# FCAT Handbook—A Resource for Educators

FCAT
Florida Comprehensive Assessment Test

# FLORIDA DEPARTMENT OF EDUCATION



John L. Winn
Commissioner of Education

Just Read, Florida!

Dear Florida Educator,

The *FCAT Handbook* provides Florida educators with a broad spectrum of information on many aspects of the Florida Comprehensive Assessment Test (FCAT). As the FCAT has grown in terms of subjects, grades, and students tested, as well as its prominence in Florida's system of school accountability, it has become increasingly important for educators to have a thorough understanding of what the FCAT measures, its methodology, and what its results mean. Such an understanding is crucial for explaining test results to students and their parents and for building confidence in the program.

One prominent feature of the FCAT that is highlighted throughout the *Handbook* is the extent to which Florida educators and other citizens are involved in every aspect of the test, from fine-tuning test items to establishing criteria for scoring student work. Extensive educator involvement helps ensure that the FCAT is an accurate reflection of the learning goals that Florida educators established for their students, contained in the *Sunshine State Standards*.

The attention given to the FCAT and its results should not obscure the fact that it is only one of several factors used to measure the progress of students, schools, and districts. Other important factors include locally-developed expectations and classroom-level assessments of student progress. While no single test can completely describe student and school progress, the various FCAT components (reading, mathematics, science, and writing) work together to provide an effective means for assessing progress and for guiding instruction.

I hope you find this publication a useful resource. Thank you for your efforts to help all Florida students reach the high standards that have been set for educational achievement.

Sincerely,

John L. Winn

John Winn
Commissioner, Florida Department of Education

# TABLE OF CONTENTS

# INDEX OF TABLES

# INDEX OF FIGURES

# PROLOGUE: THE EDUCATOR'S ROLE IN THE FCAT PROCESS

The *FCAT Handbook—A Resource for Educators (Handbook)* is written primarily for educators, but should be informative for anyone interested in the various aspects of the Florida Comprehensive Assessment Test (FCAT). The *Handbook* can serve as a reference manual for those seeking a more thorough understanding of the FCAT and for those looking only for specific information. The chapters and sections are structured to facilitate reading the *Handbook* from cover to cover; however, the organization also facilitates its use as a reference to other sources of information about the FCAT.

Included throughout the *Handbook* are profiles of people who are involved with the FCAT program. Most are classroom teachers or administrators in Florida's public schools who have served on FCAT committees. Educator involvement in the FCAT development, administration, and scoring processes is identified with an icon like the one displayed to the left.

To ensure that the FCAT is an accurate measure of the *Sunshine State Standards*, Florida educators are encouraged to become familiar with the FCAT process, remain up to date on new developments, and provide feedback via committee participation. This *Handbook* is intended to provide important background information, including further explanations of the role of educators in the FCAT process. News about the program and additional updates are posted regularly on the FCAT web site (www.firn.edu/doe/sas/fcat.htm).

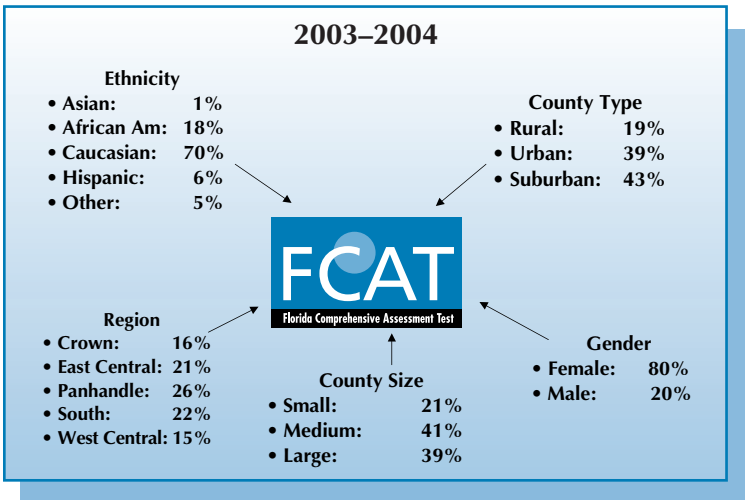
For some Florida educators, much of the information in this *Handbook* may be new; however, the development and implementation of the FCAT have been shaped by the active involvement of thousands of Florida educators serving on FCAT Committees. Since 1995, educators have guided the development of the *Sunshine State Standards (Standards* or *SSS),* the determination of which

benchmarks to assess and how to assess them on the FCAT, and how essays as well as other performance tasks should be scored. In addition, all FCAT test items are reviewed and accepted by committees of Florida educators.

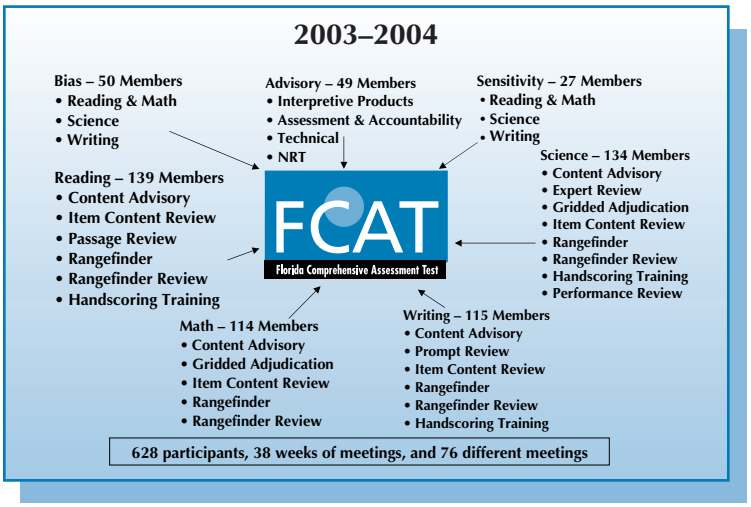
Figures 1 and 2 illustrate the extent to which the FCAT is guided annually by Florida educators. From July 1, 2003, through June 30, 2004, the Florida Department of Education (the DOE) convened and facilitated 76 different committee meetings involving more than 600 participants, representing 63 of Florida's 67 counties. A balanced representation on the basis of gender, ethnicity, geographic location, and district size is also considered when forming committees.

Other committee participants include Florida citizens who share a stake in the education of Florida's children as well as local and national experts in psychometrics. In this publication, some of the FCAT committee members are featured and quoted.

## Figure 1: FCAT Committee Demographics in 2003–2004




## Figure 2: FCAT Committees in 2003–2004



# 1.0 INTRODUCTION

This *Handbook* provides information about the beginnings of the FCAT, the considerations governing item and test development, the mechanics of item and test scoring, and the meaning of the different FCAT scores. Such an understanding can be useful for helping students prepare for the FCAT and for explaining the test and the test results to students and their parents. Much of the information here has appeared in other publications and on the DOE web site, but this is the first time this information has been consolidated and presented in a single document.

The FCAT measures student achievement of the benchmarks contained in Florida's *Sunshine State Standards,* which were developed with the goal of providing all students with an education based on high expectations. The FCAT supports and provides an objective measure of the *Standards* as the foundation for curriculum and instruction. The FCAT also provides feedback and accountability indicators to Florida educators, policy makers, students, and other citizens.

Administered annually to all Florida public school students in Grades 3–11, the FCAT includes items with a varied range of difficulty and cognitive complexity. A score in Achievement Level 2 or higher on FCAT Reading is now a requirement for student promotion from Grade 3 to Grade 4. Achieving a passing score on Grade 10 FCAT Reading and Grade 10 FCAT Mathematics is a statewide graduation requirement. FCAT results serve as a major source of data for determining the school grades that the DOE assigns and reports annually.

## FCAT development is guided by the active involvement of Florida educators.

Because the FCAT serves so many high-stakes purposes, it is important that FCAT development is guided by the active involvement of Florida educators. The DOE maintains open communication with Florida educators regarding how the FCAT and the various associated processes and activities might be improved. The DOE also ensures that the test meets external quality standards for assessments, such as "Standards for Educational and Psychological Testing" (1999) by the American Educational Research Association (AERA). As an indication of quality, *Education Week* (2004)[1] awarded an "A" to Florida for its standards and accountability policies.

[1] Skinner, Ronald A. and Staresina, Lisa N., *Education Week Special Report,* "Quality Counts 2004: State of the States," January 8, 2004. URL: http://counts.edweek.org/sreports/qc04/ (free registration required).

First administered in 1998, the FCAT program has become an integral part of Florida's public education system; however, the FCAT is only one component of Florida's quest for higher standards. Other important components include classroom tests and grades, as well as the standards and measures established by individual teachers, schools, and districts. Because it was developed at the state level, the FCAT is the component for which local educators and administrators may need more information.

Key topics covered in this publication include:

- educators' roles in the process of creating a large-scale assessment;
- background and history of Florida K–12 testing;
- test format and content;
- test development;
- test construction;
- test administration; and
- scoring and reporting results.

For more information on these and other topics, refer to the Guide to Related Resources found in Chapter 9.0.

Information about several important topics **not** covered in the *Handbook* can be found at the web sites listed in Table 1.

**TABLE 1: GETTING MORE INFORMATION ON TOPICS NOT COVERED**

| Topics Not Covered | Web Sites for More Information |
|---|---|
| Norm-referenced assessments in the FCAT program (NRT) | http://www.firn.edu/doe/sas/nrthome.htm |
| Results and trends over time | http://fcat.fldoe.org |
| Florida's activities under the federal No Child Left Behind Act (NCLB) | http://www.fldoe.org/NCLB |
| School grades and other accountability measures | http://www.firn.edu/doe/evaluation/home0018.htm |

Although some of the information about the FCAT is technical, the *Handbook* is written for those without specialized knowledge of psychometrics. Technical information is presented at the conceptual level first, as well as in the context of its relevance to the test.

### Note on Criterion-Referenced Tests and Norm-Referenced Tests

The FCAT consists of two types of tests: norm-referenced tests (NRT) in reading and mathematics, which compare the achievement of Florida students with that of their peers nationwide; and criterion-referenced tests (CRT) in reading, mathematics, science, and writing, which measure student progress toward meeting the *Sunshine State Standards* benchmarks. As illustrated in Table 2 below, both the FCAT *SSS* and the FCAT NRT are used to measure achievement and guide instruction of individual students.

**TABLE 2: COMPARISON OF THE FCAT *SSS* AND THE FCAT NRT***

| FCAT *SSS* | FCAT NRT |
|---|---|
| Scores provided relate to Florida's *Sunshine State Standards* benchmarks. | Scores provided relate the performance of Florida students to that of other students nationwide. |
| Measures achievement in reading, mathematics, science, and writing. | Measures achievement in reading comprehension and mathematics problem-solving. |
| Grades assessed are 3–10 in reading and mathematics. | Grades assessed are 3–10 in reading and mathematics. |
| Grades assessed are 4, 8, and 10 in writing. | Writing is not assessed. |
| Grades assessed are 5, 8, and 11 in science. | Science is not assessed. |
| Includes multiple-choice, gridded-response, and performance task items. | Includes only multiple-choice items. |
| Mathematics portion measures a wide range of skills and problem-solving methods. | Mathematics portion measures a wide range of skills and problem-solving methods. |
| Reading portion measures vocabulary and literary elements, along with the *Sunshine State Standards.* | Reading portion measures vocabulary in context. |
| Calculators may be used for the mathematics assessment in Grades 7–10. | Calculators may be used for the mathematics problem solving test in Grades 7–10. |
| Rulers and other measuring devices are not used for the mathematics assessment. | Rulers are used for mathematics measurement items in the problem-solving test. |
| Mathematics reference sheets are provided for students in Grades 6–10. | Mathematics reference sheets are provided for students in Grades 7–10. |

* This comparison is based on the FCAT NRT, which is the *Stanford Achievement Test, Tenth Edition* (*Stanford 10* or *SAT 10*), administered as of the date of this publication.

