6.0 Scoring the Test

The process of scoring the FCAT begins after student answer documents are returned to the DOE's contractor. Just as test construction can be viewed in terms of item development and whole-test construction, so can the scoring process be viewed in terms of item scoring and whole-test scoring. This distinction is necessary because the discussion of item scoring focuses on the methods used to rate student responses to individual items, whereas the discussion of whole-test scoring focuses on the statistical methods used to derive *scale scores* for the test overall. Several of the concepts and terms used in this chapter, such as *true score* and *developmental scale score*, are also used in Chapter 7.0, Reporting FCAT Results.

This chapter is divided into two sections, one dealing with the process and methods for scoring items and the other describing the methods used to generate scores for the test as a whole, including scale scores, developmental scale scores, and Achievement Level classifications. In addition, each section details the quality control processes used to ensure the accuracy of scores.

6.1 Scoring Multiple-Choice and Gridded-Response Items

Multiple-choice (MC) and gridded-response (GR) items are scanned and scored using automated processes. As such, these items are frequently referred to as "machine scored." Slightly different processes are used to score multiple-choice and gridded-response items.

Multiple-choice items have only one correct answer. Although rare, when a mis-keyed multiple-choice item is found, the key is corrected or the item is deleted from scoring. Because several correct answers or answer formats are possible for gridded-response items, a list of acceptable answers must be identified for use by the scoring program. The Gridded-Response Adjudication Committee works with the DOE to identify all acceptable answers and formats when other possibilities are discovered during scoring. See Section 4.1 and Appendix D for more information about this committee.

Numerous checks are incorporated in the scoring program to alert scoring personnel to any possible problems with an item, such as when a large number of otherwise high-achieving students chose or gridded an answer that was not originally identified as correct. These situations lead scoring personnel to investigate whether there is more than one correct answer to a multiple-choice item or whether the list of acceptable answers to gridded-response items may need to be expanded.

> ***Quality Assurance Measures: Statistical Reviews***—The same statistical reviews conducted on items after field testing and on test forms during test construction are conducted after operational testing. These reviews are conducted again because the

population of students taking the operational test may not have the same characteristics as the field-test population. Another purpose of these reviews is to ensure that the items and test have the characteristics that will make the FCAT an effective measure of student achievement. Any deviation from the specified criteria might compromise the accuracy of the student scores.

6.2 Scoring Short- and Extended-Response Performance Task Items and Prompted Essays (Handscoring)

Handscoring is guided by a set of *Handscoring Specifications.* Because the *Handscoring Specifications* contain detailed information about the FCAT test content, they are protected by test security statutes and are not available to the public. FCAT scoring of performance tasks is *holistic,* as opposed to *analytic,*[11] meaning that a single rating is given for the response as a whole. For FCAT Reading, FCAT Mathematics, and FCAT Science, 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. For FCAT Writing+ essays, scorers use a scale that ranges from Unscorable (0) to 6. For more information regarding handscoring, see *Florida Reads!*, *Florida Writes!*, *Florida Solves!,* and *Florida Inquires!*, which are distributed to districts each spring, after the FCAT administration. Another resource is *FCAT Performance Task Scoring—Practice for Educators* publications and software.

