.222	0.429
10	A	0.388	0.534	0.657	0.815	1.173
(41)	B	-2.483	-1.431	-0.531	-0.055	1.196
	C	0.053	0.095	0.139	0.213	0.380

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

Grade (No. of MC Items)	Parameter	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	A	0.513	0.790	0.909	1.072	1.429
(40)	B	-2.004	-1.011	-0.340	0.159	1.011
	C	0.034	0.094	0.152	0.245	0.562
4	A	0.530	0.692	0.792	0.976	1.491
(39)	B	-2.640	-1.073	-0.212	0.276	0.802
	C	0.063	0.143	0.190	0.251	0.404
5	A	0.412	0.742	0.941	1.075	1.419
(33)	B	-3.033	-0.663	0.057	0.550	0.989
	C	0.064	0.109	0.172	0.231	0.437
6	A	0.407	0.725	0.938	1.164	1.586
(33)	B	-2.545	-0.475	0.077	0.651	1.301
	C	0.060	0.122	0.187	0.237	0.316
7	A	0.556	0.731	0.942	1.120	2.387
(32)	B	-2.004	-0.346	0.200	0.646	1.253
	C	0.034	0.147	0.191	0.273	0.369
8	A	0.334	0.711	0.892	1.182	1.573
(30)	B	-3.416	-0.166	0.285	0.497	1.535
	C	0.051	0.150	0.213	0.248	0.339
9	A	0.367	0.832	0.918	1.161	1.712
(29)	B	-1.918	-0.346	0.127	0.552	1.569
	C	0.053	0.126	0.183	0.245	0.362
10	A	0.359	0.685	0.797	1.008	1.709
(28)	B	-2.600	-0.693	0.103	0.704	1.423
	C	0.033	0.105	0.184	0.287	0.437

Test characteristic curves (TCCs) were plotted using item parameters from each grade/subject test. In other words, ICCs for all items were summarized into one curve, a TCC. The results for each grade/subject are shown in Figure 3 (Reading) and Figure 4 (Mathematics). Achievement (*x*-axis) was transformed to the 100–500 scale (see next section "Scale Conversion and Test Equating"). The vertical lines on each graph mark the cutpoints for the five performance levels.

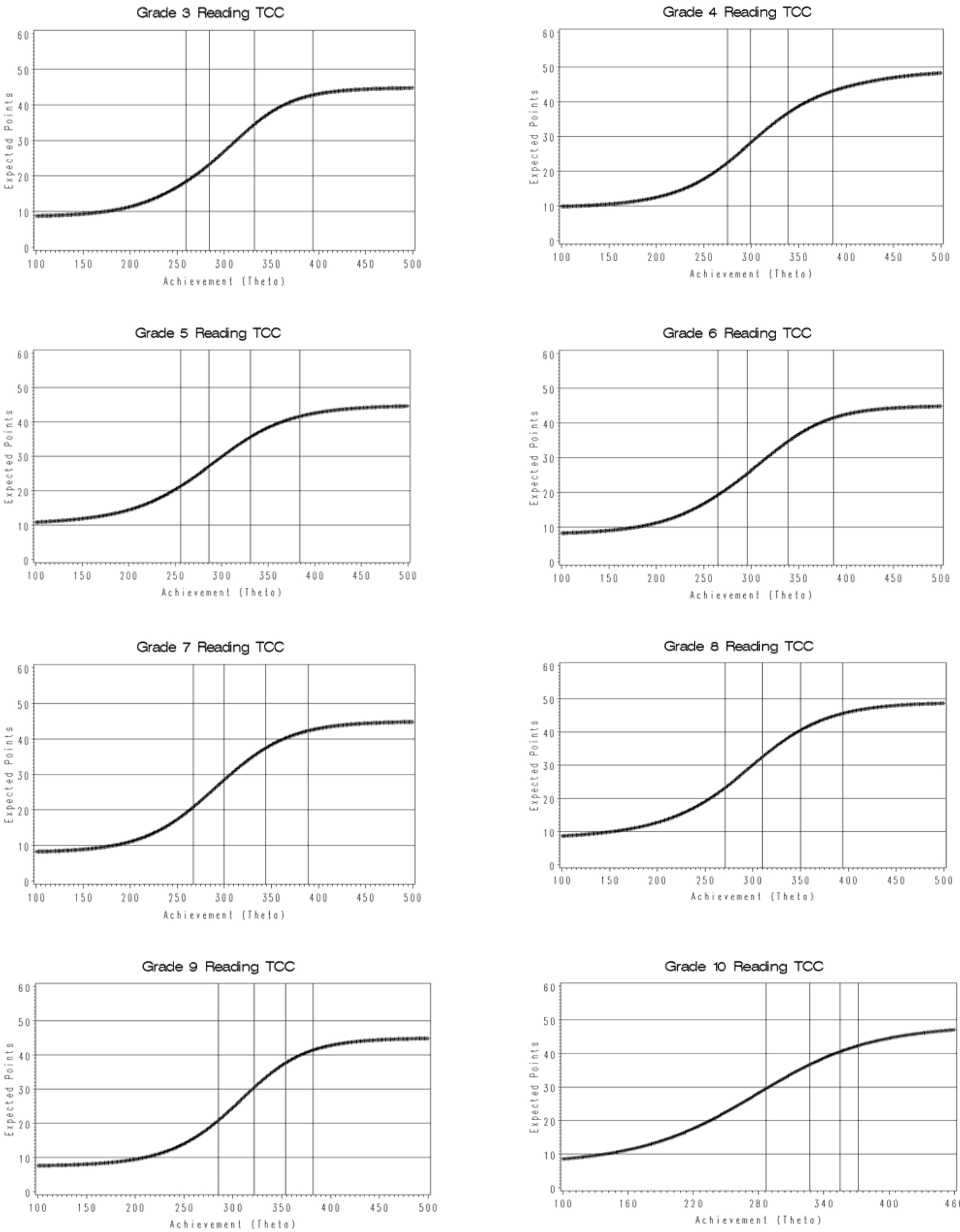

Figure 3. Test characteristic curves (TCCs) for FCAT Reading by grade.

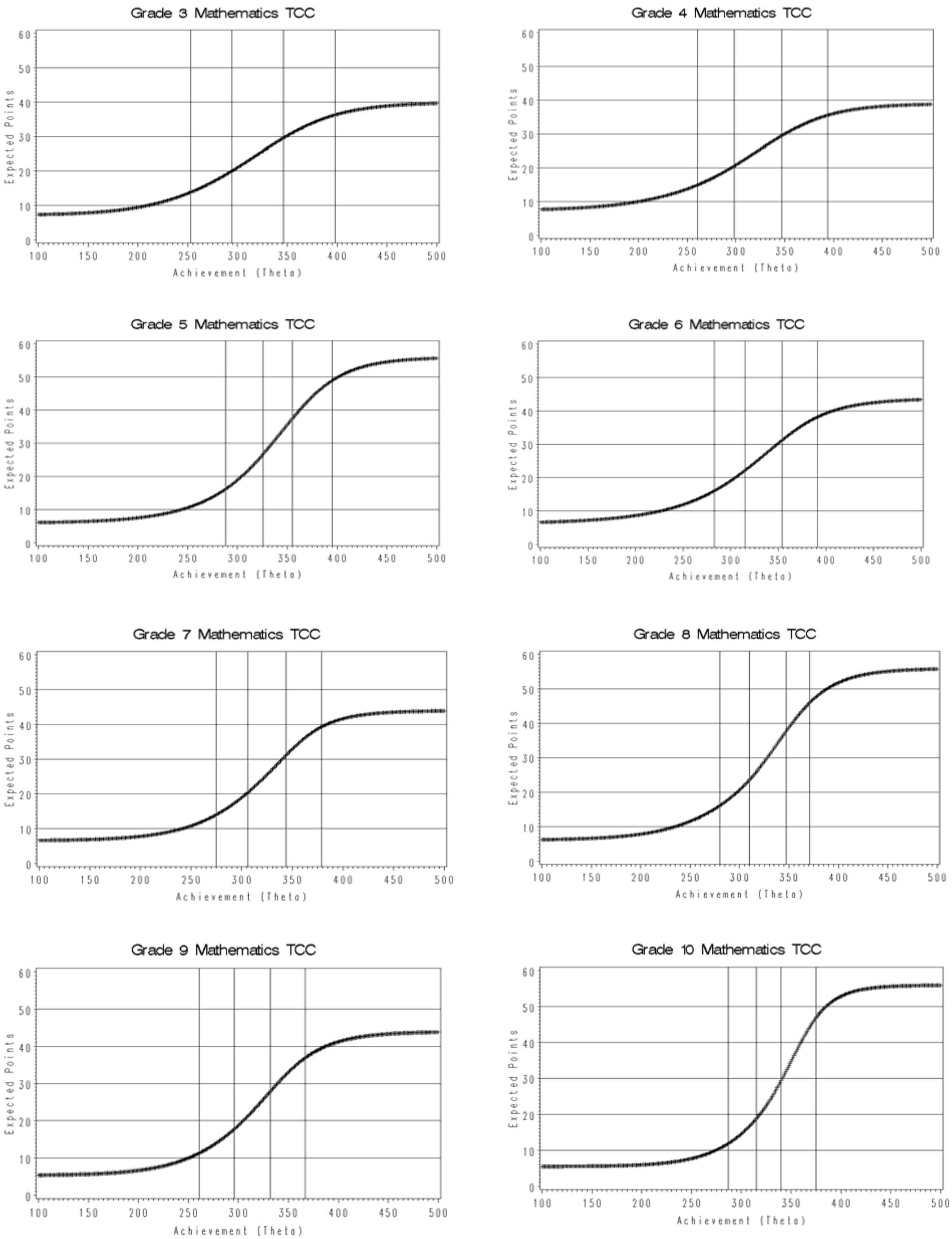

Figure 4. Test characteristic curves (TCCs) for FCAT Mathematics by grade.