**Bonnie C. Atwater**
(Assessment; Administration)
District Coordinator of Assessment, Duval County Public Schools
Jacksonville, Florida

**FCAT Committee Experience:** FCAT NRT Advisory; Bias Review; FCAT RFP Committee (FDOE)

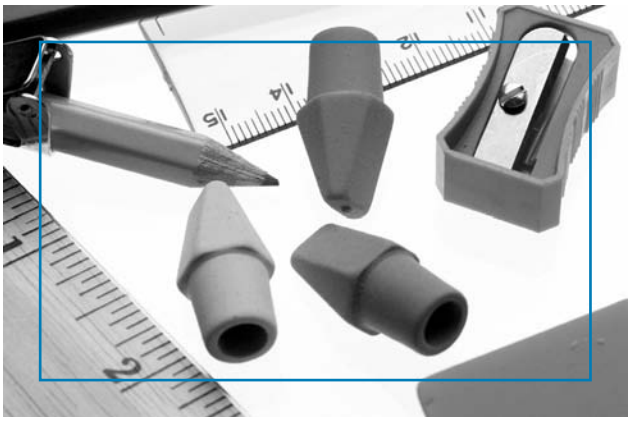
**Related Experience:** Florida Association of Test Administrators (FATA)

**"As a member of the FCAT NRT Advisory Committee, I have seen how DOE staff and contractors work diligently to incorporate district and community needs into the assessment program. Their responsiveness motivates me to encourage schools and community members to offer constructive comments that will benefit all students."**

The FCAT NRT provides information to help ensure that Florida students are keeping pace with their peers nationally. Comparing Florida students to those around the nation requires that the NRT not be too closely aligned with the curriculum of any one state, so the NRT is not necessarily aligned with the *Sunshine State Standards*. From 2000–2004, the test used for the NRT was the *Stanford Achievement Test, Ninth Edition*© (*Stanford 9* or *SAT 9*), published by Harcourt Assessment, Inc. Beginning in 2005, the *Stanford Achievement Test, Tenth Edition*© (*Stanford 10* or *SAT 10*) will be used for three to five years.

In the remainder of this *Handbook*, the term "FCAT" is used to refer only to the CRT portion, or the FCAT *SSS*. The FCAT *SSS* is based explicitly on the learning goals that Florida educators have identified in the *Sunshine State Standards* and is developed, administered, and scored with the active participation of hundreds of Florida educators and citizens.

For more information about the FCAT NRT, refer to the DOE web site, www.firn.edu/doe/sas/nrthome.htm.



# 2.0 BACKGROUND

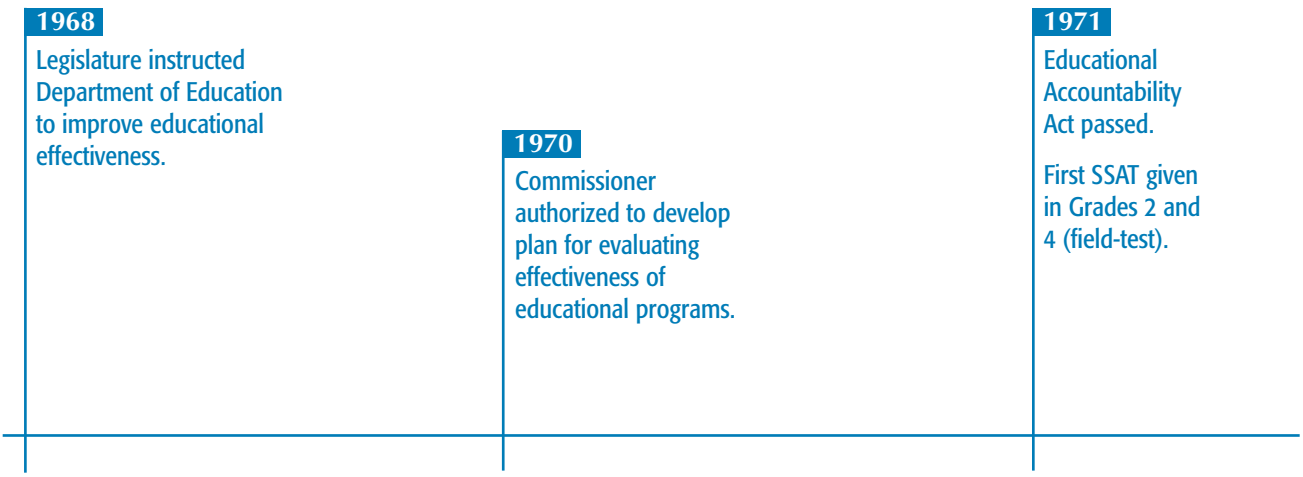
The FCAT is the latest and most comprehensive initiative in Florida's continuously developing system of statewide educational assessment. It is the result of numerous expansions and refinements of the original vision for a statewide system of educational accountability set forth in the Educational Accountability Act of 1971 (Section 229.57, Florida Statutes). The FCAT is similar to Florida's historical educational assessments in that it is a test of student achievement; however, the implementation of the FCAT was influenced by two recent trends in educational assessment: an emphasis on rigorous and clearly defined state-level standards and an emphasis on regular assessments (i.e., annually for a range of grades) of those standards. A more detailed history of Florida's statewide assessment program can be found at www.firn.edu/doe/sas/hsaphome.htm.

## 2.1 The Educational Accountability Act of 1971

The 1971 Act created the statewide assessment program by requiring:

- the establishment of basic, specific, uniform statewide educational objectives for each grade level and subject area, including, but not limited to, reading, mathematics, and writing; and
- the development and administration of a uniform and regularly administered statewide assessment to determine pupil status, pupil progress, and the degree to which pupils had achieved established educational objectives.

The resulting educational objectives included only minimum requirements, in contrast to the more extensive, detailed, and rigorous standards that have since evolved. The 1971 assessment included only a criterion-referenced test (CRT) component for reporting Florida students' progress in meeting Florida-specific objectives.



The first statewide assessment, called the State Student Assessment Tests (SSAT), was given in reading, writing, and mathematics in Grades 2 and 4 in 1971 and in Grades 3, 6, and 9 in 1972. The assessments collected data on representative samples of Florida students in each grade level tested, providing useful information at the state and district levels but not at the school or student level.

## 2.2 Expansions and Enhancements

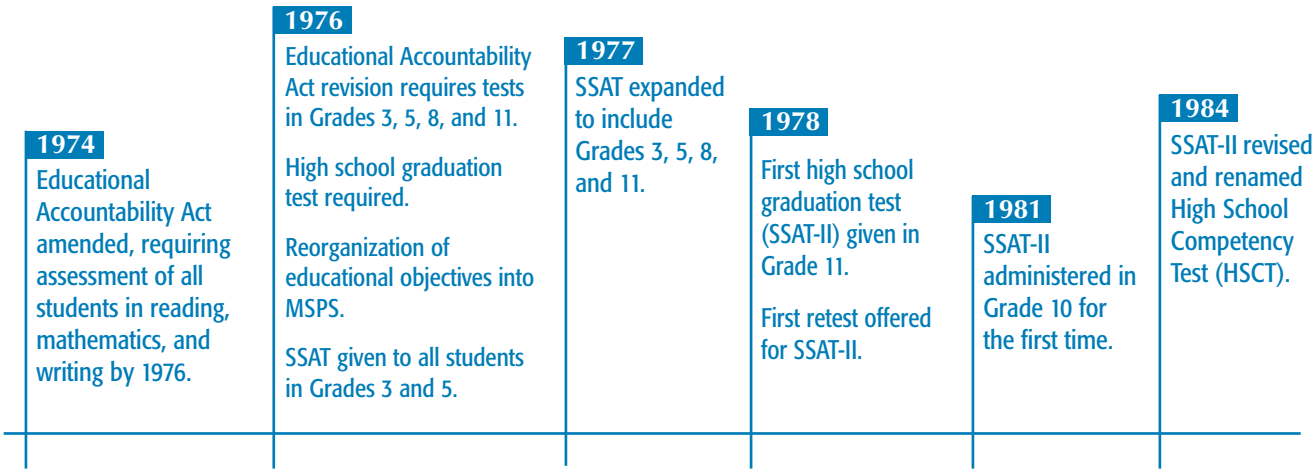
The need for school- and student-level data was quickly realized, and in 1974, the Educational Accountability Act was amended to require the assessment of all students in reading, mathematics, and writing by 1976.

In 1976, the Florida Legislature expanded the Educational Accountability Act to require assessments in Grades 3, 5, 8, and 11 and the nation's first high school graduation test, a functional literacy test, to be given in Grade 11. The Act also called for organizing educational objectives used in test development into Minimum Student Performance Standards (MSPS), which would have wider applications for curriculum and instructional planning.

The Grade 11 graduation test, which became the State Student Assessment Test, Part II (SSAT-II), was changed to Grade 10 in 1981 to allow additional opportunities for students to pass the test. After substantial revisions in 1984, the name of the test was changed to the High School Competency Test (HSCT). In 1992, the test was moved back to Grade 11.

In 1992, the Florida Writing Assessment Program was introduced in the format of a single, extended writing task based on a prompt. Originally administered only in Grade 4, the assessment was also administered in Grade 8 in 1993 and in Grade 10 in 1994.

Also in 1992, the Florida assessment program included the Grade Ten Assessment Test (GTAT), which was a customized, norm-referenced, multiple-choice test in reading comprehension and mathematics given in Grade 10. The GTAT ended with the 1996 administration.



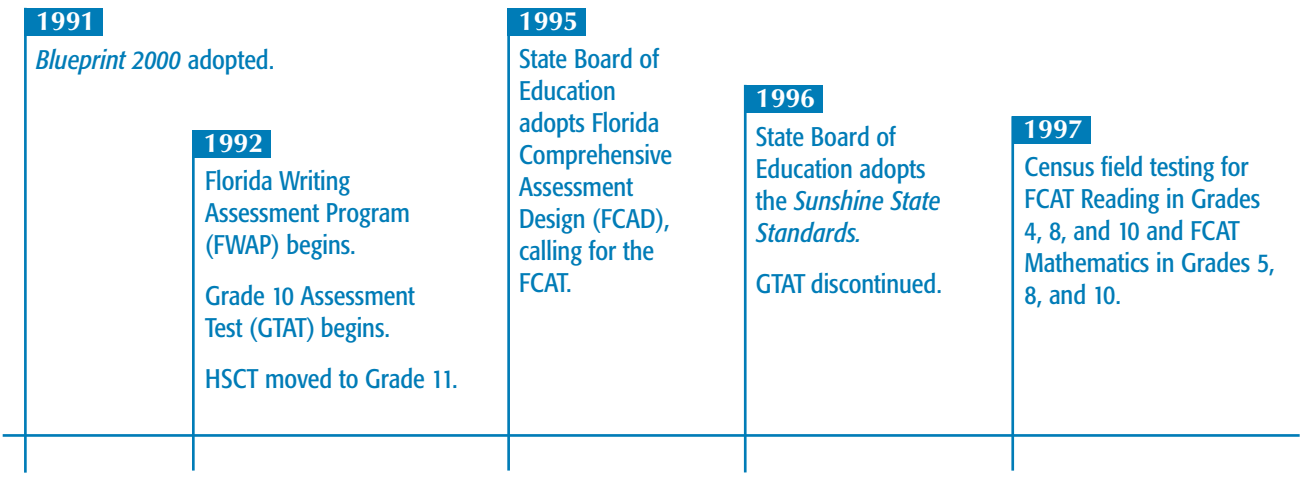
## 2.3 Birth of the FCAT: Florida Comprehensive Assessment Design and the *Sunshine State Standards*

The School Improvement and Accountability Act of 1991 called for sweeping changes by defining seven innovative and challenging goals for Florida's public educational system. The goals were further delineated by the Florida Commission on Education Reform and Accountability and were disseminated in *Blueprint 2000*. Goal 3 of the legislation was dedicated to improving student performance and included ten standards. The first four of these standards correspond to reading, writing, mathematics, and thinking skills.