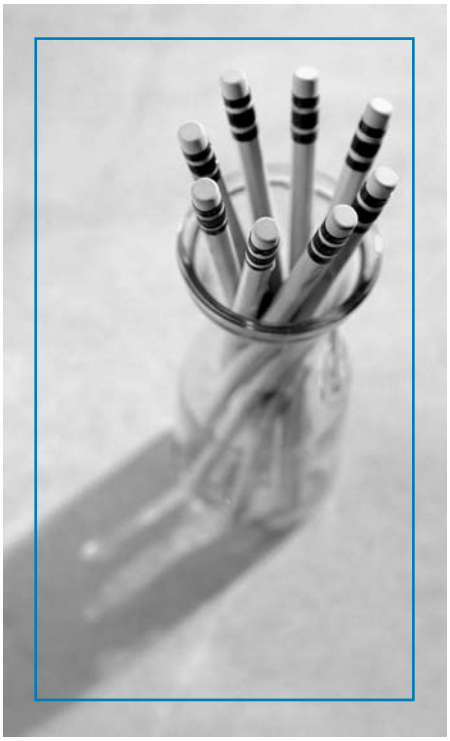
The anchor papers and item-specific criteria are developed initially by Florida educators serving on Rangefinder Committees (see page 46 and Appendix D for more information) and then reviewed and refined by other Florida educators on **Rangefinder Review Committees.** After performance task items are selected for use as operational items, Rangefinder Review Committees review the scoring guides and training materials originally established by the Rangefinder Committees. There are Rangefinder Review Committees for reading, mathematics, and science. 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.

Frank Santa Maria
(Reading and Writing Instruction Language Arts Department) Eighth-grade teacher and Department Chair, Murdock Middle School, Charlotte County Public Schools, Port Charlotte, Florida

FCAT Committee Experience: Writing Rangefinder; Writing Prompt Review; Writing Content Advisory; Prompt Writing Committee; Reading Standard Setting

"Fear not the FCAT! These exams were not designed to make us miserable. They were carefully conceived and are meticulously reviewed. They emerge each year from the coordinated efforts of the FDOE, its contractors, and professional educators. Having served on FCAT committees since 1997 has allowed me to appreciate the entire process and inspire my students to always do their best."

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



Short- and extended-response performance task items are handscored by professional scorers with the guidance of the DOE staff. These professional scorers include test contractor employees, educators who are not currently employed in the Florida public school system, retired teachers, part-time graduate students, and others. 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. Depending on the subject, applicants may be required to also take a subject-area exam or write an essay. Those selected as candidates attend a multiple-day training session at which they are provided with various materials to familiarize themselves with the scoring process and are provided multiple opportunities to practice scoring. At the end of the training, candidates must pass a qualifying examination. The examination requires them to score sets of sample essays or student responses for which scores have been established by Florida educators. To pass the examination, candidates must match the pre-established scores.

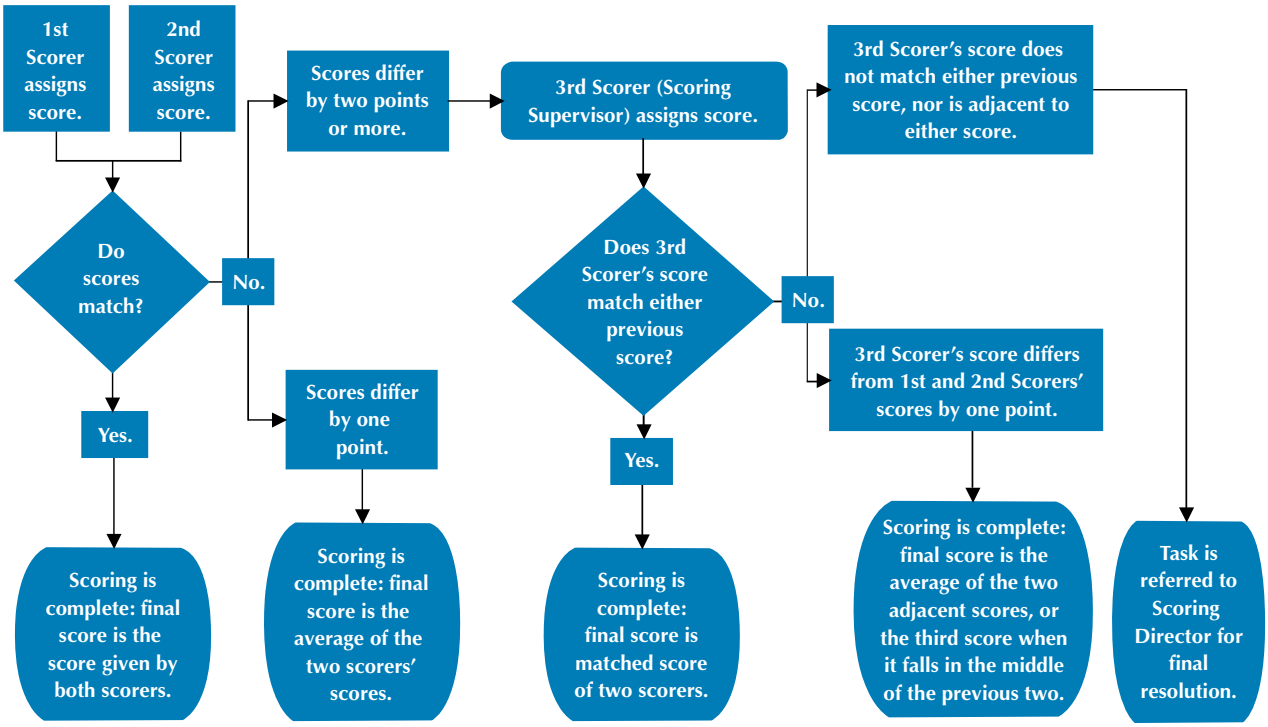
Those selected as professional scorers work in teams of 10–15 members with each team having a Scoring Supervisor. Each team specializes in a set of two to three performance task items, known as *rater item blocks* (RIBs) (for reading, mathematics, or science), or in a single writing prompt. A Scoring Director and an Assistant Scoring Director supervise all the teams assigned to a prompt or RIB. Prior to the scoring sessions, all student responses to writing prompts and performance task items 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.