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

Table 58. "A" Parameter Summary Data—Gridded-Response and Performance Task Items

Grade	Item Type (No of Items)	Minimum	25th Percentile	Median	75th Percentile	Maximum
Reading						
4	SR (3)	1.012	1.012	1.156	1.347	1.347
	ER (1)			1.020		
8	SR (3)	0.788	0.788	1.026	1.339	1.339
	ER (1)			0.700		
10	SR (3)	0.670	0.670	0.810	0.936	0.936
	ER (1)			1.076		
Mathematics						
5	GR (11)	1.006	1.051	1.315	1.573	1.982
	SR (4)	0.919	0.976	1.040	1.237	1.427
	ER (2)	0.811				1.044
6	GR (11)	0.628	1.000	1.567	1.868	2.692
7	GR (12)	1.018	1.338	1.584	2.093	2.417
8	GR (14)	0.991	1.078	1.752	2.266	2.861
	SR (4)	1.222	1.276	1.351	1.442	1.512
	ER (2)	1.190				1.447
9	GR (15)	0.973	1.145	1.492	1.887	2.358
10	GR (16)	0.978	1.330	1.501	1.901	2.370
	SR (4)	1.097	1.311	1.548	1.728	1.885
	ER (2)	1.035				1.415

Scale Conversion and Test Equating

IRT scaling produces item parameters for an achievement scale targeted to a true score mean of 0 and true score standard deviation of 1. For the FCAT, however, scores are reported on a 100–500 scale; therefore, a transformation is needed for the IRT item parameters in order for them to produce the appropriate scores.

In addition to the need for students' scores to be placed on the FCAT scale, there is also the need for those scores to be comparable to scores from past years. Even though students are expected to perform differently (presumably better) than students in previous years. To report scores in 2006 on the FCAT 100–500 scale and make them comparable to scores from past years, the data output by IRT processing needs to be equated. This equating process involves (a) repeating "anchor items" in the 2006 test that had been used in previous FCAT administrations, and (b) applying the Stocking/Lord (1983) procedure using parameters from those anchor items to adjust for the difference between students in 2006 and previous years. The anchor items and the Stocking/Lord procedure are used to equate 2006 test scores to the test scores originally reported.[14] This procedure, using different anchor items, has been conducted every year since 1998.

With the completion of the 2006 scaling, the anchor items have two sets of item parameters: (a) new parameters on the mean equal to 0 and standard deviation equal to 1 scale produced this year, and (b) old parameters that were transformed during their previous use. The old parameters are based on either the original 1998 scale or the 2001 scale. The Stocking/Lord (1983) procedure uses the old item parameters to locate the achievement scale and then searches for a transformation multiplier and additive constant that can combine to make the new parameters replicate the original achievement scale as closely as possible. This is done by attempting to match test characteristic curves (TCC), which are summations of item characteristic curves (ICC) (see Figure 1), produced by the old parameters with test characteristic curves produced by transformations of new parameters. Since the items are the same, the same scale should result.

During this equating process, item-level reviews are conducted. Specifically, item parameter estimates are reviewed for their stability before they are included in the equating process. A tool used to evaluate anchor parameter differences is a computational procedure that produces a metric indicating the difference between the shapes of the item characteristic curves produced by the current item parameters versus base-year item parameters (i.e., parameters that are equated to the base scale in the most recent administration of the items). This metric takes all item parameters into account. The procedure checks for outlier items by computing differences in response probabilities based on base-year and current-year parameter values. The items with the largest differences are identified for further review and possible elimination from equating. In order to calculate the differences, anchor parameters are placed on the current year's IRT scale. Then, values of the squared differences are calculated at 31 quadrature points (the same that are used in the Stocking/Lording procedure), and the mean of the 31 squared differences is computed for each item. Items are flagged if their mean squared difference is greater than expected, given the mean squared differences of all items. A summary log of the anchor item-level analysis can be found in Appendix C. If a particular item parameter is too low, too high, or at variance with prior parameter estimates, then the FDOE personnel make a decision as to whether the item should remain in the anchor set.

This year, numerous items intended for linking were dropped from the equating process for FCAT Reading (in Grades 6, 7, and 10). On the Grade 6 Reading test, all of the anchor items on Form 28 were dropped and replaced with back-up anchor items on Form 30. Item 15, Form 29 was dropped from the Grade 7 Reading test and all of the Form 29 items on the Grade 10 Reading test were dropped. These items were dropped because changes were made in the position of the reading passages on the test forms. In other words, their position on the 2006 test was different from previous usages and this affected students' performance. For FCAT Mathematics, only one item

[14] The FCAT became operational for Grades 4, 8, and 10 in 1998 and Grades 3, 5, 6, 7, and 9 in 2001.

was dropped from all grades: Item 6, Form 30 on the Grade 3 test. Statistical data for these dropped items are found in Appendix A.[15]

Another method used to compare old and new item parameter differences is to plot two ICCs for each anchor item; one plot is created by using the previous year's parameters, and the second is created using the current year's parameters [the probability of answering correctly is plotted on the *y*-axis, and the achievement index (theta) is plotted on the *x*-axis]. This way, the two ICCs can be compared visually. This technique adds another useful decision-making tool to those that are already in place. Figure 5 shows a comparison of two different plots: the Example A plot shows that there was little change in the way students responded to this particular question from its previous usage (ICC labeled "Old") to its current usage (ICC labeled "New"). Example B, however, shows divergence between the two ICCs, and they converge at about 1 standard deviation above the mean (0). When an anchor item shows this type of divergence, it is advisable that FDOE content experts examine the item by asking questions, such as was there a misprint in the test booklet? FDOE content experts should then make a decision as to whether the item should be included as an anchor.

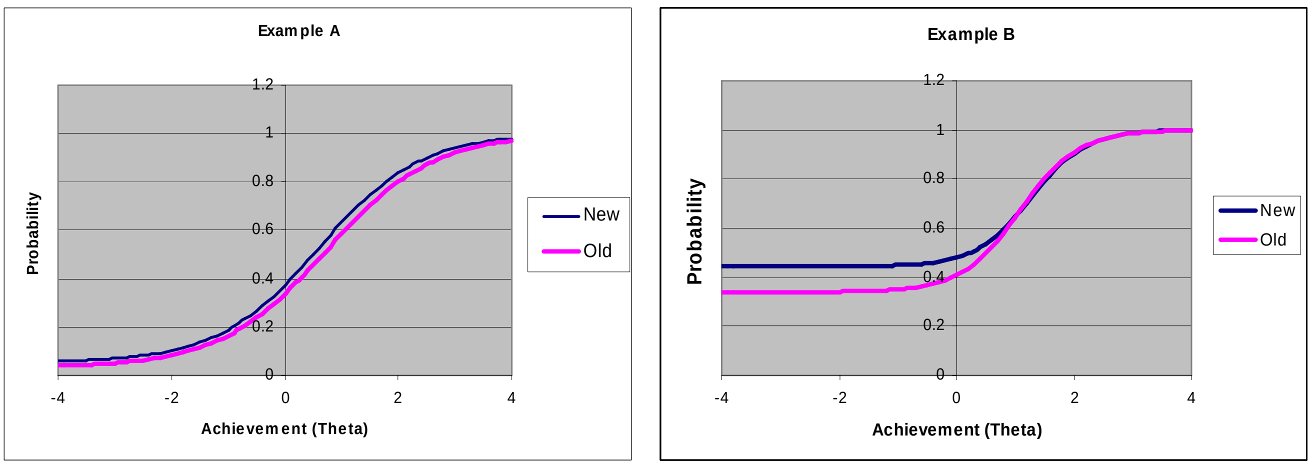

Figure 5. Sample ICC plots used to examine anchor item behavior from year to year.

Table 59 shows the item type and number of anchor parameters used in equating, along with the transformation constants [M1 (Slope) and M2 (Intercept)] that were derived to replicate the base year FCAT scale. The M2 additive constant projects the change in average true score level expected for standard curriculum students. Thus, while an average standard curriculum student would be expected to have a score of 300 for Grade 4 Reading in 1998, the average standard curriculum student in 2006 would be expected to have a score of approximately 324 (the value of M2 for Grade 4 Reading).

---

[15] Dropped items are marked with an asterisk (*) in the data tables.

Table 59. Equating Multiplicative and Additive Constants

Grade	Anchor Item Type and Number	M1 Multiplier	M2—Additive Constant
Reading			
3	21 MC	42.958	322.215
4	18 MC	40.286	323.679
5	19 MC	42.639	315.236
6	24 MC	43.896	315.000
7	22 MC	42.610	314.296
8	24 MC	39.330	306.957
9	23 MC	39.544	315.899
10	21 MC	46.666	309.782
Mathematics			
3	22 MC	52.686	332.424
4	23 MC	47.265	325.676
5	15 MC, 8 GR	39.320	337.462
6	19 MC, 8 GR	48.598	319.865
7	16 MC, 11 GR	42.231	313.921
8	13 MC, 8 GR	37.547	322.167
9	15 MC, 12 GR	40.397	311.862
10	13 MC, 12 GR	30.134	332.819

Anchor items should have as many of the relevant characteristics, to the extent possible, as the core items. Several statistical comparisons were done to examine this issue. First, a comparison of the mean proportion correct was calculated (i.e., the mean for core items answered correctly compared to the mean for anchor items answered correctly). For Reading, Table 1c in Appendix A shows that the largest difference between core and anchor item means was in Grade 3 (approximately 9 percent), where students performed slightly better on the anchor items. For mathematics, the opposite was true in terms of the largest performance difference between core and anchor items; here, students performed slightly better on the core items. Table 1c (Appendix B) shows that the largest difference was in Grade 9, where the core items' mean *p*-value is 0.517, and the anchor items' mean *p*-value is 0.447 (7 percent lower). Another statistic used to compare anchor and core item behavior is seen in Table 1d in Appendices A and B: mean points scored for core items versus anchor items. Total points from anchor items should be at least 20 percent of the total points scored on the core test. This was true for all grade/subject tests.

Anchor mapping statistics are also found in Table 1e in Appendices A and B. For each grade, the tables list the median position of anchor items in 2006 and their median position during previous usage. A rank-order correlation coefficient (*r*) shows the degree of agreement between item positions from year to year. This year's mathematics anchor items, for all grades, were in close proximity to the previous year's test position. Reading tests showed more variability, which is expected to a certain degree because reading tests are passage dependent.

Two additional tables of information are provided that (a) present comparisons of the percent of core versus anchor items by content category (Table 1f), and (b) provide comparisons of core and anchor items by item type (i.e., MC, GR, SR, and ER). Anchor items for reading are all multiple-choice, but for mathematics, anchor items can be either MC or GR. These item type comparisons are found in Table 1g in each appendix (Appendix A for Reading and Appendix B for Mathematics).