The ten standards from *Blueprint 2000* were reinforced in 1995 by the Florida Comprehensive Assessment Design (FCAD), created by the Florida Commission on Education Reform and Accountability. The FCAD called for formal development of a new statewide assessment system as part of an overall strategy to increase student achievement. This assessment system, which would be called the Florida Comprehensive Assessment Test (FCAT), was based on the first four standards of *Blueprint 2000's* Goal 3.

The FCAD was followed a year later by the adoption of the *Sunshine State Standards*, a set of learning expectations, driven by Goal 3, in seven content areas (language arts, mathematics, science, social studies, health and physical education, foreign languages, and the arts) and in four separate grade clusters (PreK–2, 3–5, 6–8, and 9–12). The benchmarks of the *Sunshine State Standards,* which represent the skills and knowledge deemed essential for Florida students, became the foundation for the FCAT.

In 1997, the FCAT was census field-tested for the first time in Grades 4 (reading), 5 (mathematics), 8 and 10 (for reading and mathematics). A *census field test* means the entire eligible population is tested. Other field tests include only select populations, a sampling of the eligible test population. Item types included multiple-choice, gridded-response (for mathematics), and performance task (short- and extended-response) items, or questions. Within a few years, the existing Florida Writing Assessment Program (FWAP) was incorporated into the FCAT and became known as FCAT Writing.



## 2.4 A+ Plan for Education

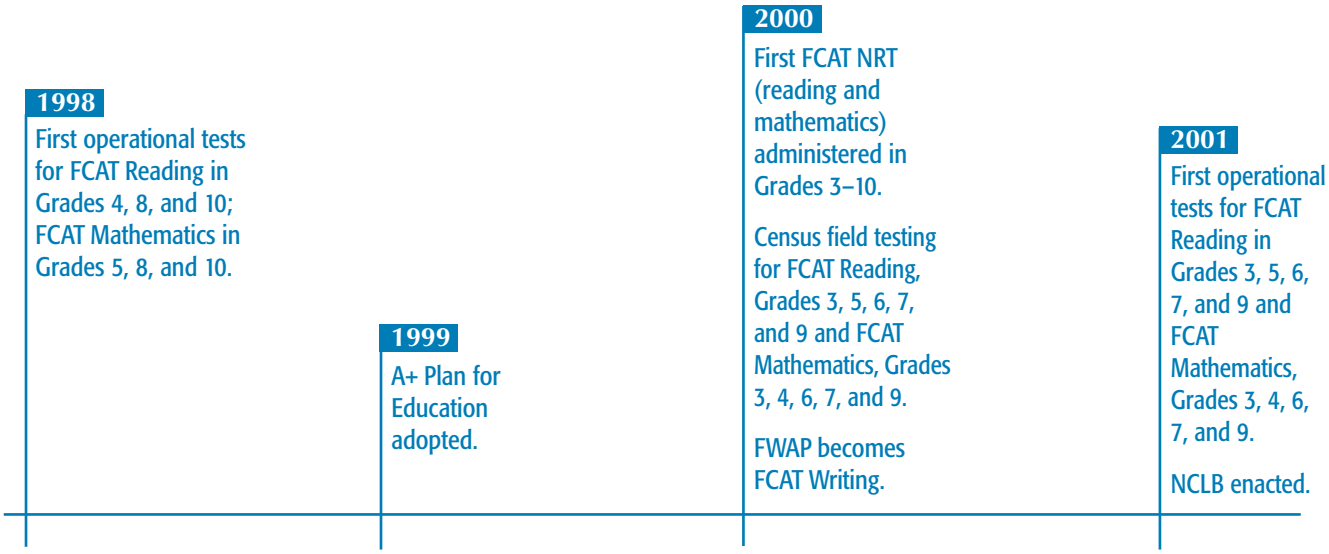
Approved by the Florida Legislature in 1999, the A+ Plan for Education expanded Florida's statewide assessment program to include the assessment of reading and mathematics in Grades 3–10, a science assessment (FCAT Science), and a system for calculating the academic growth of each student over time. It also required students to pass the Grade 10 FCAT *SSS* in reading and mathematics in order to graduate from high school. As a result of these changes, the *Sunshine State Standards* were further defined to include Grade-Level Expectations (GLEs) for Grades 3–8 in language arts, mathematics, science, and social studies.

## 2.5 No Child Left Behind

The No Child Left Behind Act of 2001 (NCLB) required the assessment of all students in Grades 3–8 in reading and mathematics. Because the FCAT assesses reading and mathematics in Grades 3–10, Florida already had an assessment system in place to provide the Adequate Yearly Progress (AYP) data required by the Act. Although NCLB increased emphasis on the FCAT and required new types of analyses, it did not require any major changes to the FCAT's content, development process, or administration.

## 2.6 FCAT Writing+

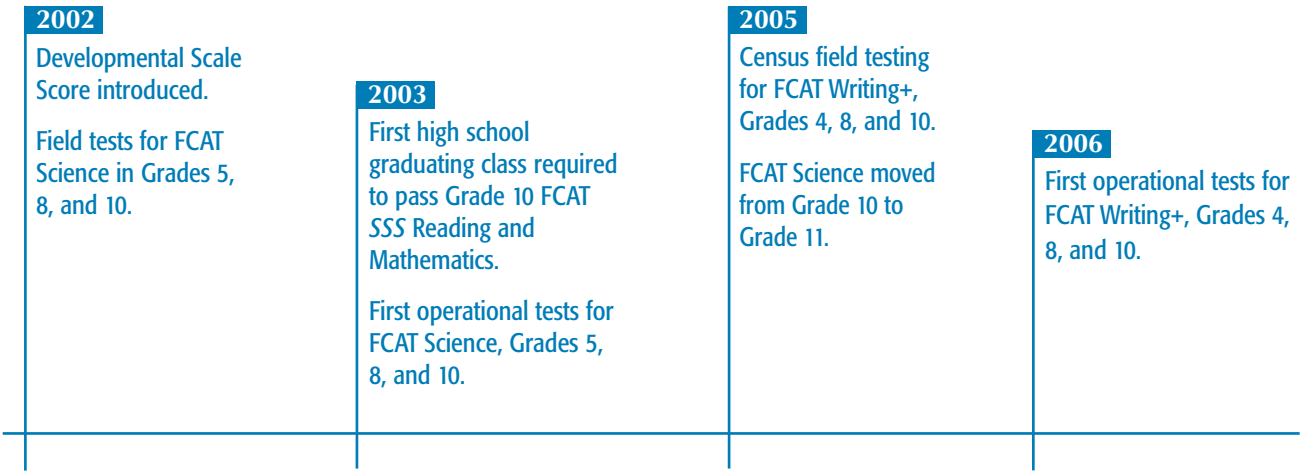
The DOE is supplementing the FCAT Writing essay test with multiple-choice items. Items were field tested on all eligible Florida students in Grades 4, 8, and 10 in February 2005. Since a multiple-choice component is being added, the test was renamed "FCAT Writing+ (plus)." The first operational administration of FCAT Writing+ (essay plus multiple-choice items) will be in February 2006. In this *Handbook,* the writing assessment will be referred to as "FCAT Writing+."



The purpose for adding a machine-scored section to FCAT Writing+ is to allow writing performance to be used to satisfy the state's graduation requirement and also to provide a more comprehensive assessment of writing. Although all decisions about FCAT Writing+ will not be finalized until after the 2005 field test, the FCAT Writing Content Advisory Committee recommended that a 100–500 whole-test scale score be reported, as well as subscores (a rubric score of 0 to 6) for the essay and for the categories of focus, organization, support, and conventions. Student scores on FCAT Writing+ will be reported for the first time in May 2006.

## 2.7 FCAT Science

The A+ Plan for Education passed by the Florida Legislature in 1999 required a science assessment for students in Grades 5, 8, and 10. Development of science test items began in 2000, and a field test of these items was conducted in a representative sample of Florida schools in April 2002. The first operational assessment and reporting of student scores took place in May 2003. Beginning in March 2005, FCAT Science was administered in Grade 11 instead of Grade 10. This change was in response to requests by Florida science educators to allow an additional year for students to receive high-school level science instruction.

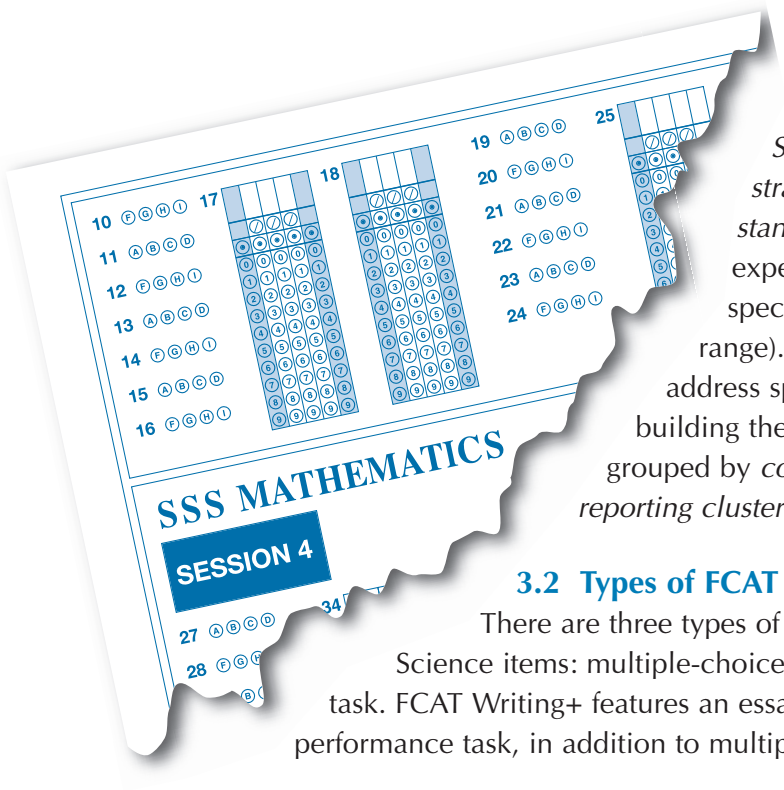


# 3.0 Test Content and Format

The FCAT is administered to students on regular school days under the supervision of each school's staff. FCAT Reading, FCAT Mathematics, and FCAT Science are given on specific days within a two-week period in the spring. The FCAT NRT is also administered during the same time frame. FCAT Writing+ is administered to students in Grade 4 over a two-day period and to Grades 8 and 10 in a single day in February. All test forms are printed in English only.