To guide them in rating responses, scorers have the following tools and references at their disposal:

- A general scoring rubric for all items of the same subject, grade level, and item type, with descriptions of work demonstrative of each point on the scale.
- Anchor papers with annotations—Actual, unedited student responses to the task or essay that illustrate typical performance for each point on the scale. Each student response is annotated with a rationale for the score given. Anchor papers are also called rangefinder papers.
- Item-specific criteria—For FCAT Reading, FCAT Mathematics, and FCAT Science, scorers have a description and example of a top-score response for each item.

As shown in Figure 19, each student response is read independently by at least two professional scorers. For short-response performance tasks, if the scorers' two scores are not identical, a third scorer reviews the response to resolve the difference. For extended-response performance tasks, a third scorer is used if the first two scores are nonadjacent, that is, if they differ by more than one point. This third scoring, called resolution scoring, is performed by a Scoring Supervisor. All scoring is carefully monitored by the DOE staff.

Figure 19: Handscoring Process for FCAT Writing+ Essays



Quality Assurance Measures for Handscoring—Numerous measures are in place to ensure scoring accuracy and consistency. Some of these have already been mentioned, such as the process for selecting and training scorers of reading, mathematics, and science performance tasks and writing essays. Additional methods of ensuring accuracy and consistency of handscoring include:

- ***Use of Same Scoring Materials Each Year***—Each time a performance task appears on the FCAT, scorers are trained using the same set of training materials and scoring guidelines that were used in previous years. The FCAT Rangefinder Review Committees may make minor revisions to these documents for clarity, but the criteria and examples for each score point remain the same every year.

- ***Backreading***—Scoring Supervisors (and Scoring Directors, as needed) check the work of individual scorers to ensure that they are scoring responses in accordance with the established guidelines. Supervisors read behind all scorers throughout the scoring session. This is called backreading, and it is done with more frequency at the beginning of the scoring session to identify scorers who may need additional training and monitoring. Supervisors ask scorers to review responses that were scored incorrectly, and then provide guidance on how to score more accurately.