IRT Fit Statistics

As previously explained, IRT scaling algorithms attempt to find item parameters (numerical characteristics) that create a match between observed patterns of item responses and theoretical response patterns defined by the selected IRT models. The Q1 statistic (Yen, 1981) may be used as an index for finding how well theoretical item curves match observed item responses. Q1 is computed by first conducting an IRT item-parameter estimation, then by estimating students' achievement using the estimated item parameters, and lastly, by using students' achievement scores in combination with estimated item parameters to compute expected performance on each item. Differences between expected item performance and observed item performance are then compared at selected intervals across the range of student achievement. Q1 is computed as a ratio involving expected and observed item performance and is, therefore, interpretable as a chi-square (χ^2) statistic, which is a statistical test that determines whether the data (observed item performance) fit the hypothesis (the expected item performance).

Q1 for each item type has varying degrees of freedom because the different types of items have different numbers of IRT parameters; therefore, Q1 is not directly comparable across item types. An adjustment or linear transformation (translation to a z-score, Z_{Q1}) is made for different numbers of item parameters and sample size to create a more comparable statistic.

Q1 can be expressed as

$$Q_{1j} = \sum_{i=1}^{I} \frac{N_{ji}\left(O_{ji} - E_{ji}\right)^2}{E_{ji}\left(1 - E_{ji}\right)},$$

where N_{ji} is the number of examinees in cell *i* for item *j*; O_{ji} and E_{ji} are the observed and predicted proportions of examinees in cell *i* that pass item *j*:

$$E_{ji} = \frac{1}{N_{ji}} \sum_{aei}^{N_{ji}} P_j\left(\hat{\theta}_a\right).$$

The generalization of Q1, or Generalized Q1, for items with multiple response categories is

$${}_{gen}Q_{1j} = \sum_{i=1}^{I} \sum_{k=1}^{N_{ji}} \frac{N_{ji}\left(O_{jki} - E_{jki}\right)^2}{E_{jki}}$$

with

$$E_{ji} = \frac{1}{N_{ji}} \sum_{aei}^{N_{ji}} P_{jk}\left(\hat{\theta}_a\right).$$

Both the Q1 and Generalized Q1 results are transformed into the statistic ZQ, and are compared to a criterion, ZQ_{crit}, to determine acceptable fit.

$$ZQ > \frac{Q - df}{\sqrt{2df}}$$

and

$$ZQ_{crit} > \frac{N}{1500} * 4,$$

where Q is either Q1 or Generalized Q1 and *df* is the degrees of freedom for the statistic (df = 10 – number of parameters estimated). Poor fit is indicated where ZQ is greater than ZQ_{crit}.

Q1, for each item type, has varying degrees of freedom because the different types of items have different numbers of IRT parameters; therefore, Q1 is not directly comparable across item types. An adjustment, or linear transformation (translation to a *z*-score, Z_{Q1}), is made for different numbers of item parameters and sample sizes to create a more comparable statistic. The FCAT has set criteria for a minimum Z_{Q1} value standard for an item to have acceptable fit (FDOE, 1998).[3] Complete item-specific Q1 results are in the Appendices. Tables 60 and 61 present the distributions of Z_{Q1} for reading and mathematics items, respectively. Table 62 presents the number of poorly fitting items by item type. For MC items, the low number of poorly fitting items is consistent with previously reported patterns of strong point-biserials and strong "A" parameters; however, for SR, ER, and GR items, the number of items with poorly fit statistics has decreased from last year. In 2005, there were 36 poorly fitting items in mathematics (across all grades), and in 2006, there are only 5 such items.

Table 60. Z_{Q1} Statistic, Summary Data—All Reading Items

Grade	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	-0.831	0.742	1.904	3.668	8.498
4	-0.924	0.142	1.444	3.105	19.853
5	-0.986	0.760	1.453	2.567	16.146
6	-1.242	0.256	0.787	1.628	3.566
7	-0.961	0.178	1.286	2.242	18.558
8	-1.371	0.100	0.692	1.551	14.707
9	-0.938	-0.046	0.801	2.259	15.217
10	-0.689	0.445	0.752	2.117	8.372

[3] If Z_{Q1} > (sample size • 4)/1500, then fit is rated as "poor."

Table 61. Z_{Q1} Statistic, Summary Data—All Mathematics Items

Grade	Minimum	25th Percentile	Median	75th Percentile	Maximum
3	-1.086	0.094	1.321	2.734	6.989
4	-1.059	-0.064	1.037	1.578	3.426
5	-1.152	0.081	0.820	2.724	15.159
6	-1.118	-0.177	0.886	2.785	10.192
7	-0.748	0.246	0.931	3.026	25.266
8	-1.023	-0.146	1.379	3.777	15.317
9	-0.908	0.140	1.323	4.460	34.344
10	-1.109	0.285	2.049	4.008	22.714

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

	Reading			Mathematics			
Grade	MC	SR	ER	MC	GR	SR	ER
3	0/45			0/40			
4	0/41	0/3	0/1	0/39			
5	0/45			0/33	0/11	0/4	0/2
6	0/45			0/33	0/11		
7	0/45			1/32	1/12		
8	0/41	0/4	0/1	0/30	0/14	0/4	0/2
9	0/45			0/29	2/15		
10	0/41	0/3	0/1	0/28	1/16	0/4	0/2

Note: Numbers shown represent "Number of items with 'poor fit'/Total number of items."

Achievement Scale Unidimensionality

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

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

When calculating the level of local item dependence for two items (*i* and *j*), the Q3 statistic is

$$Q_3 = r_{d_i d_j},$$

a correlation between d_i and d_j values. For test-taker *k*,

$$d_{ik} = u_{ik} - P_i(\theta_k),$$

u_{ik} is the score of the k^{th} test taker on item *i* (one if correct, zero if incorrect), and $P_i(\theta_k)$ represents the probability of test-taker *k* responding correctly to item *i*.

With *n* items, there are $n(n-1)/2$ Q3 statistics. For example, Grade 3 Reading has 45 items and 990 Q3 values. The Q3 values should all be small. Q3 data are summarized in Tables 63 and 64 by minimum, 5th percentile, median, 95th percentile, and maximum values for each FCAT grade/subject combination. To add perspective to the meaning of the Q3 distributions, the average zero-order correlation (item intercorrelation) among item responses is also shown. If the achievement construct is "accounting for" the relationships among the items, Q3 values should be much smaller than the zero-order correlations. These tables indicate that for all grades/subjects, at least 90 percent (between the 5th and 95th percentile) of the items are expectedly small, showing Q3 values between -0.066 and 0.025 for both reading and mathematics. These data, coupled with the Q1 data above, indicate that the unidimensional IRT model provides a very reasonable solution for capturing the essence of student achievement defined by the carefully selected set of items for each grade and subject.

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

		Q3 Distribution				
Grade	**Average Zero-order Correlation**	**Minimum**	**5th Percentile**	**Median**	**95th Percentile**	**Maximum**
3	0.170	-0.086	-0.060	-0.021	0.023	0.232
4	0.141	-0.091	-0.056	-0.021	0.021	0.082
5	0.128	-0.108	-0.056	-0.019	0.014	0.140
6	0.156	-0.095	-0.055	-0.022	0.025	0.209
7	0.163	-0.095	-0.057	-0.022	0.022	0.072
8	0.136	-0.096	-0.056	-0.020	0.014	0.123
9	0.163	-0.088	-0.052	-0.022	0.017	0.087
10	0.134	-0.121	-0.059	-0.020	0.011	0.165

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

Grade	Average Zero-order Correlation	Q3 Distribution				
		Minimum	5th Percentile	Median	95th Percentile	Maximum
3	0.185	-0.119	-0.066	-0.022	0.013	0.139
4	0.158	-0.091	-0.059	-0.024	0.007	0.117
5	0.204	-0.103	-0.054	-0.018	0.018	0.266
6	0.180	-0.097	-0.058	-0.020	0.018	0.099
7	0.197	-0.087	-0.053	-0.020	0.011	0.127
8	0.206	-0.098	-0.052	-0.017	0.017	0.137
9	0.192	-0.105	-0.052	-0.021	0.011	0.226
10	0.210	-0.123	-0.056	-0.017	0.020	0.223

Item Bias Analyses

FCAT test items receive intensive, qualitative reviews by expert panels before being placed into field tests, including review for possible gender or ethnicity bias (FDOE, May 2002). In addition, items are examined after each use for quantitative evidence of differential performance by various subgroups of examinees representing both genders and the racial and ethnic groups whose achievement levels are assumed to be comparable. Thus, test scores for female students are compared with those for male students, test scores for African-American students are compared with those for White students, and test scores for Hispanic students are compared with those for White students.

The analyses of differential item functioning (DIF) were done using two methods that are described by Zwick, Donoghue, and Grima (1993). Both methods compare performance on each item with performance on the test as a whole. For any given Achievement Level, as defined by the FCAT scale score, performance on each item should be the same for females as for males. Similarly, at any given level of overall achievement, performance on each item should be similar for African-American or Hispanic students when compared with the White student population. The Mantel (1963) statistic [a version of the common Mantel-Haenszel (1959) statistic that accommodates PT items] is a chi-square statistic that tests the statistical significance (or probability) of differences in item performance. Using standardized mean difference (SMD) is particularly useful with the large FCAT calibration sample sizes because a statistically significant difference may appear between two groups responding to an item. That difference (reviewed by educators and policymakers), however, may not be deemed large enough to cause concern from a practical testing and decision-making perspective. For this reason, an SMD rating system was put into place (FDOE, 1998) that groups items into one of seven categories according to their demonstrated differential functioning. Items that fall into the 1, 2, or 3 categories have small SMD, therefore, they show little performance difference between the groups of interest.

Tables 65 and 66 present the distribution of SMD summary ratings. For reading, all but seven items (across all grades) are in the lowest two categories of SMD. All but three of the mathematics items fall into the two lowest SMD categories. These items had already been through a rigorous review, including field testing in previous years, so the infrequent incidence of large DIF ratings is not surprising. Mantel and SMD results for each item are presented in Appendices A and B.