## 3.1 Test Format

FCAT items[2] are based directly on individual benchmarks found in the *Sunshine State Standards*. Within each subject, items are developed to represent the complete range of content associated with the benchmarks. A few benchmarks are not easily assessed within the time limitations or through the format of FCAT items and, therefore, are not assessed on the test. Because of the FCAT's direct link to the *Standards*, students who have mastered the *Standards* and have practiced with FCAT item formats should perform well on the FCAT.




Within subjects and established grade ranges (i.e., PreK–2, 3–5, 6–8, 9–12), there are three categories for *Sunshine State Standards* expectations: *strand* (broad category of knowledge), *standard* (general statement of expectation), and *benchmark* (more specific level of expectation for each grade range). All FCAT items are designed to address specific benchmarks. For the purpose of building the test, scoring, and reporting, items are grouped by *content clusters* (sometimes called *reporting clusters* or strands).

## 3.2 Types of FCAT Items

There are three types of FCAT Reading, Mathematics, and Science items: multiple-choice, gridded-response, and performance task. FCAT Writing+ features an essay component, which is considered a performance task, in addition to multiple-choice items. These item types

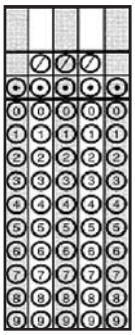
[2] In assessment terminology, an *item* is any question, essay prompt, or other task to which a student is expected to respond. Not all items are presented as questions, so the term *item* is used.

differ not only in format, but in the amount of time students should need to respond to them and in the number of points a correct response to each item is worth.[3] The time estimates for item types are used to establish the test administration schedule and to ensure that students have ample time to complete the test. Although the FCAT is a timed test, the time allotted is intended to be sufficient for almost all students. In students' test booklets, special icons are used to identify gridded-response items, short-response performance tasks, and extended-response performance tasks.

***Multiple-choice items*** (FCAT Reading, Reading Retakes, Mathematics, Mathematics Retakes, Science, Writing+)—Students choose the correct answer from three or four possible choices and mark the choice by filling in a bubble in the test booklet or answer document. Three-option multiple-choice items are found only in FCAT Writing+. (See Section 3.8 for more information about FCAT Writing+ multiple-choice items.) Multiple-choice items require approximately one minute to answer and are each worth one raw score point.

Mathematics and science gridded-response icon

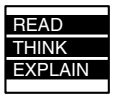
***Gridded-response items*** (FCAT Mathematics, Mathematics Retake, Science)—Students solve problems or answer questions requiring a numerical response and bubble or mark their numerical answers in response grids. Answers may be gridded using several correct formats. Students must accurately fill in the bubbles below the grids to receive credit for their answers. Students are provided with detailed instructions for filling in the bubbles in the *FCAT Sample Test Materials.* Additional instructions are also included in the front of the test book. Gridded-response items require approximately one and a half minutes to answer and are each worth one raw score point.



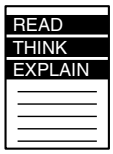
Sample answer grid for Grades 6–10 mathematics and science

## Performance Tasks

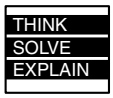
***Short- and extended-response items*** (FCAT Mathematics, Reading, Science)—Students respond to items in their own words or show their solutions to problems. Short-response tasks require approximately five minutes to complete, and students may receive a raw score of 0, 1, or 2 points. Extended-response tasks require approximately 15 minutes to complete, and students may receive a raw score of 0, 1, 2, 3, or 4 points.



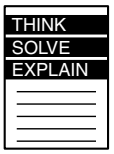

Reading short-response icon



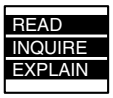

Reading extended-response icon



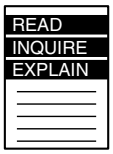

Mathematics short-response icon




Mathematics extended-response icon




Science short-response icon




Science extended-response icon

[3] Multiple-choice and gridded-response items are worth one point each; short-response performance tasks are worth two points each; extended-response performance tasks are worth four points each; and essay responses are worth six points each.

***Prompted essay*** (FCAT Writing+)—Each FCAT Writing+ prompt has two parts: the *writing situation* and the *directions for writing*. The *writing situation* orients the students to the subject about which they are to write. The *directions for writing* guide the students to think about the topic before they begin to write. Essays are scored on a scale ranging from 0 points (unscorable) to 6 points. Students are given 45 minutes to complete their writing.

Calculators are provided to students in Grades 7 and higher on the mathematics and science portions of the FCAT. (See pages 28 and 34.) Dictionaries[4] and other reference materials are **not** allowed on any test at any grade level.

## 3.3 Cognitive Complexity

The benchmarks in the *Sunshine State Standards* identify knowledge and skills that students are expected to acquire, with the underlying expectation that students also demonstrate critical thinking. Goal 3, Standard 4 of Florida's *System of School Improvement and Accountability* makes this expectation clear:

> **"Florida students use creative thinking skills to generate new ideas, make the best decisions, recognize and solve problems through reasoning, interpret symbolic data, and develop efficient techniques for lifelong learning."**

The degree of challenge of FCAT items is currently categorized in two ways: cognitive complexity and item difficulty. Cognitive complexity refers to the cognitive level associated with the item. Since the inception of the FCAT, Bloom's Taxonomy[5] has been used for this purpose; however, Bloom's Taxonomy is difficult to use because it requires an inference about the skill, knowledge, and background of the students responding to the

[4] Limited English proficient (LEP) students may use an English-to-heritage-language dictionary, but not an English language dictionary. For more information on LEP accommodations: http://www.firn.edu/doe/omsle/pdf/lep_factsheet.pdf

[5] Bloom, B.S., et al. *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain.* New York: McKay, 1956.

item. Beginning in 2004, a new cognitive classification system is being used that is based, in part, on Dr. Norman L. Webb's work with "Depth of Knowledge" levels.[6] This change in classification systems has not changed the difficulty of the FCAT.

The transition to a new cognitive classification system was made to focus on the expectations of the item, not the ability of the student. The demands on thinking that an item makes—that is, what it asks the student to recall, understand, reason about, and do—are determined with the assumption that the student is familiar with the knowledge and skills the item assesses.

The categories—low complexity, moderate complexity, and high complexity—form an ordered description of the cognitive demands an item makes on a student. Items at the low level of complexity require a simple skill, such as locating details in a text or solving a one-step problem. At the moderate level, an item can ask the student to summarize a passage or retrieve information from a graph and use it to solve a problem. At the high level, an item may require a student to analyze cause-and-effect relationships or justify a solution to a problem. The distinctions made in item complexity are intended to provide a balance across the tasks administered at each grade level.

Item difficulty has two meanings, depending on the stage of item development. At the item review stage (before use on the test), item difficulty is based on professional judgment about how hard an item is for students working at grade level. At this point, items are classified as easy, medium, or hard. After field testing, item difficulty refers to the percentage of students who actually chose the correct answer. At this stage, item difficulty is referred to as the *p*-value. (See Chapter 4.0 or Appendix A for more information about *p*-values.)

While an item can be classified as having a low level of challenge, in terms of cognitive complexity, it can still be difficult in terms of *p*-value. In general, if 70 percent or more of the students answered the item correctly, it is considered easy. If 40–69 percent of the students answered the item correctly, it is considered average. If less than 40 percent of the students answered the item correctly, it is considered challenging.

[6] Webb, N.L., (1999). *Alignment Between Standards and Assessment,* University of Wisconsin Center for Educational Research.

## 3.4 Test Forms, Operational Items, Field-Test Items, and Anchor Items

When taking the FCAT, all students of the same grade level respond to a common set of items on each test. These common items are called *operational items* and count toward students' scores. Either *field-test items* or *anchor items* are also found on all students' tests, but do not count toward students' scores. Field-test items are administered to students only to gather data on the items. Items found to be acceptable may be considered for future use on the FCAT operational test. Anchor items are those which have appeared on the FCAT in previous years and are used to ensure that the scores on the test can be equated or made comparable from year to year.[7]

**TABLE 3: FCAT BY SUBJECT, GRADE, AND ITEM TYPE**

| Grade | Reading | Writing+* | Mathematics | Science |
|---|---|---|---|---|
| 3 | MC | | MC | |
| 4 | MC, SR, ER | WP, MC | MC | |
| 5 | MC | | MC, GR, SR, ER | MC, SR, ER |
| 6 | MC | | MC, GR | |
| 7 | MC | | MC, GR | |
| 8 | MC, SR, ER | WP, MC | MC, GR, SR, ER | MC, GR, SR, ER |
| 9 | MC | | MC, GR | |
| 10 | MC, SR, ER | WP, MC | MC, GR, SR, ER | |
| 11 | | | | MC, GR, SR, ER |
| Retake | MC | | MC, GR | |

**Key**
**MC** multiple-choice
**GR** gridded-response
**SR** short-response performance task
**ER** extended-response performance task
**WP** writing prompt or prompted essay

* Beginning with the field test in 2005, FCAT Writing+ includes multiple-choice items at the same grade levels, in addition to the prompted essay.

Table 3, above, lists the types of items used in each content area, by grade. The next four sections of the *Handbook* provide additional information about the different content areas and detail the knowledge and skills assessed in each area. Examples of sample test items are included. In addition, certain content-specific features of the FCAT are examined, such as calculator use at given grade levels, and the types of reading passages that are used on the test.

[7] Prior to 2004, anchor items counted toward students' scores.

## 3.5 Reading Content

FCAT Reading employs a wide variety of written material to assess students' reading comprehension as defined in the *Sunshine State Standards.* FCAT Reading is composed of about 6–8 reading passages with sets of 6–11 items based on each passage. There are two types of reading passages: informational and literary.

Informational passages provide readers with facts about a particular subject and may include magazine and newspaper articles, editorials, and biographies. Literary passages are written primarily for readers' enjoyment and may include short stories, poems, folk tales, and selections from novels. Table 4, on the next page, shows the different types of passages students may encounter on the test. Most passages are selected from published sources, although some may be written expressly for the FCAT.