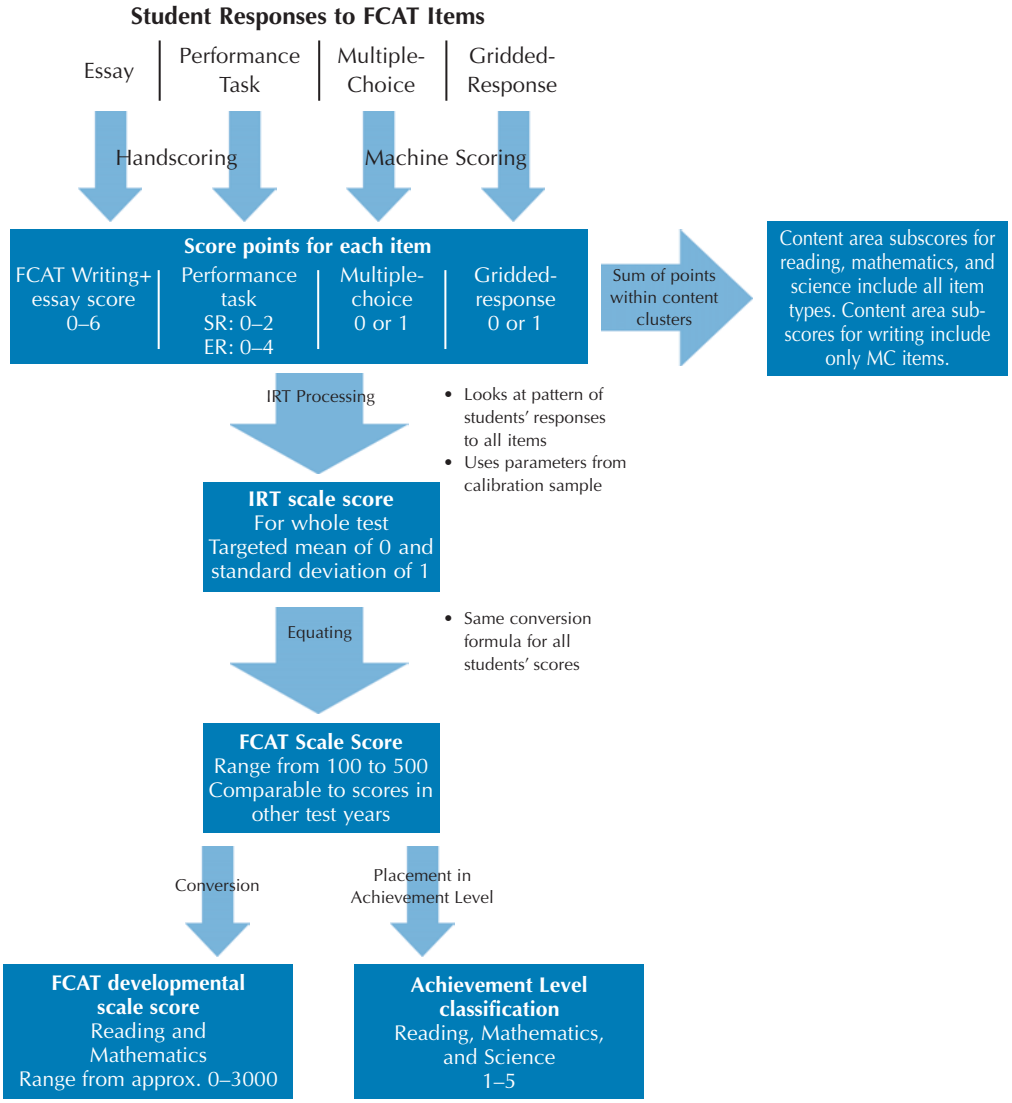
- ***Daily Review of Training Materials***—At the beginning of each scoring session, team members spend at least 15 minutes reviewing their training materials and scoring guidelines, including anchor papers and item-specific criteria.

- ***Calibration Sessions (Retraining)***—Scorers meet periodically as a team to review scoring guidelines. They review anchor papers, which represent the range of responses for each possible score point and have been pre-scored by the FCAT Rangefinder and Rangefinder Review Committees. The anchor papers provide scorers with a clear definition of each score point. This process and the quality control measures (reliability and validity checks) implemented during scoring ensure that all performance tasks are scored according to Florida's standards. Retraining is also conducted for scorers whose scores are consistently inaccurate or fall below acceptable standards. If retraining is unsuccessful, scorers are dismissed from the program.