Table 65. Item DIF Rating Summary—Reading

Grade	Standardized Mean Difference (SMD) Rating						
	Low					High	
	1	2	3	4	5	6	7
3	45	0	0	0	0	0	0
4	43	1	1	0	0	0	0
5	44	0	1	0	0	0	0
6	44	1	0	0	0	0	0
7	44	0	1	0	0	0	0
8	43	1	1	0	0	0	0
9	43	2	0	0	0	0	0
10	39	3	0	3	0	0	0

Table 66. Item DIF Rating Summary—Mathematics

Grade	Standardized Mean Difference (SMD) Rating						
	Low					High	
	1	2	3	4	5	6	7
3	39	1	0	0	0	0	0
4	39	0	0	0	0	0	0
5	47	2	1	0	0	0	0
6	42	2	0	0	0	0	0
7	43	1	0	0	0	0	0
8	49	1	0	0	0	0	0
9	44	0	0	0	0	0	0
10	47	1	1	1	0	0	0

Test Reliability, Standard Error of Measurement, and Information

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

Test reliability concerns the concept that a test score results from some true level of achievement plus measurement error. For a population of students, reliability is a ratio of variation in true achievement compared with variation in observed test scores. The less 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; rather, standard error of measurement (SEM) is a function of how well a student's pattern of item responses matches the expected response pattern uncovered by the IRT-modeling processes. In other words, with IRT modeling, score assignment is more accurate for a student who correctly answers the easy items and misses the difficult items than for a student who gets as many easy items correct as difficult items. Furthermore, score assignment tends to be more accurate for students toward the center of the distribution than for students with more extreme scores. Another way to determine the amount of precision in estimating achievement is to look at information. In IRT, a test's information is inversely related to SEM ($1/\sigma^2$); therefore, if the amount of information on the ability scale is large, then ability can be estimated with precision for students whose true ability is at that level (Baker, 2001).

Conditional standard error curves, depicted for FCAT Reading in Figure 6 and FCAT Mathematics in Figure 7, are used to depict test reliability. The curves plot the average SEM extracted from student score records as a function of achievement level. SEM is like a standard deviation because approximately two-thirds of the students with a given level of achievement will have observed test scores within one SEM of the given true score. For example, in Figure 6 the Grade 3 Reading SEM plots show that a student whose true achievement level is 200 will have an SEM of approximately 25. That means that approximately two-thirds of those students will have test scores between 170 and 230. The remaining one-third of the students with a true achievement level of 200 will have test scores more than 25 points away from 200. As expected, SEM is larger at the tails of the achievement level distribution and smaller in the center. Most students, however, are in the center of the distribution. Cutpoints, represented by vertical lines on each graph, are used to demarcate student performance categories (1–5). Notice that cutpoints are located in the center of the distribution where the vast majority of students fall (see Table 67).

Test information functions (TIFs), seen in Figures 8 and 9, show the amount of information as plotted on the 100–500 achievement scale. For reading, the TIFs generally peak around an achievement value of 300, but the TIFs peak slightly higher in Grades 3 and 9 than for the other grades. The peaks can be interpreted to mean that these tests estimate achievement more precisely around 300 than at other achievement levels. A flatter curve means a test estimates achievement with more equal precision across that range of achievement (such as Grade 10 Reading). For mathematics, the TIFs generally peak around an achievement value of 350 on the achievement scale. Grades 5, 7, 8, and 10 appear to contain more information between 300 and 400 on the mathematics tests achievement scale than do the reading tests. This is especially true in Grade 10.

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 68.

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 nonstandard curriculum students. This was done in order to make reliability data consistent across grades and subjects and is not confounded by any differences in calibration samples. In addition, these estimates are consistent with the application of the FCAT; they characterize test results for all students who receive scores.

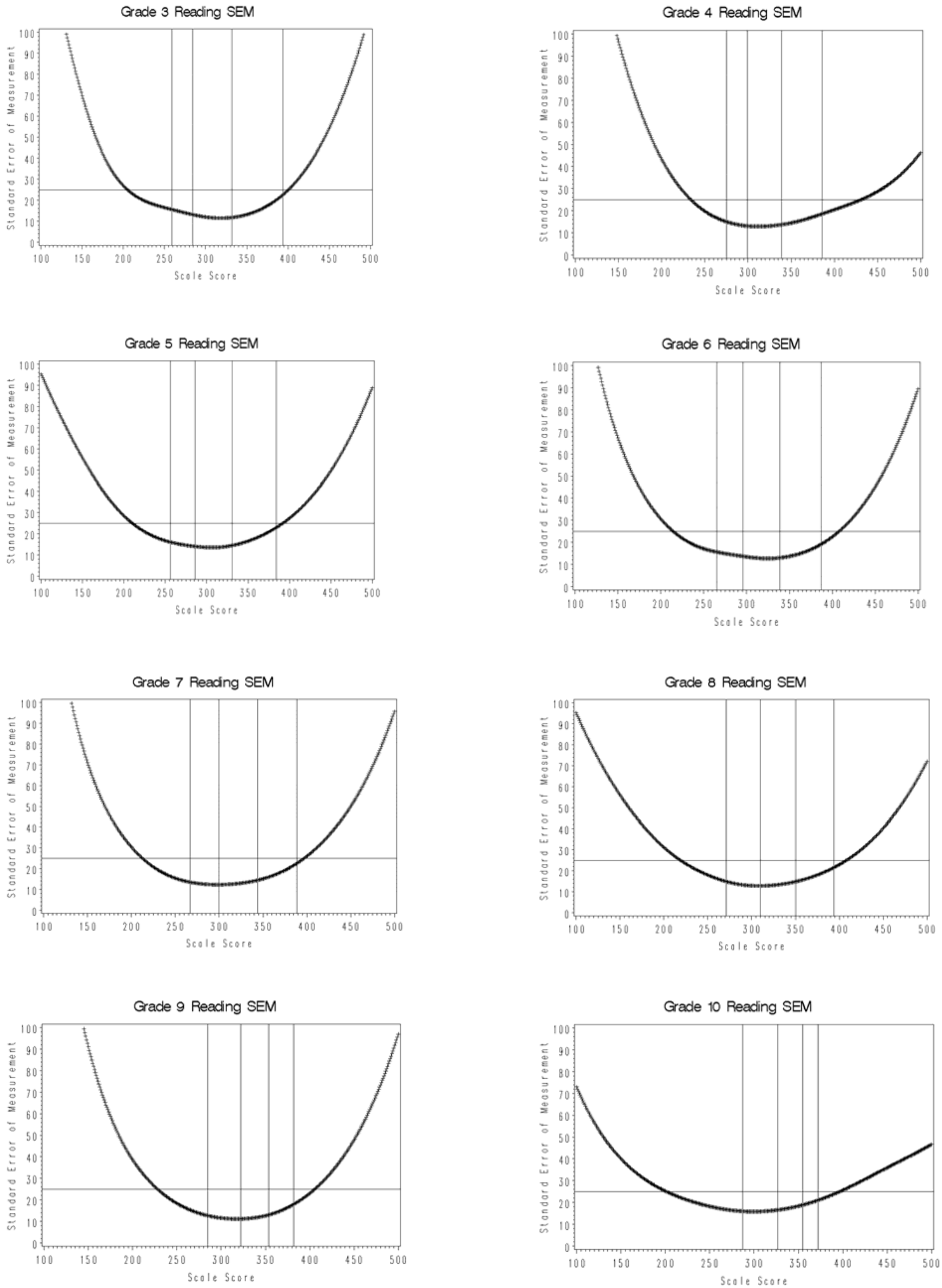

Figure 6. Standard error of measurement (SEM) plots for 2006 FCAT Reading by grade.

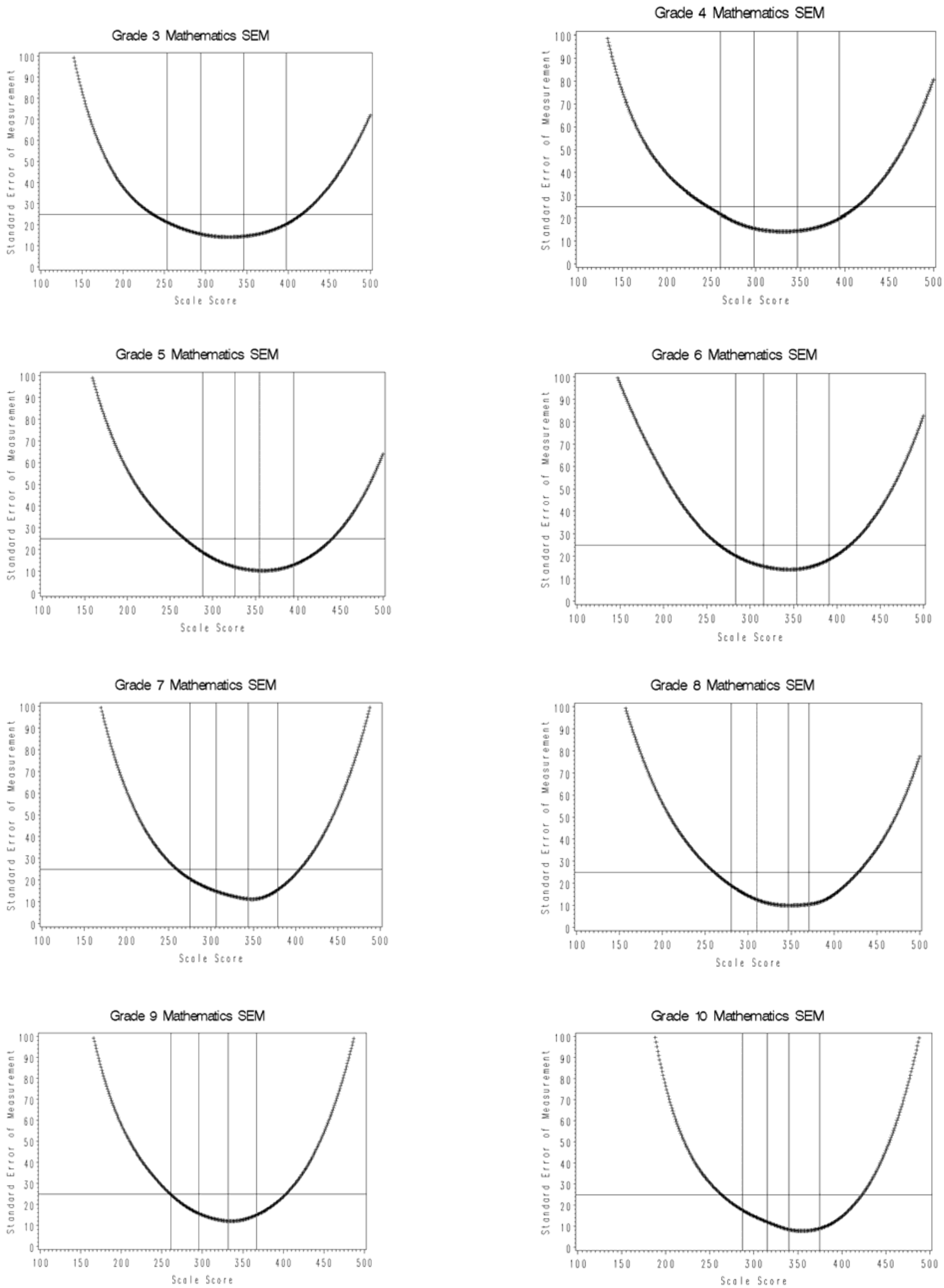

Figure 7. Standard error of measurement (SEM) plots for 2006 FCAT Mathematics by grade.