**Max Hutto**
Writing and Reading Supervisor, Middle School Language Arts, School District of Hillsborough County
Tampa, Florida

**FCAT Committee Experience:** Reading Content Advisory; Reading Passage and Item Review; Reading Rangefinder and Rangefinder Review; Reading Standard Setting; Writing Content Advisory; Writing Prompt Review; Prompt Writing; Writing Rangefinder and Rangefinder Review; Writing Item Review; Writing Handscoring Training

**"FCAT has made alignment of the curriculum and training to the *Sunshine State Standards* a must for all districts. Being involved with FCAT at the district level has made me realize the importance of raising expectations for all students and the importance of providing meaningful instruction to help them meet these high expectations. Serving on FCAT committees over the years has helped me to know the importance of all educators working together, both at the state and district levels, to ensure the success of all Florida students."**

The Orlando Sentinel—Florida

February 11, 2003, FINAL

**Study Praises FCAT as Indicator of Learning**

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

| Table 4: Types of Reading Passages | |
|---|---|
| Types of Informational Texts | Types of Literary Texts |
| Subject-area text (e.g., science, history) | Short stories |
| Magazine and newspaper articles | Literary essays (e.g., critiques, personal narratives) |
| Diaries | Excerpts from novels |
| Editorials | Poems |
| Informational essays | Historical fiction |
| Biographies and autobiographies | Fables and folktales |
| Primary sources (e.g., Bill of Rights) | Plays |
| Consumer materials | |
| How-to articles | |
| Advertisements | |
| Tables and graphic presentations of text (e.g., illustrations, photographs, and captions) | |

Table 5 below shows the percentage of FCAT Reading items on a test for literary and informational text, as well as the passage length for each grade level. As students progress beyond the early grades, they will read informational text with increasing frequency in and out of school. The percentage of informational text students will encounter on the FCAT also increases as they progress through the grades. Likewise, the range of words per passage increases across the grade levels.

| Table 5: Distribution of FCAT Reading Test Items by Passage Type and Length | | | | |
|---|---|---|---|---|
| | Percentage Distribution of Reading Test Items by Passage Type | | Number of Words per Passage | |
| Grade | Informational | Literary | Average | Range |
| 3 | 40% | 60% | 350 | 100–700 |
| 4 | 50% | 50% | 400 | 100–900 |
| 5 | 50% | 50% | 450 | 200–900 |
| 6 | 50% | 50% | 500 | 200–1000 |
| 7 | 60% | 40% | 600 | 300–1100 |
| 8 | 60% | 40% | 700 | 300–1100 |
| 9 | 70% | 30% | 800 | 300–1400 |
| 10 | 70% | 30% | 900 | 300–1700 |

## Knowledge and Skills Tested

FCAT Reading is based on the benchmarks found in the Reading and Literature strands of the Language Arts *Sunshine State Standards*. The four reading content clusters used for the FCAT are: (1) Words and Phrases in Context; (2) Main Idea, Plot, and Purpose; (3) Comparison and Cause/Effect; and (4) Reference and Research.

Table 6 indicates the relative emphasis on each cluster by providing the percentage of raw score points available in each cluster assessed on the FCAT at the different grade levels. As students progress through the grades, more emphasis is placed on higher level thinking skills, which predominate in the Reference and Research cluster. Some of the benchmark skills addressed at each grade level are shown on these pages. For more detailed information, refer to the *FCAT Reading Test Item Specifications,* available at http://www.firn.edu/doe/sas/fcat/fcatis01.htm.

Table 6 also indicates a range of percentages for score points in each cluster by grade. This range is necessary because each passage identified for use on FCAT Reading is unique and has varied potential for assessing benchmarks and for the number and type of possible items. Since each year's test has a different selection of passages, the variance in this potential creates shifts in the percentage of score points in a cluster.

**TABLE 6: APPROXIMATE PERCENTAGE DISTRIBUTION OF RAW SCORE POINTS ACROSS FCAT READING CONTENT CLUSTERS BY GRADE LEVEL**

| Grade | Words and Phrases In Context | Main Idea, Plot, and Purpose | Comparison and Cause/Effect | Reference and Research |
|---|---|---|---|---|
| 3–5 | 15–20% | 30–55% | 20–45% | 5–15% |
| 6–8 | 15–20% | 30–55% | 15–25% | 10–30% |
| 9–10 | 15–20% | 20–50% | 10–25% | 20–40% |

FCAT Reading includes multiple-choice items at all grades. At Grades 4, 8, and 10, it also includes short- and extended-response performance tasks, scored using two- or four-point rubrics. Rubrics are the scoring guidelines or criteria used to evaluate all of the FCAT performance tasks and essays. The rubric describes what is required for each possible score point. For example, a short-response task may require the student to describe how a character in a story changes or shows growth. An extended-response task requires a longer and more detailed response, such as a comparison of traits or actions of two different characters. Students are provided eight lines on which to write their answers for short-response items and 14 lines for extended-response items. Table 7, on the next page, presents the number of items per type at each grade level, as well as the total time needed to take a test at each grade level. Sample items are also presented to illustrate each item type. Additional sample items are included in the *FCAT Sample Test Materials* posted on the DOE web site (www.firn.edu/doe/sas/fcat/fcatsmpl.htm).

## Table 7: Number of Reading Items per Item Type and Total Test Time by Grade

| Grade | Multiple-Choice | Performance Tasks | Total Minutes per Test |
|---|---|---|---|
| 3 | 50–55 | 0 | 120 |
| 4 | 45–50 | 5–7 | 160 |
| 5 | 50–55 | 0 | 120 |
| 6 | 50–55 | 0 | 120 |
| 7 | 50–55 | 0 | 120 |
| 8 | 45–50 | 5–7 | 160 |
| 9 | 50–55 | 0 | 120 |
| 10 | 45–50 | 5–7 | 160 |
| Retake | 55–60 | 0 | 160 |

Note: Total testing time is divided into two testing sessions, except for the retake test, which only has one session. Students taking the retake test may receive additional time to complete the test. The data in this table give ranges for the approximate number of items by item type. These ranges include both operational and field-test or anchor items.

**Figure 3: Example of a Grade 8 FCAT Reading Multiple-Choice Item**

According to the story, why do the inhabitants of Earth and Kaan say that this has been the "very best Zoo"?

**A.** Both groups felt safe because of the protective bars.
**B.** Both groups felt the zoo was worth the money spent.
**C.** Both groups considered each other frightening creatures behind bars.
**D.** Both groups considered each other the strangest creatures they had ever seen.

**Figure 4: Example of a Grade 10 FCAT Reading Short-Response Performance Task**

READ THINK EXPLAIN

How did William Fee contribute to the cotton industry and everyday life? Support your answer with details and information from the article.

**Figure 5: Example of a Grade 8 FCAT Reading Extended-Response Performance Task**

READ THINK EXPLAIN

How have sea gulls contributed to or affected the development of Salt Lake City? Use details and information from the article to support your answer.

At Grades 3, 4, and 5, FCAT Reading assesses the following skills:

### *Words and Phrases in Context*

- uses strategies to increase vocabulary through word structure clues (prefixes, suffixes, roots), word relationships (antonyms, synonyms), and words with multiple meanings
- uses context clues to determine word meanings

### *Main Idea, Plot, and Purpose*

- determines main idea or essential message in a text
- identifies relevant details and facts
- recognizes and arranges events in chronological order
- identifies author's purpose in a text
- understands plot development and conflict resolution in a story

### *Comparisons and Cause/Effect*

- recognizes the use of comparison and contrast
- recognizes cause-and-effect relationships
- identifies similarities and differences among characters, settings, and events in various texts

### *Reference and Research*

- uses maps, charts, photos, or other multiple representations of information
- reads, organizes, and interprets written information for various purposes, such as making a report, conducting an interview, taking a test, or performing a task

At Grades 6, 7, and 8, FCAT Reading assesses the following skills:

### *Words and Phrases in Context*

- uses various strategies, including contextual and word structure clues, to analyze words and text
- draws conclusions from a reading text

### *Main Idea, Plot, and Purpose*

- determines the stated or implied main idea or essential message in a text
- identifies relevant details and facts
- recognizes organizational patterns
- identifies and uses the author's purpose and point of view to construct meaning from text
- recognizes persuasive text
- recognizes and understands how literary elements support text (e.g., character and plot development, point of view, tone, setting, and conflicts and resolutions)

### *Comparisons and Cause/Effect*

- recognizes comparison and contrast
- recognizes cause-and-effect relationships

### *Reference and Research*

- locates, organizes, and interprets written information for a variety of purposes
- synthesizes information within or across texts
- checks validity and accuracy of research information
- synthesizes strong versus weak arguments

At Grades 9 and 10, FCAT Reading assesses the following skills:

### *Words and Phrases in Context*

- selects and uses strategies to understand words and text
- makes and confirms inferences from a reading text
- interprets data presentations (e.g., maps, diagrams, graphs, and statistical illustrations)

### *Main Idea, Plot, and Purpose*

- determines stated or implied main idea
- identifies relevant details
- identifies methods of development
- determines author's purpose and point of view
- identifies devices of persuasion and methods of appeal
- identifies and analyzes complex elements of plot (e.g., setting, tone, major events, and conflicts and resolutions)

### *Comparisons and Cause/Effect*

- recognizes the use of comparison and contrast
- recognizes cause-and-effect relationships

### *Reference and Research*

- locates, gathers, analyzes, and evaluates information for a variety of purposes
- selects and uses appropriate study and research skills and tools according to the type of information being gathered or organized
- analyzes the validity and reliability of primary source information and uses the information appropriately
- synthesizes information from multiple sources to draw conclusions

## 3.6 Mathematics Content

### Knowledge and Skills Tested

FCAT Mathematics addresses almost all of the *Sunshine State Standards* benchmarks at the associated grade levels. Most items address a single benchmark, but some items, especially extended-response performance tasks, can address more than one related benchmark. The five mathematics content strands used for FCAT design, scoring, and reporting are the same as the five strands under which the benchmarks are grouped in the *Standards*. The five strands are: (1) Number Sense, Concepts, and Operations; (2) Measurement; (3) Geometry and Spatial Sense; (4) Algebraic Thinking; and (5) Data Analysis and Probability. Table 8, below, shows the relative emphasis on each strand by providing the percentage of raw score points available in each at the different grade levels. At Grades 9 and 10, the Geometry and Spatial Sense strand and the Algebraic Thinking strand have slightly more items than the other three strands. A summary of the content assessed at each grade level is provided on the next few pages. For more detailed information, refer to the *FCAT Mathematics Test Item Specifications* available at http://www.firn.edu/doe/sas/fcat/fcatis01.htm.

**Roberta Dilocker**
Administrator for secondary curriculum; Mathematics Coordinator of Central Region and Secondary Education, Citrus County School District
Inverness, Florida

**FCAT Committee Experience:** Mathematics Content Advisory; Mathematics Item Review; Mathematics Rangefinder; *Lessons Learned* Committee

**Related Experience:** Mathematics Region II Leadership Team; Florida Association of Mathematics Supervisors (FAMS), Secretary

**"Serving on a variety of FCAT committees has provided me with many insights into the FCAT processes. The opportunity to share these experiences with others has greatly influenced both staff development and curriculum alignment projects within our district. Rangefinding committees were most valuable to me as I learned how to design rubrics to objectively assess performance task responses."**

**TABLE 8: APPROXIMATE PERCENTAGE DISTRIBUTION OF RAW SCORE POINTS ACROSS FCAT MATHEMATICS CONTENT STRANDS BY GRADE LEVEL**

| Grade | Number Sense, Concepts, and Operations | Measurement | Geometry and Spatial Sense | Algebraic Thinking | Data Analysis and Probability |
|---|---|---|---|---|---|
| 3 | 30% | 20% | 17% | 15% | 18% |
| 4 | 28% | 20% | 17% | 17% | 18% |
| 5–8 | 20% | 20% | 20% | 20% | 20% |
| 9–10 | 17% | 17% | 25% | 25% | 16% |

## Mathematics Content Tested

FCAT Mathematics assesses the following skills at Grades 3–10:

### *Number Sense, Concepts, and Operations*

- identifies operations (+, –, ×, ÷) and the effects of operations
- determines estimates
- knows how numbers are represented and used

### *Measurement*

- recognizes measurements and units of measurement
- compares, contrasts, and converts measurements

### *Geometry and Spatial Sense*

- describes, draws, identifies, and analyzes two- and three-dimensional shapes
- visualizes and illustrates changes in shapes
- uses coordinate geometry