- ***Validity and Reliability Reports***—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. Comparisons of the scores assigned by a scorer with the established scores are compiled as validity reports and presented to Scoring Directors and DOE staff throughout the scoring sessions. From the validity reports, Scoring Directors can see which responses are most often scored incorrectly and which scorers are most often in disagreement with the established scores. Reliability (consistency) of handscoring is monitored using reports of inter-rater reliability. Each scorer's (or rater's) score on a student response is compared to the other score given to that response. A cumulative percent of agreement between the two scores on every response (as opposed to validity responses only) is reported for each scorer as the inter-rater reliability percent. The information on this report indicates whether a scorer is agreeing with other scorers scoring the same responses. Analysis of the report is used to determine if a scorer or group of scorers is drifting from the established guidelines and require additional training.

6.3 Whole-Test Scoring

Figure 20: Derivation of FCAT Scores



Beginning in Spring 2006, FCAT Writing+ student performance will be reported using a scale score of 100 to 500. This scale score will encompass performance on the essay as well as the multiple-choice items. A developmental scale score is not available for either science or writing. Figure 20 above displays the derivation of FCAT scores across content areas and item types.

Content subscores are provided for each subject area test. These subscores are provided as the number of points correct compared to the number of points possible. Chapter 3, Test Content and Format, provides the content categories for each subject with the range of points possible in each category.

Quality Assurance Measures—For most statistical indicators, post-operational test reviews are conducted on data from a carefully selected group of students representative of all students tested. A notable exception is Standard Error of Measurement (SEM),

a reliability indicator that is calculated using data from the entire tested population. Although the SEM is derived differently for tests scored using IRT, the meaning is similar. That is, if a student were to take the same test over and over (without additional learning between the tests or without remembering any of the questions from the previous tests), the indicator of the variance in the resulting test scores is called the standard error of measurement. If the reviews find that the test displays less-than-ideal characteristics, adjustments can be made during scoring, e.g., an item can be excluded from scoring; however, because of the stringent selection criteria for operational items, such cases are rare.

Mark D. Reckase, Ph.D.
(Design and development of large scale assessments)
Professor, Michigan State University, Okemos, Michigan

FCAT Committee Experience: Technical Advisory Committee

Related Experience: America Educational Resource Association (AERA), Vice President of Division D; National Assessment Governing Board—Executive Committee

"As a university professor, it is important for me to keep up to date on the technical and policy issues related to large scale assessment so I can pass that information along to my students. The FCAT is one of the best state testing programs in the country, and it serves as a good example of ways such programs should be implemented."

Scale Scores

FCAT scale scores are the result of a two-step process that analyzes student responses using Item Response Theory (IRT) and uses the resulting item parameters to convert student responses to a scale score that is comparable across test years.

IRT Scoring

As described in Section 4.5 (IRT Framework, page 60), the IRT model used to develop and score the FCAT is based on the idea that each student possesses a certain level of knowledge and skill, what IRT calls *ability*. The goal of the FCAT and of the quality control process described in this *Handbook* is to accurately report a score as close to the true level of ability as possible. The IRT model is widely used because it produces the most accurate score estimates possible.

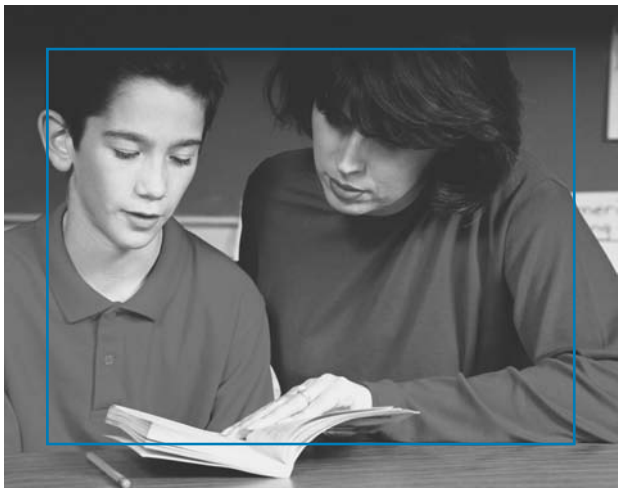
Another key feature of the IRT model is that ability and item difficulty exist on a single dimension so that students with low scores[12] will generally succeed on less difficult items, students with moderate scores will typically succeed on items with low to moderate difficulty, and students with high scores