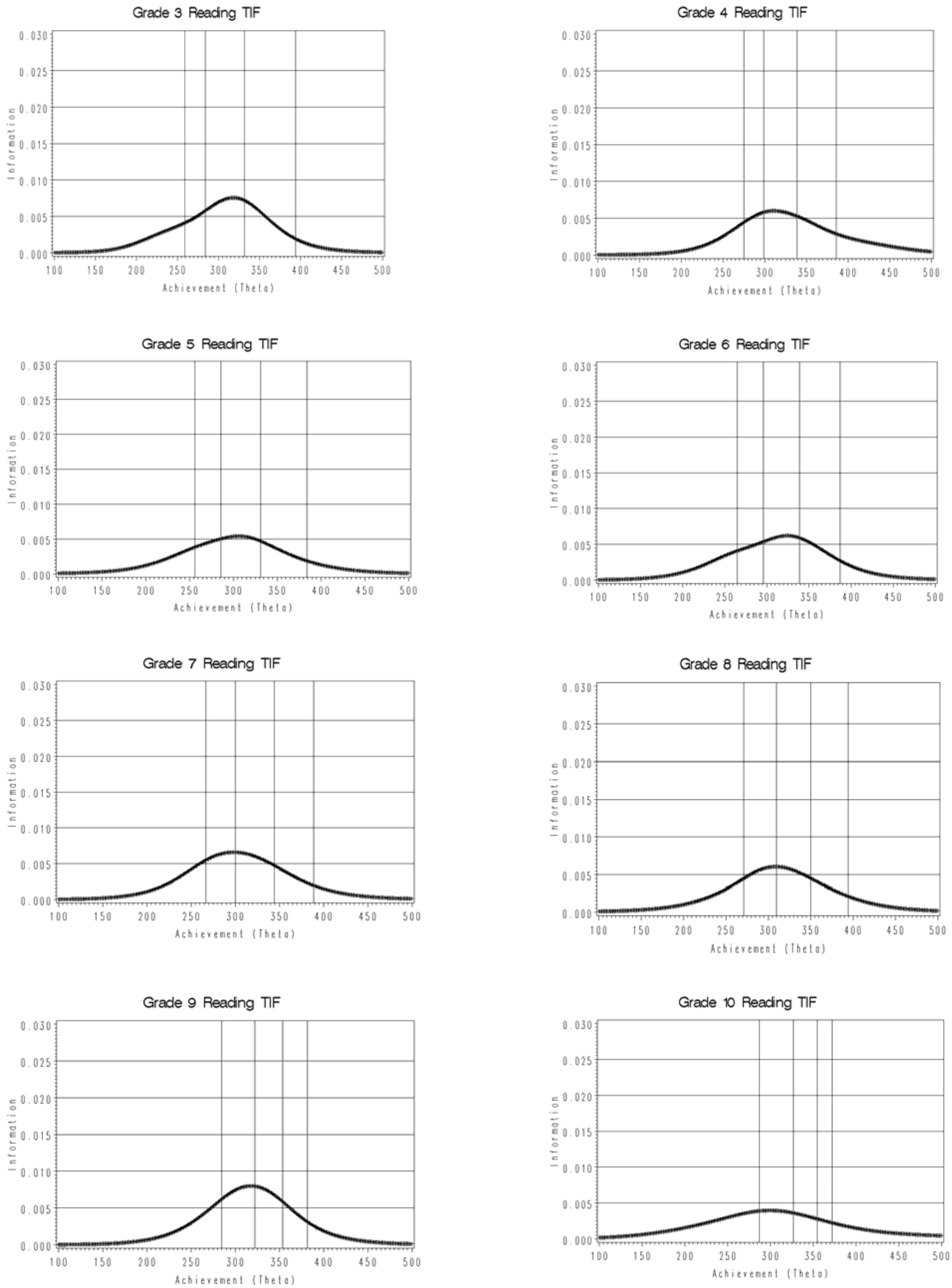

Figure 8. Test information functions (TIFs) for 2006 FCAT Reading by grade.

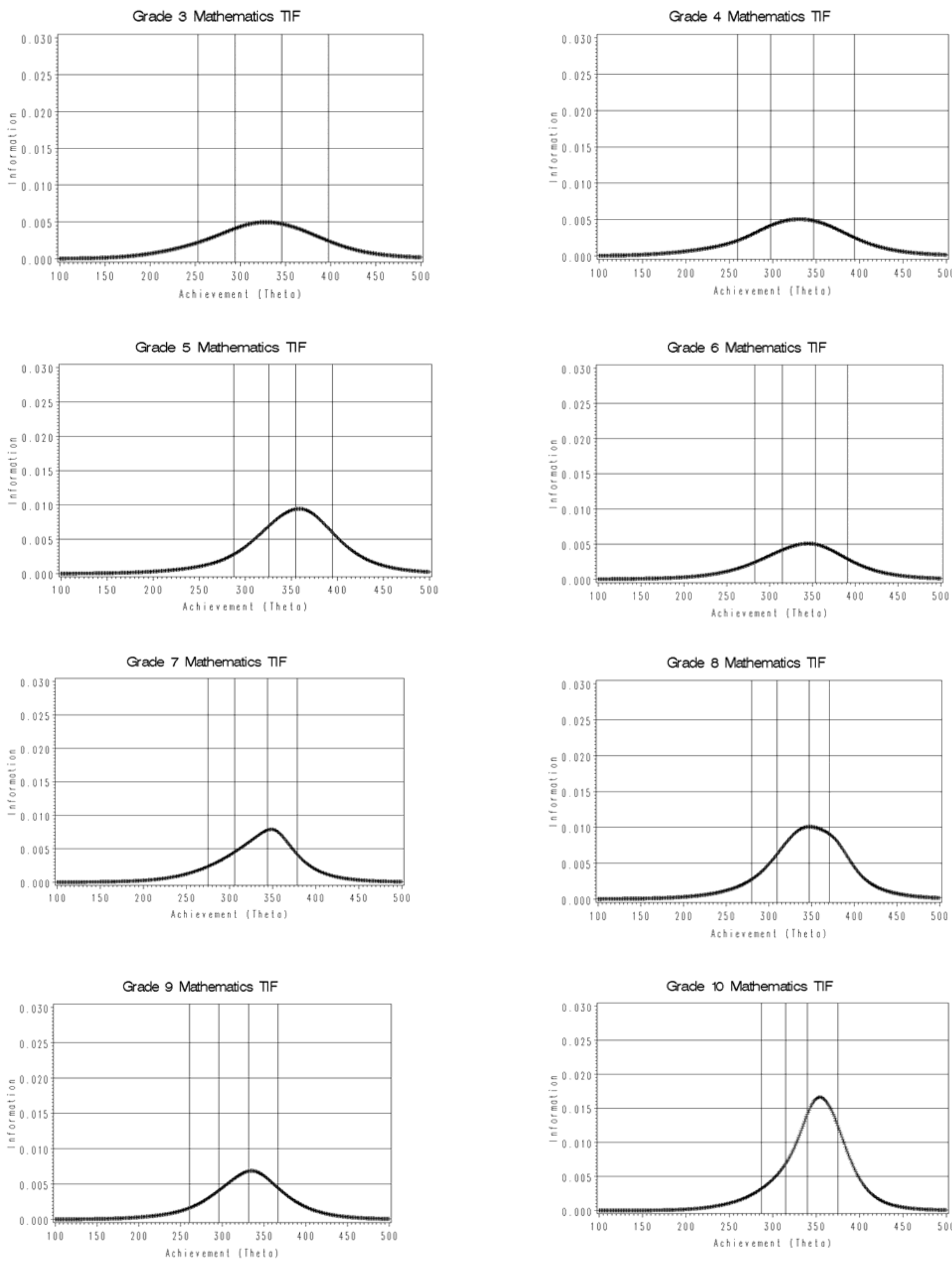

Figure 9. Test information functions (TIFs) for 2006 FCAT Mathematics by grade.

Table 67. Standard Error of Measurement (SEM) at Cutpoints for Score Categories 1–5

Grade	Reading		Mathematics	
	Cutpoint	SEM	Cutpoint	SEM
3	259	16	253	21
	284	13	294	16
	332	12	346	15
	394	23	398	20
4	275	15	260	22
	299	13	298	15
	339	14	347	14
	386	19	394	20
5	256	16	288	19
	286	14	326	12
	331	15	355	10
	384	23	395	13
6	265	16	283	20
	296	14	315	15
	339	13	354	14
	387	19	391	18
7	267	13	275	21
	300	12	306	15
	344	14	344	11
	389	23	379	16
8	271	15	280	19
	310	13	310	13
	350	15	347	10
	394	22	371	10
9	285	13	261	25
	322	11	296	16
	354	13	332	12
	382	18	367	15
10	287	16	287	18
	327	17	315	12
	355	19	340	8
	372	21	375	9
PASS (10 only)	300	16	300	15

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

Table 68 also shows traditional Cronbach's alpha reliability statistics. These estimates are based on raw scores only and have been calculated for the total set of items and for the items that comprise each of the separate reporting categories. For reading, all but Grade 8 have slightly higher marginal reliabilities than last year. For mathematics, the opposite is true. However, the reliabilities, though lower in five of the eight grades in 2006 than in 2005, are only minimally lower (the largest decrease is 0.006).