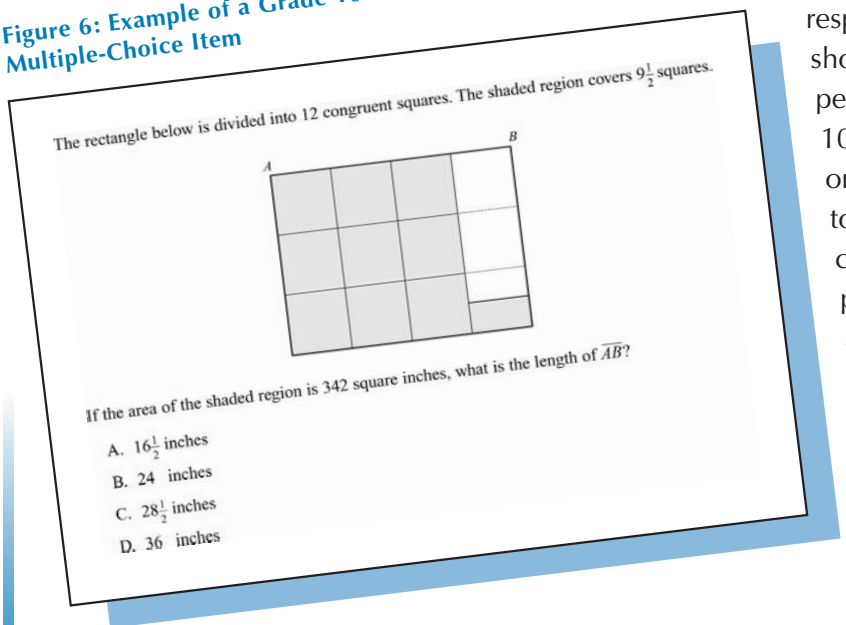
### *Algebraic Thinking*

- describes, analyzes, and generalizes patterns, relations, and functions
- writes and uses expressions, equations, inequalities, graphs, and formulas

### *Data Analysis and Probability*

- analyzes, organizes, and interprets data
- identifies patterns and makes predictions, inferences, and valid conclusions
- uses probability and statistics

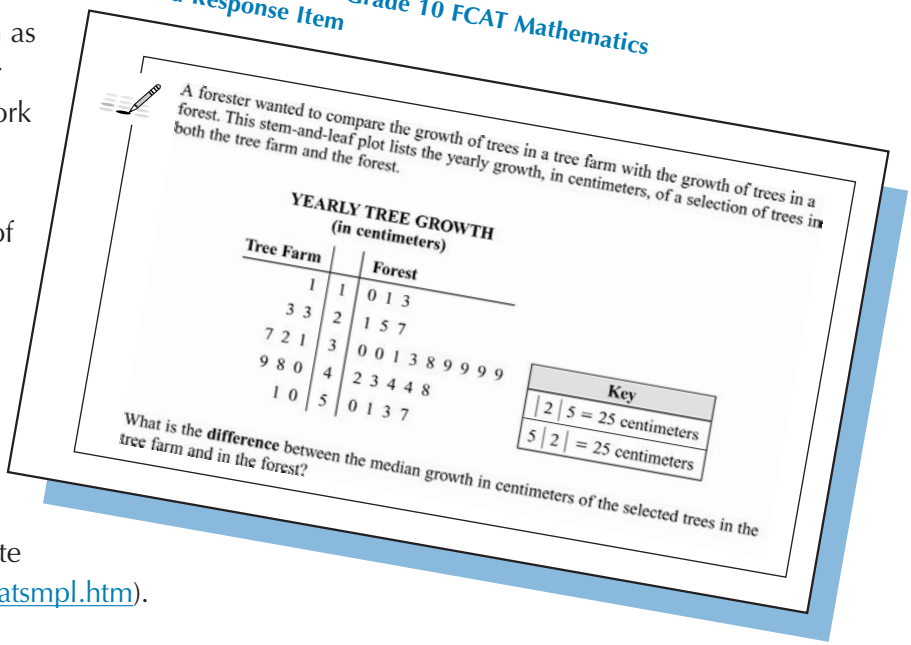
**Figure 6: Example of a Grade 10 FCAT Mathematics Multiple-Choice Item**




FCAT Mathematics includes multiple-choice items in Grades 3–10, gridded-response items in Grades 5–10, and short- and extended-response performance tasks in Grades 5, 8, and 10. Performance tasks, scored on two- or four-point rubrics, require students to read all parts of the question carefully, think about and analyze the problem, determine a way to solve it, and write a detailed solution or describe an answer to the problem in their own words. A short-response performance task may ask for an equation that represents a problem

situation. An extended-response item requires a longer, more detailed response, such as constructing a graph. Answer spaces may include blank work space, charts or graphs, or lined answer space. Table 9, below, displays the number of items per item type and total test time for each grade. Examples of mathematics items are shown on the next few pages. Additional sample items are included in the *FCAT Sample Test Materials* on the DOE web site (www.firn.edu/doe/sas/fcat/fcatsmpl.htm).

**Figure 7: Example of a Grade 10 FCAT Mathematics Gridded-Response Item**


A forester wanted to compare the growth of trees in a tree farm with the growth of trees in a forest. This stem-and-leaf plot lists the yearly growth, in centimeters, of a selection of trees in both the tree farm and the forest.

YEARLY TREE GROWTH
(in centimeters)

| Tree Farm | | Forest |
|---|---|---|
| 1 | 1 | 0 1 3 |
| 3 3 | 2 | 1 5 7 |
| 7 2 1 | 3 | 0 0 1 3 8 9 9 9 9 |
| 9 8 0 | 4 | 2 3 4 4 8 |
| 1 0 | 5 | 0 1 3 7 |

Key
2 | 5 = 25 centimeters
5 | 2 | = 25 centimeters

What is the **difference** between the median growth in centimeters of the selected trees in the tree farm and in the forest?

**Table 9: Number of Mathematics Items per Item Type and Total Test Time by Grade**

| Grade | Multiple-Choice | Gridded-Response | Performance Tasks | Total Minutes per Test |
|---|---|---|---|---|
| 3 | 45–50 | 0 | 0 | 120 |
| 4 | 45–50 | 0 | 0 | 120 |
| 5 | 35–40 | 10–15 | 5–8 | 160 |
| 6 | 35–40 | 10–15 | 0 | 120 |
| 7 | 35–40 | 10–15 | 0 | 120 |
| 8 | 30–35 | 10–15 | 5–8 | 160 |
| 9 | 30–35 | 15–20 | 0 | 120 |
| 10 | 30–35 | 15–20 | 5–8 | 160 |
| Retake | 25–30 | 25–30 | 0 | 160 |

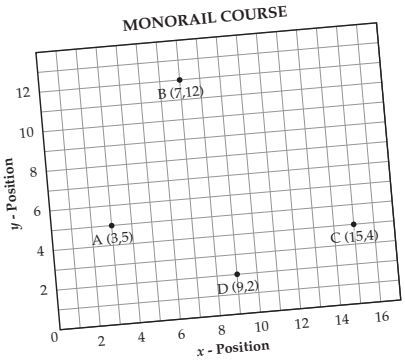
Note: Total testing time is divided into two testing sessions, except for the retake test, which only has one session. Students taking the retake test may receive additional time to complete the test. The data in this table give ranges for the approximate number of items by item type. These ranges include both operational and field-test or anchor items.

**Figure 8: Example of a Grade 10 FCAT Mathematics Short-Response Performance Task with Two Parts**

THINK SOLVE EXPLAIN

The course of the monorail at an amusement park must be changed to make room for a new parking lot. Engineers have decided that only the main supporting column located at point C on the grid below should be relocated. They have also decided that the rebuilt course should be in the shape of a parallelogram.

*Part A* Plot the new location of the supporting column and write its coordinates. Label the new location C′.




*Part B* Use the definition or properties of a parallelogram to verify that the new monorail course is a parallelogram. You must use the slopes of the sides, the lengths of the sides, or both, to help verify your answer.

## Calculators, Reference Sheets, and Rulers

Items for Grades 3–6 are designed to not require calculators, and students in those grades may not use them. In Grades 7–10, four-function calculators are provided to all students for use on all items in all testing sessions. Visually impaired students in these grades are provided with "talking calculators." A reference sheet of appropriate formulas and conversions is provided to students in Grades 6–10 for use during testing. If any formula is needed in Grades 3–5, the appropriate formula is included with the test item. Although rulers may be used on the NRT portion of the FCAT, they are not required and may not be used during FCAT Mathematics.

## 3.7 Science Content

### Knowledge and Skills Tested

FCAT Science measures student achievement of the science benchmarks contained in the *Sunshine State Standards* at Grades 5, 8, and 11. The eight science strands found in the *Standards* are grouped into four reporting clusters: (1) Physical and Chemical Sciences; (2) Earth and Space Sciences; (3) Life and Environmental Sciences; and (4) Scientific Thinking. Items in all clusters may require scientific thinking, although success on the first three clusters depends primarily upon content knowledge. Items classified as Scientific Thinking may be presented in the context of another cluster, but success on these items depends primarily on scientific thinking skills rather than content knowledge. At all three grade levels tested, score points are distributed approximately evenly across the four clusters.

Because of the large number of *Sunshine State Standards* science benchmarks assessed by the FCAT (58 at Grade 5, 70 at Grade 8, and 74 at Grade 11), some benchmarks are assessed annually while the content of others is sampled (assessed) only periodically.

Some of the benchmarks addressed annually in each science cluster for Grades 5, 8, and 11 are described on the next few pages. For more detailed information, refer to the *FCAT Science Test Item Specifications* available at http://www.firn.edu/doe/sas/fcat/fcatis01.htm.

**Mark Tohulka**
(Biology and Life Sciences)
High-school level
Science Teacher, MAST Academy
Miami-Dade County Public Schools
Miami, Florida

**FCAT Committee Experience:** Science Content Advisory; Science Item Review; Science Performance Review

**Related Experience:** Florida Association of Science Teachers (FAST), past President; curriculum writer with NOAA, NASA, and the University of Miami

**"From the very beginning of the FCAT Science test development, I have been impressed by the ability and diligence of the people involved. Every effort is being made to construct an accurate test of a student's scientific literacy, not recall of isolated facts and terms."**

## Science Content Tested

At Grade 5, FCAT Science annually assesses the following skills:

### *Physical and Chemical Sciences*

- understands that matter can be described, classified, and compared
- traces the flow of energy in a system
- identifies the differences between renewable and non-renewable energy sources
- describes, predicts, and measures the types of motion and effects of forces
- identifies the types of force that act upon an object

### *Earth and Space Sciences*

- understands that changes in climate, geological activity, and life forms can be traced and compared
- recognizes that Earth's systems change over time
- identifies the cause of the phases of the Moon and seasons
- recognizes the role of Earth in the vast universe

### *Life and Environmental Sciences*

- understands that living things are different but share similar structures
- recognizes that many characteristics of an organism are inherited
- explains the relationship and interconnectedness of all living things to their environments
- understands that plants use carbon dioxide, minerals, and sunlight to produce food (photosynthesis)

### *Scientific Thinking*

- uses scientific methods and processes to solve problems
- recognizes that most natural events occur in consistent patterns
- understands the interdependence of science, technology, and society

At Grade 8, FCAT Science annually assesses the following skills:

### *Physical and Chemical Sciences*

- recognizes the differences between solids, liquids, and gases
- contrasts physical and chemical changes
- identifies atomic structures
- recognizes properties of waves
- describes how energy flows through a system
- describes, measures, and predicts the types of motion and effects of force