[12] In this case "low scores" (and "moderate scores" and "high scores") refers to a student's true level of ability, which the test attempts to estimate. It does not refer to any other assessment of student achievement, such as scores on other tests, report card grades, or teacher assessments. If a student with a history of poor academic performance performs well on the FCAT, for the purposes of this discussion, he or she is a student with high ability.

will succeed on items at all levels of difficulty. Ideally, any test constructed using the IRT model will include items that clearly distinguish between students with increasing levels of ability.

Two important aspects of IRT processing contrast with traditional methods of test scoring. One aspect is that items are given different considerations based on their differing IRT parameters when calculating the overall score. For example, relatively more consideration might be given to items with a greater discrimination (a high *a*-parameter) and relatively less consideration might be given to items on which a lot of guessing occurs (a high *c*-parameter). In situations like these, different considerations apply in the same way to the calculation of scores for all students.

Another important contrast between IRT scoring and traditional methods is the use of *pattern scoring*. That is, the pattern of correct and incorrect answers provided by a student is analyzed in combination with the IRT item parameters. Students who know the correct answer may inexplicably miss easy items, and sometimes students who do not know the answer get difficult items correct. Information about the pattern of answers and the test items is used to evaluate the likelihood of individual student responses. This is called pattern scoring. As a result of this method of scoring, students with the same raw score may have similar, but not necessarily identical, scale scores. Different scale scores result because the students' patterns of correct answers were different.



The Miami Herald February 11, 2003 Tuesday BR EDITION

FCAT Gets High Marks in Measuring Achievement

For the complete text of this article, see Appendix C.

IRT pattern scoring is used with the FCAT because it produces more accurate depictions of students' true levels of ability (knowledge and skill).

IRT pattern scoring may result in situations in which students answering the same number of items correctly would receive different scale scores because the pattern of their answers (which questions were answered correctly or incorrectly) is different. Students who correctly answer exactly the same items would, of course, receive the same scale score. Using IRT pattern scoring is an important method of ensuring the most accurate measure of student achievement possible.

Process

In the first step of scoring, each item's IRT parameters are calculated using a carefully selected sample of schools that represents the total state population. This is called the *calibration sample* and the schools selected as part of this sample are often referred to as "early-return" schools. The role that the calibration schools play is critical to the scoring process because the item parameters that are calculated based upon this sample are used to generate scores for all students.

Equating

After IRT calibration, the process of *equating* is used to place IRT-processed scores on the FCAT scale of 100 to 500 and to ensure that the resulting scores are comparable to those of previous years. Making scores comparable allows comparisons between, for example, the achievement of Grade 8 students in 2004 and the achievement of Grade 8 students in 2001. The FCAT is designed to be of similar difficulty each year; however, slight differences in test difficulty (the content of the test items) may influence student scores. Without equating, it would be difficult to determine whether differences in scores between years are the result of these slight differences in the test difficulty or differences in students' true levels of knowledge and skill.