Table 68. IRT Marginal Reliabilities and Cronbach's Alpha

Reading Grade	**IRT Marginal r_{ii}**	**Cronbach's Alpha**					
		Total	**Word and Text**	**Main Idea**	**Recognizing Relationships**	**Research Reference**	
3	0.920	0.890	0.607 (7)	0.814 (22)	0.698 (12)	0.472 (4)	
4	0.915	0.853	0.366 (6)	0.711(19)	0.679 (15)	0.406 (5)	
5	0.902	0.865	0.497 (7)	0.737 (17)	0.674 (15)	0.413 (6)	
6	0.928	0.891	0.649 (11)	0.710 (15)	0.693 (11)	0.602 (8)	
7	0.919	0.895	0.567 (7)	0.780 (20)	0.655 (9)	0.610 (9)	
8	0.910	0.853	0.473 (6)	0.637 (18)	0.552 (8)	0.651 (13)	
9	0.922	0.896	0.549(4)	0.785 (20)	0.625 (10)	0.696 (11)	
10	0.916	0.852	0.405 (6)	0.584 (15)	0.640 (12)	0.653 (12)	
Mathematics Grade	**IRT Marginal r_{ii}**	**Total**	**Number Sense, Concepts, Operations**	**Measure-ment**	**Geometry and Spatial Sense**	**Algebraic Thinking**	**Data Analysis/ Probability**
3	0.927	0.900	0.736 (12)	0.682 (8)	0.507 (7)	0.592 (6)	0.668 (7)
4	0.923	0.880	0.651 (10)	0.607 (8)	0.524 (7)	0.605 (7)	0.605 (7)
5	0.947	0.873	0.641 (12)	0.629 (11)	0.479 (9)	0.442 (10)	0.635 (8)
6	0.935	0.862	0.604 (9)	0.564 (9)	0.374 (9)	0.541 (8)	0.623 (9)
7	0.938	0.862	0.548 (9)	0.602 (9)	0.550 (8)	0.504 (9)	0.583 (9)
8	0.947	0.885	0.552 (12)	0.645 (11)	0.536 (8)	0.647 (10)	0.559 (9)
9	0.940	0.845	0.502 (8)	0.515 (7)	0.516 (11)	0.616 (10)	0.410 (8)
10	0.949	0.882	0.522 (11)	0.420 (9)	0.670 (10)	0.620 (12)	0.667 (8)

Note: Numbers in parentheses are the number of items per category.

Intercorrelations among Reporting Categories and Scale Scores

Tables 69–84 present intercorrelations among the 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 (0.91 to 0.98). Comparisons of the correlations among reporting category scales are affected by differences in scale reliabilities that result from differences in numbers of items in the categories (see Table 69). For example, the observed correlations for Grade 3 Reading in the Research and Reference category would be expected to be lower than the other categories because it is measured with fewer items than the other categories. This means that all of the correlations among the reporting categories are underestimated due to lower reliabilities of corresponding subscores. Also, it should be noted that the number of students reported in the following tables are not the same as the number of students reported in the calibration samples of the demographic tables above (see Tables 2–49) due to the fact that only standard curriculum students are included in the intercorrelations among reporting categories while all students with reportable scores who were in the calibration schools are included in the demographic tables.

Table 69. Grade 3 Reading Reporting Category and Scale Score Intercorrelations

	Total Raw Score (45)	Word & Text (7)	Main Ideas (22)	Relation-ships (12)	Research & Ref. (4)
Scale Score	0.946	0.760	0.898	0.830	0.690
Total Raw Score	1	0.792	0.949	0.883	0.732
Word & Text	--	1	0.669	0.630	0.537
Main Ideas	--	--	1	0.747	0.632
Relationships	--	--	--	1	0.585

Note: Number of items in parentheses; N = 8,588.

Table 70. Grade 4 Reading Reporting Category and Scale Score Intercorrelations

	Total Raw Score (45)	Word & Text (6)	Main Ideas (19)	Relation-ships (15)	Research & Ref. (5)
Scale Score	0.970	0.644	0.910	0.870	0.727
Total Raw Score	1	0.676	0.938	0.894	0.741
Word & Text	--	1	0.553	0.522	0.422
Main Ideas	--	--	1	0.747	0.617
Relationships	--	--	--	1	0.577

Note: Number of items in parentheses; N = 7,594.

Table 71. Grade 5 Reading Reporting Category and Scale Score Intercorrelations

	Total Raw Score (45)	Word & Text (7)	Main Ideas (17)	Relation-ships (15)	Research & Ref. (6)
Scale Score	0.954	0.681	0.881	0.843	0.662
Total Raw Score	1	0.728	0.914	0.884	0.703
Word & Text	--	1	0.565	0.538	0.428
Main Ideas	--	--	1	0.713	0.546
Relationships	--	--	--	1	0.537

Note: Number of items in parentheses; N = 8,088.

Table 72. Grade 6 Reading Reporting Category and Scale Score Intercorrelations

	Total Raw Score (45)	Word & Text (11)	Main Ideas (15)	Relation-ships (11)	Research & Ref. (8)
Scale Score	0.965	0.833	0.872	0.850	0.800
Total Raw Score	1	0.865	0.900	0.875	0.838
Word & Text	--	1	0.692	0.677	0.649
Main Ideas	--	--	1	0.707	0.665
Relationships	--	--	--	1	0.671

Note: Number of items in parentheses; N = 7,688.

Table 73. Grade 7 Reading Reporting Category and Scale Score Intercorrelations

	Total Raw Score (45)	Word & Text (7)	Main Ideas (20)	Relation-ships (9)	Research & Ref. (9)
Scale Score	0.958	0.781	0.900	0.827	0.803
Total Raw Score	1	0.816	0.940	0.864	0.834
Word & Text	--	1	0.688	0.649	0.614
Main Ideas	--	--	1	0.743	0.693
Relationships	--	--	--	1	0.647

Note: Number of items in parentheses; N = 8,297.

Table 74. Grade 8 Reading Reporting Category and Scale Score Intercorrelations

	Total Raw Score (45)	Word & Text (6)	Main Ideas (18)	Relation-ships (8)	Research & Ref. (13)
Scale Score	0.973	0.740	0.871	0.773	0.883
Total Raw Score	1	0.756	0.899	0.783	0.911
Word & Text	--	1	0.605	0.544	0.608
Main Ideas	--	--	1	0.630	0.709
Relationships	--	--	--	1	0.623

Note: Number of items in parentheses; N = 7,756.

Table 75. Grade 9 Reading Reporting Category and Scale Score Intercorrelations

	Total Raw Score (45)	Word & Text (4)	Main Ideas (20)	Relation-ships (10)	Research & Ref. (11)
Scale Score	0.956	0.715	0.893	0.804	0.839
Total Raw Score	1	0.748	0.940	0.833	0.877
Word & Text	--	1	0.644	0.555	0.592
Main Ideas	--	--	1	0.690	0.743
Relationships	--	--	--	1	0.655

Note: Number of items in parentheses; N = 8,300.

Table 76. Grade 10 Reading Reporting Category and Scale Score Intercorrelations

	Total Raw Score (45)	Word & Text (6)	Main Ideas (15)	Relation-ships (12)	Research & Ref. (12)
Scale Score	0.979	0.711	0.858	0.834	0.871
Total Raw Score	1	0.721	0.876	0.861	0.884
Word & Text	--	1	0.552	0.518	0.546
Main Ideas	--	--	1	0.660	0.676
Relationships	--	--	--	1	0.678

Note: Number of items in parentheses; N = 7,193.

Table 77. Grade 3 Mathematics Reporting Category and Scale Score Intercorrelations

	Total Raw Score (40)	Number Sense (12)	Measure-ment (8)	Geometry (7)	Algebra (6)	Data Analysis (7)
Scale Score	0.963	0.876	0.814	0.721	0.787	0.805
Total Raw Score	1	0.903	0.851	0.759	0.809	0.835
Number	--	1	0.718	0.581	0.686	0.674
Measurement	--	--	1	0.550	0.636	0.619
Geometry	--	--	--	1	0.518	0.583
Algebra	--	--	--	--	1	0.610

Note: Number of items in parentheses; N = 8,615.

Table 78. Grade 4 Mathematics Reporting Category and Scale Score Intercorrelations

	Total Raw Score (39)	Number Sense (10)	Measure-ment (8)	Geometry (7)	Algebra (7)	Data Analysis (7)
Scale Score	0.962	0.834	0.780	0.725	0.788	0.792
Total Raw Score	1	0.865	0.813	0.764	0.817	0.815
Number	--	1	0.626	0.563	0.640	0.626
Measurement	--	--	1	0.528	0.583	0.574
Geometry	--	--	--	1	0.528	0.561
Algebra	--	--	--	--	1	0.588

Note: Number of items in parentheses; N = 7,441.

Table 79. Grade 5 Mathematics Reporting Category and Scale Score Intercorrelations

	Total Raw Score (50)	Number Sense (12)	Measure-ment (11)	Geometry (9)	Algebra (10)	Data Analysis (8)
Scale Score	0.965	0.879	0.850	0.806	0.850	0.853
Total Raw Score	1	0.911	0.884	0.835	0.877	0.885
Number	--	1	0.767	0.680	0.760	0.754
Measurement	--	--	1	0.669	0.730	0.730
Geometry	--	--	--	1	0.657	0.672
Algebra	--	--	--	--	1	0.730

Note: Number of items in parentheses; N = 8,066.

Table 80. Grade 6 Mathematics Reporting Category and Scale Score Intercorrelations

	Total Raw Score (44)	Number Sense (9)	Measure-ment (9)	Geometry (9)	Algebra (8)	Data Analysis (9)
Scale Score	0.949	0.800	0.843	0.801	0.743	0.817
Total Raw Score	1	0.826	0.891	0.840	0.810	0.850
Number	--	1	0.670	0.610	0.591	0.634
Measurement	--	--	1	0.687	0.660	0.705
Geometry	--	--	--	1	0.595	0.642
Algebra	--	--	--	--	1	0.607

Note: Number of items in parentheses; N = 7,624.

Table 81. Grade 7 Mathematics Reporting Category and Scale Score Intercorrelations

	Total Raw Score (44)	Number Sense (9)	Measure-ment (9)	Geometry (8)	Algebra (9)	Data Analysis (9)
Scale Score	0.925	0.784	0.836	0.718	0.796	0.798
Total Raw Score	1	0.822	0.897	0.785	0.875	0.872
Number	--	1	0.671	0.557	0.651	0.641
Measurement	--	--	1	0.639	0.737	0.737
Geometry	--	--	--	1	0.602	0.606
Algebra	--	--	--	--	1	0.702

Note: Number of items in parentheses; N = 8,197.