### *Earth and Space Sciences*

- recognizes that forces within and on Earth result in geologic structures, weather, erosion, and ocean currents
- explains the relationship between the Sun, Moon, and Earth
- understands that activities of humans affect ecosystems
- compares and contrasts characteristics of planets, stars, and satellites

### *Life and Environmental Sciences*

- identifies the structure and function of cells
- compares and contrasts structures and functions of living things
- understands the importance of genetic diversity
- recognizes how living things interact with their environments

### *Scientific Thinking*

- uses scientific methods and processes to solve problems
- recognizes that most natural events occur in consistent patterns
- understands the interdependence of science, technology, and society

At Grade 11, FCAT Science annually assesses the following skills:

### *Physical and Chemical Sciences*

- describes and explains the structure of an atom and its interactions with other atoms
- recognizes and explains chemical reactions
- describes how energy flows through a system
- describes, measures, and predicts the types of motion and effects of force

### *Earth and Space Sciences*

- recognizes that forces within and on Earth result in geologic structures, weather, erosion, and ocean currents
- identifies and explains the interconnectedness of Earth's systems
- understands that human activities affect ecosystems
- compares and contrasts characteristics of planets, stars, and satellites

### *Life and Environmental Sciences*

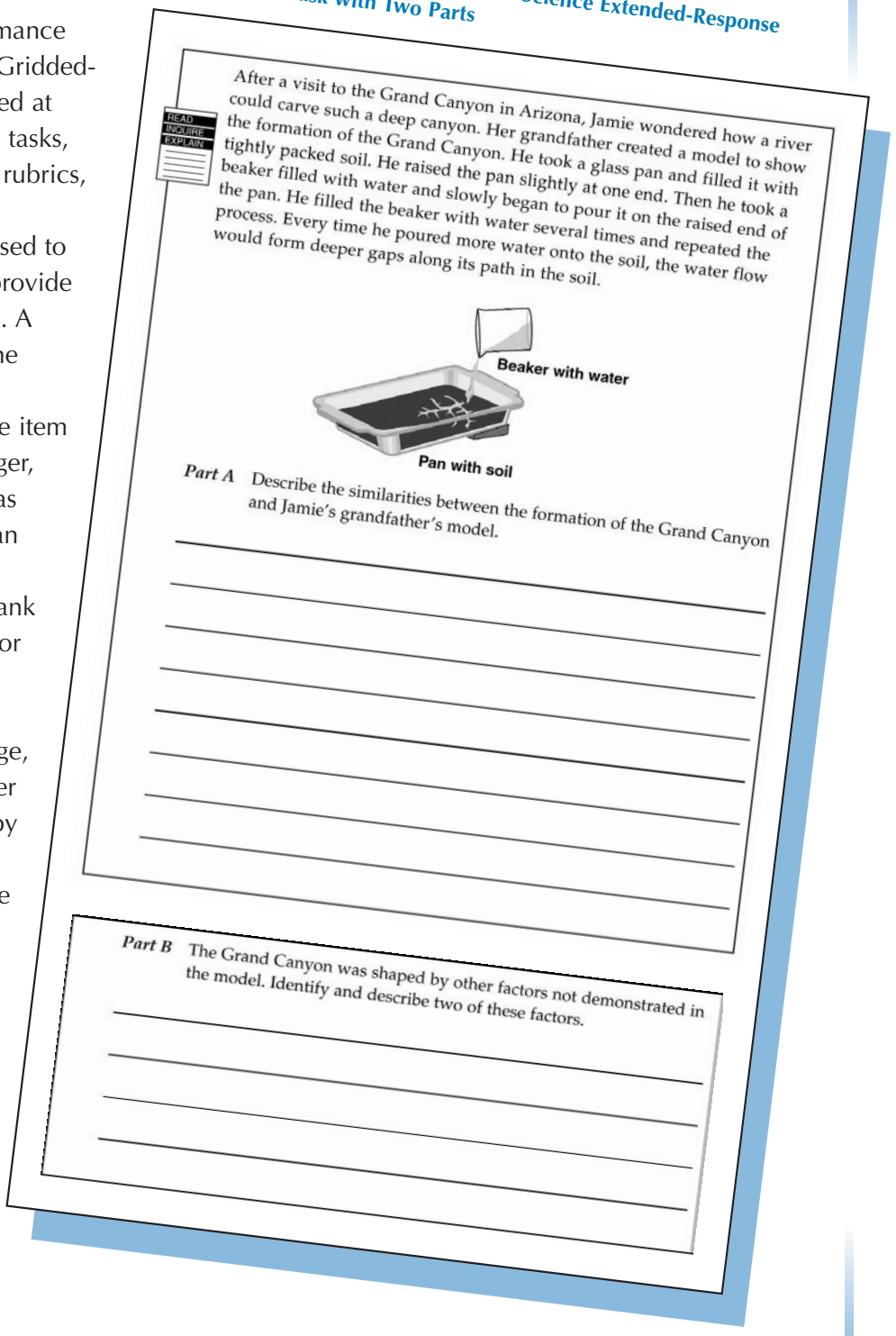
- compares and contrasts the structure and function of major body systems
- recognizes that structures, physiology, and behaviors of living things are adapted to their environments
- identifies and explains the role of DNA
- explains the relationship and interdependence of all living things and their environments

### *Scientific Thinking*

- uses scientific methods and processes to solve problems
- recognizes that most natural events occur in consistent patterns
- understands the interdependence of science, technology, and society

FCAT Science includes multiple-choice and short- and extended-response performance tasks at all three grade levels. Gridded-response items are also included at Grades 8 and 11. Performance tasks, scored with two- or four-point rubrics, require students to explain the scientific concept or process used to determine the answer and to provide the answer in their own words. A short-response item may ask the student to explain a scientific concept. An extended-response item (shown at right) requires a longer, more detailed response, such as describing the steps to use in an experiment. Performance task answer spaces may include blank work space, charts, drawings, or lined answer space, based on what is required to answer the item. Table 10, on the next page, illustrates the range of items per item type as well as test time by grade. One sample multiple-choice item is presented on the next page, and additional sample items are included on the DOE web site (www.firn.edu/doe/sas/fcat/fcatsmpl.htm).

**Figure 9: Example of a Grade 5 FCAT Science Extended-Response Performance Task with Two Parts**



READ
INQUIRE
EXPLAIN

After a visit to the Grand Canyon in Arizona, Jamie wondered how a river could carve such a deep canyon. Her grandfather created a model to show the formation of the Grand Canyon. He took a glass pan and filled it with tightly packed soil. He raised the pan slightly at one end. Then he took a beaker filled with water and slowly began to pour it on the raised end of the pan. He filled the beaker with water several times and repeated the process. Every time he poured more water onto the soil, the water flow would form deeper gaps along its path in the soil.



*Part A* Describe the similarities between the formation of the Grand Canyon and Jamie's grandfather's model.

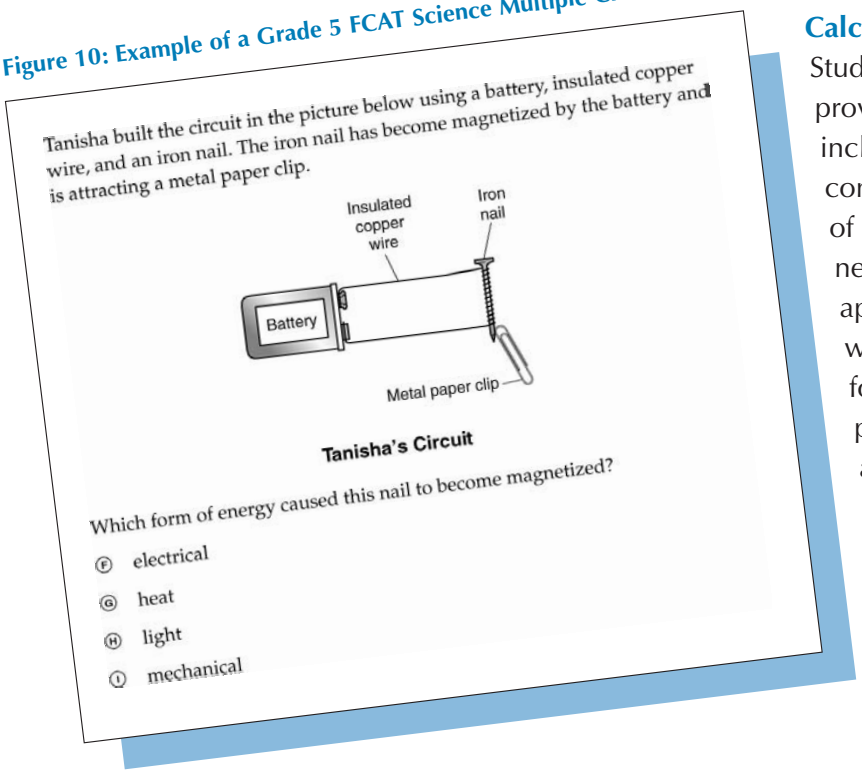
*Part B* The Grand Canyon was shaped by other factors not demonstrated in the model. Identify and describe two of these factors.

**Table 10: Number of Science Items per Item Type and Total Test Time by Grade**

| Grade | Multiple-Choice | Gridded-Response | Performance Tasks | Total Minutes per Test |
|---|---|---|---|---|
| 5 | 45–55 | 0 | 5–7 | 120 |
| 8 | 40–45 | 3–6 | 5–7 | 120 |
| 11 | 40–45 | 3–6 | 5–7 | 150 |

Note: Total testing time is divided into two testing sessions. The data in this table give ranges for the approximate number of items by item type. These ranges include both operational and field-test or anchor items.

**Figure 10: Example of a Grade 5 FCAT Science Multiple-Choice Item**




## Calculators and Reference Sheets

Students in Grades 8 and 11 are provided with reference sheets that include important formulas and conversions and a periodic table of the elements. If any formula is needed in Grade 5, the appropriate formula is included with the test item. Although four-function calculators are provided to students in Grades 8 and 11, use of calculators is not essential because of item design.

## 3.8 Writing Content

### Knowledge and Skills Tested

On FCAT Writing+, students are asked to write an essay within a 45-minute testing session on a single assigned topic. The test is based on the benchmarks describing the writing process in the writing strand of the Language Arts *Sunshine State Standards*. For the purpose of scoring and describing the quality of student essays, four elements of writing inherent in the writing process and benchmarks are considered. These are: (1) focus; (2) organization; (3) support; and (4) conventions.

**Gayle J. Cowley**
(Language Arts, Reading, and Writing)
Coordinator of Language Arts, Reading, and ESOL, Santa Rosa School District
Milton, Florida

**FCAT Committee Experience:** Writing Content Advisory; Prompt Writing; *Lessons Learned* Committee

**Related Experience:** Florida Council of Language Arts Supervisors, President; FDOE Middle Grades Reform Task Force

**"FCAT tests what we *should* be teaching: students *should* be able to read and understand and then explain their thinking in a reasonable format. My involvement in FCAT processes has given me a clear way to distinguish what's essential from what's 'nice to know.'"**

FCAT Writing+ prompts require students to respond with a narrative, expository, or persuasive essay. At Grade 4, prompts are written to elicit either a narrative or an expository response, and at Grades 8 and 10, the prompts are written to elicit either an expository or a persuasive response. A narrative response tells a story, an expository response explains an idea, and a persuasive response attempts to convince an audience to agree with a given position (see Figure 11 below).