Test developers can isolate the influence of differences in student ability through the use of *anchor items*—items that appear identically in tests of consecutive years. Because these items are identical, differences in achievement between groups can be more clearly identified. Using the Stocking/Lord[13] procedure, the procedure used to maintain the FCAT scale year after year, a

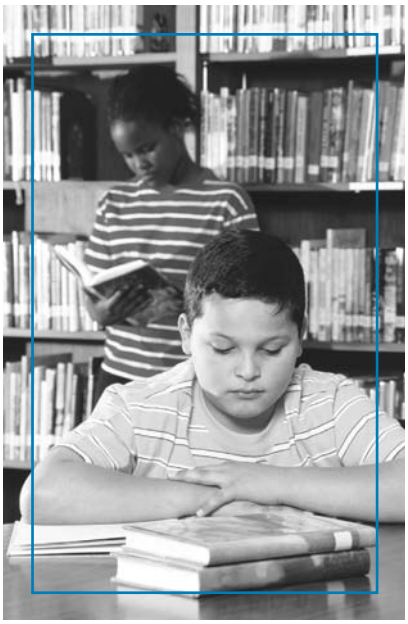
[13] Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Measurement*, 7, 201–210.

statistical relationship is established between the performance of current year students on these anchor items and the performance of students in the first year of operational testing. This relationship enables the item parameters from the current test form to be expressed on the same scale as the first operational (base) test form. Numerous steps are taken to ensure that the anchor items sufficiently represent the tests so that this relationship can be applied to the entire test for current year students. After this equating process, it is possible to report scores on a scale of 100 to 500 that are comparable to scores of previous years. This means that any differences in scores, such as the difference between mean scores for any two years, can be attributed to differences in student achievement and not to differences in the test difficulty. Anchor items are not included as part of a student's score; they are used only for the purpose of equating.

It is important to emphasize that the cross-year comparability of scores does not extend to the *content cluster subscores*. The content area cluster subscores are simply the total of all score points awarded in a given content cluster. Although anchor items are designed to be representative of the test overall, they are not sufficient for making comparisons across years within content clusters. Such a comparison would require a greater number of anchor items.

Developmental Scale Scores

In reading and mathematics, scale scores, ranging from 100 to 500 for each grade level, are converted to *developmental scale scores* (*DSS* or *vertical scale scores*), which place the scores of students on a scale ranging from 0 to 3000 for all grade levels tested. This continuous scale allows student progress to be tracked from one tested grade to the next. Placing scores on a vertical scale allows grade-to-grade growth to be represented more clearly and easily than piecing together data from several different scales. Without the FCAT developmental scale, individual students would know their scores for each year in which they took the test; however, because the score on each test would be on a 100–500 point scale, it would be difficult to chart progress over time.



The method for creating the developmental scale is similar to the method of equating described in the previous section. In equating, anchor items are placed on tests given in different years to relate the scores of the current year to the scores of the first year of operational testing. In a similar manner, the developmental scale is based on *linking items*—items that appear identically on the tests of adjacent grade levels—to relate the scores from one grade to those in the grades one year above and one year below it. With the scale score from each grade

level successively linked to those above and below it, a single scale is created. Linking is conducted to create the developmental scale score and is conducted periodically to verify or refine the scale. Linking items do not contribute to a student's score if items are not on grade level.

The intended use of the developmental scale score, also called the FCAT Score, is to monitor the progress of individual students over time. By comparing a student's scores in the same FCAT subject for two or more years with the associated mean scores (or with the various *Achievement Levels*, described in the following section) for those years, it is possible to identify whether a student's performance improved, declined, or remained consistent.

The developmental scale, however, is not intended to compare the achievement of different students in different grade levels or to make claims about a student's grade-level performance, such as a Grade 4 student attaining a score at the Grade 7 level. This is because the items used to link the tests are not representative of the broad spectrum of content at nonadjacent grade levels. As a result, a Grade 6 student's developmental scale score of 1600 on FCAT Mathematics cannot be compared to a Grade 8 student's score of 1600 because, besides linking items, the content of the FCAT Mathematics test at Grade 8 is quite different from the content at Grade 6. For both of these students, what will be important is whether or not their developmental scale scores over the next several years indicate improved performance.