Table 82. Grade 8 Mathematics Reporting Category and Scale Score Intercorrelations

	Total Raw Score (50)	Number Sense (12)	Measure-ment (11)	Geometry (8)	Algebra (10)	Data Analysis (9)
Scale Score	0.947	0.826	0.840	0.799	0.780	0.852
Total Raw Score	1	0.878	0.902	0.842	0.886	0.868
Number	--	1	0.733	0.662	0.726	0.717
Measurement	--	--	1	0.729	0.742	0.722
Geometry	--	--	--	1	0.666	0.663
Algebra	--	--	--	--	1	0.722

Note: Number of items in parentheses; N = 7,672.

Table 83. Grade 9 Mathematics Reporting Category and Scale Score Intercorrelations

	Total Raw Score (44)	Number Sense (8)	Measure-ment (7)	Geometry (11)	Algebra (10)	Data Analysis (8)
Scale Score	0.941	0.770	0.817	0.817	0.829	0.786
Total Raw Score	1	0.820	0.857	0.884	0.878	0.827
Number	--	1	0.647	0.642	0.652	0.610
Measurement	--	--	1	0.712	0.695	0.643
Geometry	--	--	--	1	0.694	0.656
Algebra	--	--	--	--	1	0.666

Note: Number of items in parentheses; N = 8,087.

Table 84. Grade 10 Mathematics Reporting Category and Scale Score Intercorrelations

	Total Raw Score (50)	Number Sense (11)	Measure-ment (9)	Geometry (10)	Algebra (12)	Data Analysis (8)
Scale Score	0.911	0.788	0.775	0.816	0.843	0.792
Total Raw Score	1	0.835	0.867	0.916	0.921	0.858
Number	--	1	0.665	0.696	0.721	0.668
Measurement	--	--	1	0.757	0.756	0.676
Geometry	--	--	--	1	0.791	0.719
Algebra	--	--	--	--	1	0.749

Note: Number of items in parentheses; N = 7,063.

Student Classification Accuracy and Consistency

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

Accuracy of Classification

According to Livingston and Lewis, 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 scores, if their true scores could somehow be known." Accuracy estimates are calculated from cross-tabulations between ". . . classifications based on an observable variable (scores on . . . a test) and classifications based on an unobservable variable (the test takers' true scores)." True score is also referred to as a hypothetical mean of scores from all possible forms of the test, if they could be somehow obtained (Young and Yoon, 1998). Since these true scores are not available, Livingston and Lewis provide a method to estimate the true score distribution of a test and create the cross-tabulation of the true score and observed score classifications. The example of the 5 × 5 cross-tabulation of the true score versus 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. The detailed procedure for calculating accuracy of classification is presented in Appendix E.

Table 85. 2006 FCAT Grade 3 Reading True Scores vs. Observed Scores Cross-

True Score	Observed Score					
	LEVEL 1	**LEVEL 2**	**LEVEL 3**	**LEVEL 4**	**LEVEL 5**	**Total**
LEVEL 1	0.146	0.020	0.003	0.000	0.000	0.168
LEVEL 2	0.031	0.048	0.034	0.000	0.000	0.113
LEVEL 3	0.006	0.036	0.211	0.050	0.000	0.302
LEVEL 4	0.000	0.000	0.055	0.293	0.035	0.384
LEVEL 5	0.000	0.000	0.000	0.012	0.021	0.033
Total	0.182	0.103	0.303	0.356	0.056	1.000

Note: Column and row totals are computed from nonrounded values. Shaded cells are used for computing overall accuracy index.

Consistency of Classification

Consistency is "the agreement between classifications based on two non-overlapping, equally difficult forms of the test." 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 5 × 5 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. The detailed procedure for calculating consistency of classification is presented in Appendix E.

Note that the consistency table is symmetrical; however, the accuracy table is nonsymmetrical because it compares classifications based on two different types of scores. Also note that agreement rates are lower in the consistency table because both classifications contain measurement error; whereas, in the accuracy table, true score classification is assumed to be errorless.

Table 86. 2006 FCAT Grade 3 Reading True Scores vs. Observed Scores Cross-Tabulation (Consistency Table)

Form Taken	Alternate Form					
	LEVEL 1	**LEVEL 2**	**LEVEL 3**	**LEVEL 4**	**LEVEL 5**	**Total**
LEVEL 1	0.140	0.028	0.014	0.000	0.000	0.182
LEVEL 2	0.028	0.033	0.040	0.002	0.000	0.103
LEVEL 3	0.014	0.040	0.178	0.071	0.001	0.303
LEVEL 4	0.000	0.002	0.071	0.249	0.033	0.356
LEVEL 5	0.000	0.000	0.001	0.033	0.023	0.056
Total	0.182	0.103	0.303	0.356	0.056	1.000

Note: Column and row totals are computed from nonrounded values. Shaded cells are used for computing consistency index conditional on level.

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 *cutpoint*. In order to facilitate their interpretations by explaining how to understand them correctly, a brief outline of computational procedures used to derive accuracy indices will be presented using the example of the FCAT Grade 3 Reading test.

The *overall accuracy* of performance level classifications is computed as a sum of the proportions on the diagonal of the joint distribution of true score and observed score levels, as indicated by the shaded area in Table 85. Actually, it is a proportion (or percentage) of correct classifications across all the levels. In the particular example, the overall accuracy index for the FCAT Grade 3 Reading test equals 0.719 (71.9 percent). This means that 71.9 percent of students are classified in the same performance categories based on their observed scores, as they would be classified based on their true scores, if they could be known.

The *overall consistency* index is analogously computed as a sum of the diagonal cells in the consistency table. Using the data from Table 86, it can be determined that the *overall consistency* index for the FCAT Grade 3 Reading test equals 0.623 (62.3 percent). In other words, 62.3 percent of Grade 3 students would be classified in the same performance levels based on the alternate form, if they would have taken it. Another way to express overall consistency is to use Cohen's *kappa (κ)* coefficient (Cohen, 1960). The overall coefficient kappa when applying all cutoff scores together is

$$k = \frac{P - P_c}{1 - P_c} \ ,$$

where P is the probability of consistent classification, and P_c is the probability of consistent classification by chance (Lee, 2000).

Kappa is a measure of "how much agreement exists beyond chance alone..." (Fleiss, 1973), which means that it assesses the proportion of consistent classifications between two forms after removing the proportion of consistent classifications expected by chance alone. Using the data from Table 88, it was computed that Cohen's κ for FCAT Grade 3 Reading equals 0.479. 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 of 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 0.249 (proportion of correct classifications) and 0.356 (total proportion of students classified into the LEVEL 4) yields 0.699, which represents the index of consistency of classification for FCAT Grade 3 Reading that is conditional on LEVEL 4. It indicates that 70 percent of all of the students whose performance is classified as LEVEL 4 would be classified in the same level based on the alternate form, if an alternate form were taken.

Accuracy conditional on level is analogously computed. The only difference is that in the consistency table, both row and column marginal sums are the same; whereas, in the accuracy table, the sum that is based on true status is used as a total for computing accuracy conditional on level. For example, in Table 85 the proportion of agreement between true score status and observed score status at LEVEL 1 is 0.146, whereas the total proportion of students with true score status at this level is 0.182. The accuracy conditional on level is equal to the ratio between those two proportions, which yields 0.869. This indicates that 87 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.

Perhaps the most important indices for accountability systems are those for the accuracy and consistency of classification decisions made at specific cutpoints. To evaluate decisions at specific cutpoints, the joint distribution of all performance levels is collapsed into a dichotomized distribution around that specific cutpoint. For example, the dichotomization at the cutpoint that separates LEVEL 1 through LEVEL 3 (combined) from LEVEL 4 and LEVEL 5 (combined) for FCAT Grade 3 Reading is depicted in Table 87. The proportion of correct classifications below that particular cutpoint is equal to the sum of the cells in the upper-left shaded area (0.535), and the proportion of correct classifications above the particular cutpoint is equal to the sum of the cells in the lower right shaded area (0.361).

Table 87. 2006 FCAT Grade 3 Reading 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	0.146	0.020	0.003	0.000	0.000	0.168
LEVEL 2	0.031	0.048	0.034	0.000	0.000	0.113
LEVEL 3	0.006	0.036	0.211	0.050	0.000	0.302
LEVEL 4	0.000	0.000	0.055	0.293	0.035	0.384
LEVEL 5	0.000	0.000	0.000	0.012	0.021	0.033
Total	0.182	0.103	0.303	0.356	0.056	1.000

Note: Columns and row totals are computed from nonrounded values. Shaded cells are used for computing accuracy at specific cutpoints.

The *accuracy index at cutpoint* is computed as the sum of the proportions of correct classifications around a selected cutpoint. In our example from Table 87, the sum of both shaded areas (upper-left shaded areas added to lower-right shaded areas) equals 0.896, which means that 89.6 percent of the students were correctly classified either above or below the particular cutpoint. The sum of the proportions in the upper-right nonshaded area (0.050) indicates false positives (i.e., 5 percent of the students are classified above the cutpoint by their observed score but are falling below the cutpoint by their true score); the sum of the lower-left nonshaded area (0.055) is the proportion of false negatives (i.e., 5.5 percent of students are observed below the cutpoint level, but their true level is above the cutpoint).

The *consistency index at cupoint* is obtained in an analogous way. For example, by taking data from Table 86 and dichotomizing the distribution at the cutpoint between LEVEL 1 and all other levels combined, it can be determined that the proportion of correct classifications around that cutpoint equals 0.911. This means that 91.1 percent of the students would be classified by an alternate form (if they had taken it) in the same two categories (LEVEL 1 or LEVEL 2 through LEVEL 5 combined) as they were classified by the actual form taken.

Accuracy and Consistency Results for 2006 FCAT

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

The overall indices of accuracy and consistency of classification for the FCAT 2006 tests are presented in Table 88.

Table 88. Estimates of Accuracy and Consistency of Performance-Level Classification by Grade and Subject

Grade	Subject	Accuracy	Consistency	Kappa (κ)
3	**Reading**	0.718	0.623	0.486
	Mathematics	0.701	0.596	0.481
4	**Reading**	0.666	0.553	0.396
	Mathematics	0.674	0.566	0.437
5	**Reading**	0.659	0.559	0.419
	Mathematics	0.666	0.553	0.410
6	**Reading**	0.964	0.589	0.463
	Mathematics	0.608	0.513	0.362
7	**Reading**	0.700	0.598	0.475
	Mathematics	0.621	0.515	0.366
8	**Reading**	0.642	0.551	0.392
	Mathematics	0.618	0.516	0.369
9	**Reading**	0.682	0.582	0.444
	Mathematics	0.618	0.508	0.363
10	**Reading**	0.623	0.542	0.380
	Mathematics	0.669	0.550	0.371

It can be seen from the above table that overall accuracy indices are in the range between 0.618 and 0.718, overall consistency indices range between 0.508 and 0.623, and κ coefficients fall in the range between 0.362 and 0.486.

In addition to overall ratings of decision accuracy, the levels of agreement at each performance level are also of interest. Table 89 displays the probability of students being classified as being in a particular performance level, given that their "true status" is the same category. It can be seen that in most tests, the accuracy indices at the lowest performance level (LEVEL 1) are substantially higher than at other levels. Similarly, the accuracy at the highest performance level is also elevated, but not so evidently as at the lowest level. This effect is due to the fact that extreme performance levels usually cover a wider range of the measured construct than the intermediate levels, and misclassification can occur in only one direction. It should be noted that the percentage of students whose observed scores are classified in the highest performance level is relatively low (it is below 10 percent for most of the tests; see Appendix D), which makes indices conditional at that level less reliable. In one instance (Grade 6 Mathematics) the percentage of students whose estimated true scores fall in LEVEL 5 equals zero, which makes it impossible to estimate the accuracy at that level; however, it is possible to estimate accuracy of decisions at the cutpoint between LEVEL 4 and LEVEL 5. Moreover, this estimate can be high (see Table 90).

Table 89. Accuracy of Classification at each Proficiency Level for each Grade and Subject

Grade	Subject	Level 1	Level 2	Level 3	Level 4	Level 5
3	Reading	0.864	0.421	0.697	0.764	0.634
	Mathematics	0.856	0.609	0.670	0.690	0.742
4	Reading	0.858	0.454	0.597	0.673	*
	Mathematics	0.844	0.569	0.644	0.673	0.632
5	Reading	0.877	0.408	0.623	0.668	0.591
	Mathematics	0.871	0.627	0.516	0.679	*
6	Reading	0.852	0.564	0.662	0.698	0.615
	Mathematics	0.873	0.442	0.583	0.475	*
7	Reading	0.864	0.544	0.683	0.697	0.672
	Mathematics	0.879	0.473	0.523	0.574	*
8	Reading	0.872	0.607	0.548	0.540	*
	Mathematics	0.892	0.542	0.613	0.481	*
9	Reading	0.895	0.604	0.584	0.571	*
	Mathematics	0.862	0.520	0.514	0.602	*
10	Reading	0.869	0.545	0.435	0.288	0.611
	Mathematics	0.902	0.574	0.497	0.678	*

*No accuracy estimates were calculated at LEVEL 5 for Grades 5, 6, 7, 8, 9, and 10 Mathematics and Grades 4, 8, and 9 Reading 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 cutpoint, which makes it desirable to know the accuracy and consistency of dichotomous decisions made around that specific cutpoint. For instance, if a college gave credits to advanced and proficient students who achieved LEVEL 5 and LEVEL 4 but not to those in LEVEL 1 through LEVEL 3, the focus of interest would be in accuracy and consistency of dichotomous decisions below, versus at and above the LEVEL 4 threshold. Reporting in a "percent at-or-above cut" (PAC) metric requires a judgment about whether a student's score is below or at-or-above a particular cutpoint. Table 90 presents the accuracy and consistency information for these dichotomous categorizations.

Table 90. Accuracy and Consistency of Dichotomous Categorizations by Grade and Subject (PAC Metric)

Grade	Subject	Accuracy				Consistency			
		1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5	1 / 2+3+4+5	1+2 / 3+4+5	1+2+3 / 4+5	1+2+3+4 / 5
3	Reading	0.941	0.921	0.894	0.953	0.916	0.888	0.852	0.933
	Mathematics	0.948	0.919	0.901	0.931	0.926	0.886	0.861	0.903
4	Reading	0.937	0.908	0.843	0.963	0.911	0.870	0.781	0.934
	Mathematics	0.940	0.909	0.888	0.931	0.915	0.872	0.843	0.905
5	Reading	0.927	0.906	0.880	0.933	0.898	0.867	0.832	0.907
	Mathematics	0.943	0.899	0.853	0.961	0.919	0.857	0.794	0.932
6	Reading	0.937	0.910	0.894	0.948	0.911	0.873	0.852	0.927
	Mathematics	0.925	0.895	0.835	0.930	0.893	0.850	0.780	0.889
7	Reading	0.933	0.907	0.901	0.954	0.905	0.870	0.862	0.935
	Mathematics	0.931	0.898	0.832	0.939	0.902	0.854	0.771	0.900
8	Reading	0.927	0.686	0.842	0.997	0.896	0.812	0.802	0.994
	Mathematics	0.945	0.912	0.828	0.919	0.922	0.874	0.767	0.873
9	Reading	0.924	0.894	0.886	0.970	0.892	0.851	0.843	0.950
	Mathematics	0.931	0.893	0.849	0.930	0.901	0.848	0.790	0.893
10	Reading	0.894	0.873	0.886	0.917	0.851	0.823	0.843	0.886
	Mathematics	0.962	0.924	0.807	0.961	0.945	0.888	0.728	0.929
10 P / F	Reading	0.894				0.851			
	Mathematics	0.955				0.936			

The data in Table 90 reveal that the level of agreement in terms of both accuracy and consistency for these dichotomous categorizations is very high. Although the rates of agreement for decision consistency are slightly lower, in no cases does the rate of agreement fall below 80 percent. In general, this means high rates of accuracy and consistency are available to support decisions about PACs.

The conclusion about high accuracy of PAC decisions is also supported by data on the percentages of false positives and false negatives derived from the dichotomized "true status" versus "observed status" categorizations (see Table 91). On average, only 4.69 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 0.000 to 0.139, indicating that not more than 13.9 percent of students were classified differently.

Table 91. Accuracy of Dichotomous Categorizations: False Positive and False Negative Rates (PAC Metric)

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	0.023	0.038	0.050	0.035	0.037	0.042	0.055	0.012
	Mathematics	0.021	0.033	0.047	0.041	0.031	0.048	0.052	0.028
4	Reading	0.029	0.035	0.060	0.037	0.034	0.057	0.097	*
	Mathematics	0.027	0.041	0.059	0.046	0.033	0.049	0.053	0.023
5	Reading	0.025	0.045	0.059	0.056	0.048	0.049	0.061	0.011
	Mathematics	0.024	0.040	0.063	0.039	0.033	0.061	0.085	*
6	Reading	0.030	0.040	0.050	0.037	0.034	0.051	0.056	0.015
	Mathematics	0.035	0.050	0.060	0.070	0.040	0.055	0.104	*
7	Reading	0.029	0.048	0.053	0.032	0.038	0.044	0.046	0.014
	Mathematics	0.030	0.044	0.082	0.061	0.039	0.059	0.086	0.000
8	Reading	0.032	0.047	0.120	0.003	0.041	0.085	0.038	*
	Mathematics	0.023	0.036	0.070	0.081	0.032	0.052	0.102	*
9	Reading	0.032	0.052	0.061	0.030	0.044	0.054	0.053	*
	Mathematics	0.029	0.048	0.079	0.070	0.041	0.060	0.072	*
10	Reading	0.050	0.067	0.069	0.063	0.056	0.060	0.045	0.020
	Mathematics	0.016	0.025	0.054	0.039	0.022	0.051	0.139	*
10 P / F	Reading	0.050				0.056			
	Mathematics	0.018				0.026			

* False negatives could not be estimated at 1+2+3+4 vs. 5 cutpoint for Grades 4, 8, and 9 Reading and Grades 5, 6, 7, 8, 9, and 10 Mathematics because the number of estimated true scores in the LEVEL 5 cell is 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 test. Students hoping to receive a regular diploma are required, among other things, to achieve a score of 300 or better on the FCAT Reading and Mathematics tests. In principle, it is possible for the following three situations to be found:

1. A student's observed performance is accurately reflected in terms of the standard and in terms of his or her true level of ability. (A student whose ability is at or above the minimum acceptable standard achieves a test score at or above that standard. A student whose true ability is below the standard achieves a score below the standard.)

2. A student whose true ability is below the standard receives a score that is, in fact, above the standard ("false positives").

3. A student whose true ability is, in fact, above the standard, but whose observed scores indicate (inaccurately) that he or she has not met the standard ("false negatives" that will, inappropriately, require the student to take the test again).

False-positive and false-negative rates for all dichotomous classifications for FCAT tests are presented in Table 91.

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

- Grade 10 Reading has the fail-pass threshold that is the same as the threshold between performance LEVELs 1 and 2. The accuracy of fail-pass decisions for this test is equal to the accuracy of dichotomous categorization between LEVEL 1 and LEVELs 2, 3, 4, and 5 combined. It can be seen from Table 90 that 89 percent of the students are correctly classified into either the pass or fail category (situation 1) based on their observed performance on Grade 10 Reading.

- Because the threshold score for fail-pass decisions for Grade 10 Mathematics falls in the middle of performance LEVEL 2, a separate analysis to estimate the accuracy of fail-pass decisions for this test was performed. The analysis shows that 96 percent of students were classified correctly into either a pass or fail category (situation 1) based on their observed performance on Grade 10 Mathematics.

REFERENCES

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

Baker, Frank (2001). *The Basics of Item Response Theory.* ERIC Clearinghouse on Assessment and Evaluation, College Park, Md.: University of Maryland.

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

Cohen, J. and Cohen, P. (1975). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum.

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

Florida Department of Education (1996). *Sunshine State Standards.* Retrieved September 20, 2002, from the Florida Department of Education website: http://www.fldoe.org/bii/curriculum/sss/

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

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

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

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

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

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

Florida Department of Education (November, 2005). *Plan for Selecting the Calibration Sample for the 2006 FCAT Administration.* Unpublished. Tallahassee, Fla.: Author.

Florida Department of Education (2005). *The FCAT 2006 Test Construction Specifications.* Unpublished. Tallahassee, Fla.: 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, Iowa: 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), pp. 179–197.

Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *Journal of American Statistical Association*. 58, pp. 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, pp. 719–748.

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

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

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, Calif.

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

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

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, pp. 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, pp. 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, Calif.: 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), pp. 233–251.