Achievement Level Classifications

Based on their scale scores, students are assigned one of five *Achievement Level Classifications*. Achievement Levels are ranges of scores within the 100 to 500 point FCAT scale (or, after conversion, within the developmental scale). The *cut point scores* (numerical borders) between each level were established by a special committee, the Standards Setting Committee comprised of Florida educators, as well as DOE staff, the Florida Education Commissioner, and the State Board of Education. The levels range from the lowest level (Level 1) to the highest level (Level 5). Determining a student's Achievement Level classification involves locating the score in one of the five Achievement Levels. Table 13 on the next page presents the developmental scale score ranges for each Achievement Level for FCAT Reading and FCAT Mathematics for all grades tested. Achievement Levels will be reported for FCAT Science beginning in 2006 and for FCAT Writing+ beginning in 2007. See Section 4.2 and Appendix D for more information about the Standards Setting Committees.

TABLE 13: ACHIEVEMENT LEVELS IN FCAT READING AND FCAT MATHEMATICS (DEVELOPMENTAL SCALE SCORES)

Reading						Mathematics				
Level 1	Level 2	Level 3	Level 4	Level 5	Grade	Level 1	Level 2	Level 3	Level 4	Level 5
86–1045	1046–1197	1198–1488	1489–1865	1866–2514	3	375–1078	1079–1268	1269–1508	1509–1749	1750–2225
295–1314	1315–1455	1456–1689	1690–1964	1965–2638	4	581–1276	1277–1443	1444–1657	1658–1862	1863–2330
474–1341	1342–1509	1510–1761	1762–2058	2059–2713	5	569–1451	1452–1631	1632–1768	1769–1956	1957–2456
539–1449	1450–1621	1622–1859	1860–2125	2126–2758	6	770–1553	1554–1691	1692–1859	1860–2018	2019–2492
671–1541	1542–1714	1715–1944	1945–2180	2181–2767	7	958–1660	1661–1785	1786–1938	1939–2079	2080–2572
886–1695	1696–1881	1882–2072	2073–2281	2282–2790	8	1025–1732	1733–1850	1851–1997	1998–2091	2092–2605
772–1771	1772–1971	1972–2145	2146–2297	2298–2943	9	1238–1781	1782–1900	1901–2022	2023–2141	2142–2596
844–1851	1852–2067	2068–2218	2219–2310	2311–3008	10	1068–1831	1832–1946	1947–2049	2050–2192	2193–2709

Achievement Level classifications provide a clearer statement than the scale score in regard to a student's performance. For schools, districts, and the state, monitoring changes in the percentages of students in each level provides a convenient method of comparing progress over time.

Quality Assurance Measures—One statistical review conducted after operational testing is accuracy and consistency of the Achievement Level classifications. Because placement in or above a specified Achievement Level is a requirement for high school graduation (on Grade 10 FCAT Reading and Grade 10 FCAT Mathematics) and is also used in decisions regarding promotion from Grade 3 to Grade 4, the accuracy and consistency of these classifications is extremely important.

Table 14 lists the major statistical indicators generated for each test. For a more detailed discussion of these indicators, refer to Chapter 4.0, Test Development and Construction, and Appendix A.

TABLE 14: STATISTICAL INDICATORS REVIEWED AFTER OPERATIONAL TESTING

Characteristic	Indicator
Appropriate Level of Difficulty	*p*-values IRT *b*-parameters, Test Characteristic Curves (TCC)
Item-Test & Item-Strand Correlations	Item-total correlations, biserial correlations, IRT *a*-parameters, TCC
Minimal Gain from Guessing	IRT *c*-parameters, TCC
Fit to IRT Model	Q_1 (Z_{Q1}) fit statistics
Statistical Bias & Other Non-content Influences	Differential Item Functioning (DIF) analysis (Mantel-Haenszel statistic, Mantel statistic, SMD rating)
Reliability	Test information curves, SEM curves Marginal reliability index, Cronbach's alpha
Unidimensionality of Achievement Scale	Q_3 statistics
Accuracy and Consistency of Achievement Level Classification	Indices of overall, conditional-on-level, and by-cut-point accuracy and consistency