More information on FCAT Writing+, including sample prompts and scored responses, may be found in *Florida Writes! Report on the FCAT Writing+ Assessment*, published by the DOE each year for Grades 4, 8, and 10. Additional information about FCAT Writing+ may be found in the FCAT *Sample Test Materials* and the *Keys to FCAT*, available on the DOE web site in PDF format.

The new section of the test includes multiple-choice items with three- and four-answer options. The test includes the following sample types on which items are based: writing samples that model student draft writing (see Figure 12, page 37); stand-alone samples that provide a succinct context for measuring knowledge of conventions (see Figure 13, page 37); cloze samples that contain high-interest material and numbered blanks (see Figure 14, page 38); and writing plans that provide a prewriting structure (see Figure 15, page 38).

**Figure 11: Example of a Grade 4 FCAT Writing+ Expository Writing Prompt**

> Most students like to do something to help the teacher at school.
>
> Think about what you like to do to help the teacher.
>
> Now write to explain what you like to do to help the teacher.

At Grade 4, FCAT Writing+ assesses the following skills:

### *Writing Process*

The student drafts and revises writing (in cursive*) that

- focuses on the topic;
- provides a logical organizational pattern, including a beginning, middle, conclusion, and transitional devices;
- includes ample development of supporting ideas;
- demonstrates a sense of completeness or wholeness;
- demonstrates a command of language, including precision in word choice;
- indicates a general knowledge of the correct use of subject/verb agreement and verb and noun forms;
- includes, with few exceptions, sentences that are complete, excluding purposefully used fragments; and
- uses a variety of sentence structures and demonstrates a knowledge of the basic conventions of punctuation, capitalization, and spelling.

* Note: One Language Arts writing benchmark for Grade 4 states that students should write in cursive. For FCAT Writing+, students may print or write in cursive.

At Grade 8, FCAT Writing+ assesses the following skills:

### *Writing Process*

The student drafts and revises writing that

- focuses on the topic, is purposeful, and reflects insight into the writing situation;
- conveys a sense of completeness and wholeness and adherence to the main idea;
- provides an organizational pattern with a logical progression of ideas;
- includes support that is substantial, specific, relevant, concrete, and/or illustrative;
- demonstrates a commitment to and an involvement with the subject;
- presents ideas with clarity;
- employs creative writing strategies appropriate to the purpose of the paper;
- demonstrates a command of language (word choice) with freshness of expression;
- includes sentences that are complete except when fragments are used purposefully;
- uses a variety of sentence structures; and
- contains few, if any, convention errors in mechanics, usage, punctuation, and spelling.

**Figure 12: Example of a Grade 8 FCAT Writing+ Sample-Based Multiple-Choice Item with Excerpted Writing Sample**

**The article below is a first draft that Antonio wrote for his teacher. The article contains errors. Read the article to answer questions 4–9.**

**The Beginning of Organized Baseball**

[1] The first organized baseball teams and their rules go back to the 1840s. [2] At that time, a New Yorker named Alexander J. Cartwright wrote the first-known written rules of the game.

[3] Baseball had been played for fun in America since the early 1800s.

Which sentence should be deleted because it presents a detail that is unimportant to the article?

**F.** sentence 18
**G.** sentence 19
**H.** sentence 20
**I.** sentence 21

**Figure 13: Example of a Grade 10 FCAT Writing+ Stand-Alone Multiple-Choice Item**

In which sentence below is all **capitalization** correct?

**F.** Kodiak island is off the Coast of Alaska.
**G.** Kodiak Island is off the coast of alaska.
**H.** Kodiak Island is off the coast of Alaska.

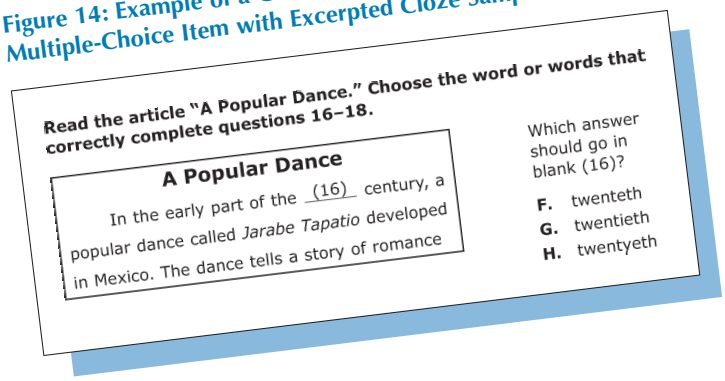
At Grade 10, FCAT Writing+ assesses the following skills:

### *Writing Process*

The student drafts and revises writing that

- focuses on the topic, is purposeful, and reflects insight into the writing situation;
- provides an organizational pattern with a logical progression of ideas;
- includes effective use of transitional devices that contribute to a sense of completeness;
- includes support that is substantial, specific, relevant, and concrete;
- demonstrates a commitment to and an involvement with the subject;
- employs creative writing strategies appropriate to the purpose of the paper;
- demonstrates a mature command of language with freshness of expression;
- uses a variety of sentence structures; and
- contains few, if any, convention errors in mechanics, usage, punctuation, and spelling.

**Figure 14: Example of a Grade 10 FCAT Writing+ Cloze-Based Multiple-Choice Item with Excerpted Cloze Sample**



**Read the article "A Popular Dance." Choose the word or words that correctly complete questions 16–18.**

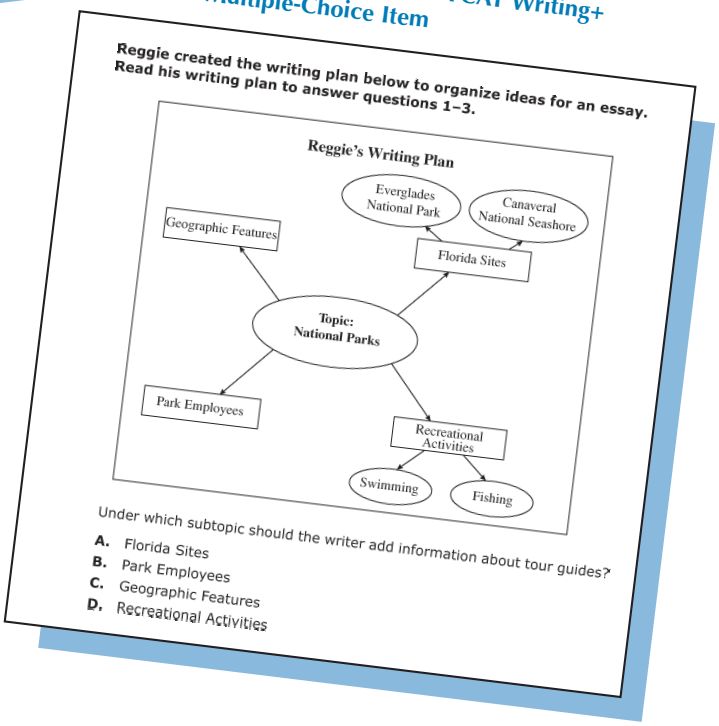
**A Popular Dance**

In the early part of the (16) century, a popular dance called *Jarabe Tapatío* developed in Mexico. The dance tells a story of romance

Which answer should go in blank (16)?

**F.** twenteth
**G.** twentieth
**H.** twentyeth

**Figure 15: Example of a Grade 10 FCAT Writing+ Plan-Based Multiple-Choice Item**



**Reggie created the writing plan below to organize ideas for an essay. Read his writing plan to answer questions 1–3.**

**Reggie's Writing Plan**



Under which subtopic should the writer add information about tour guides?

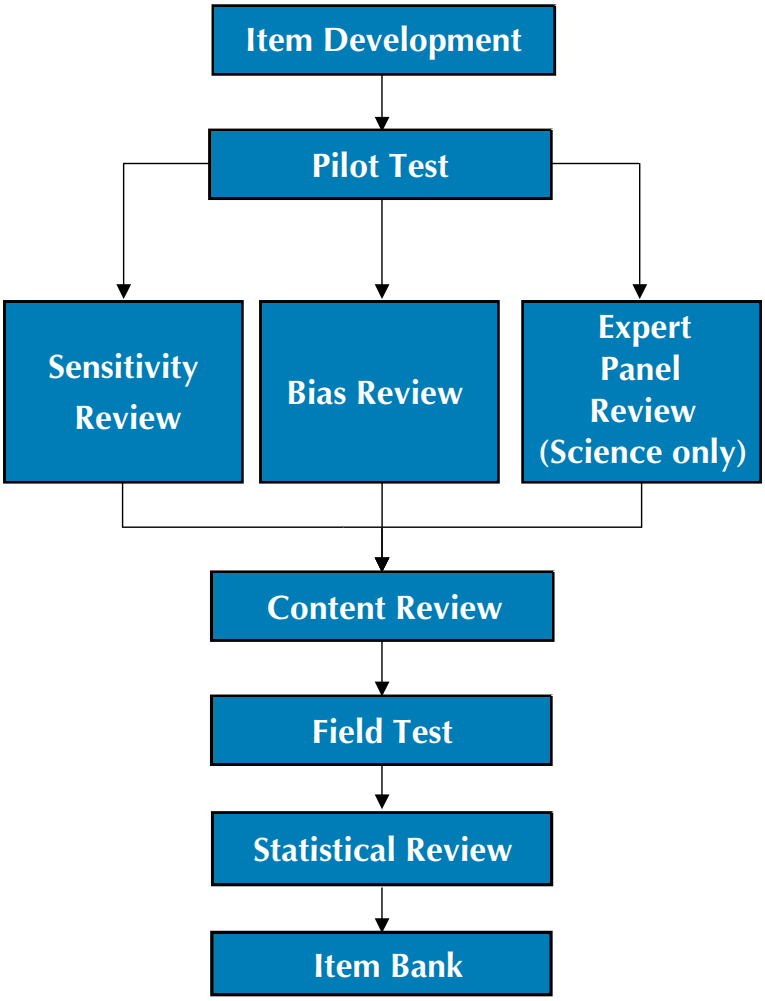
**A.** Florida Sites
**B.** Park Employees
**C.** Geographic Features
**D.** Recreational Activities

# 4.0 Test Development and Construction

Developing an annual statewide assessment to accurately measure achievement and accurately compare results from one year to the next requires an extensive process involving many people with varied expertise. This process is overseen by the Florida Department of Education and annually integrates the work of the DOE's Test Development Center (TDC), outside contractors, and several hundred Florida educators and citizens. Figure 16 briefly illustrates the item development process used for the FCAT. This chapter provides details about each step in this process.

Before reading about the FCAT development and construction processes, you should understand two key concepts. The first relates to field testing items. When an item first appears on the FCAT, it is as a *field-test* item and does not count toward a student's score. After field testing, if the item is statistically sound, then it may be used on the test as an *operational item*, which counts toward a student's score.

**Figure 16: Summary of FCAT Item Development**



The second key concept relates to the nature of the item writing and test construction processes. Item writers do not write a complete test in any given year. Instead, they write individual items that will go through a series of reviews. If items are accepted and have passed through each review successfully, they become part of the *item bank*. The item bank is a database of items serving as the source for constructing the test each year. The process of test construction involves selecting a set of items from the item bank that meets the established content and statistical guidelines of the test. The operational items on the FCAT in any given year will likely have been written in another year and may appear on the FCAT several times before being retired or released as sample items in FCAT interpretive materials for students, teachers, parents, or the general public.

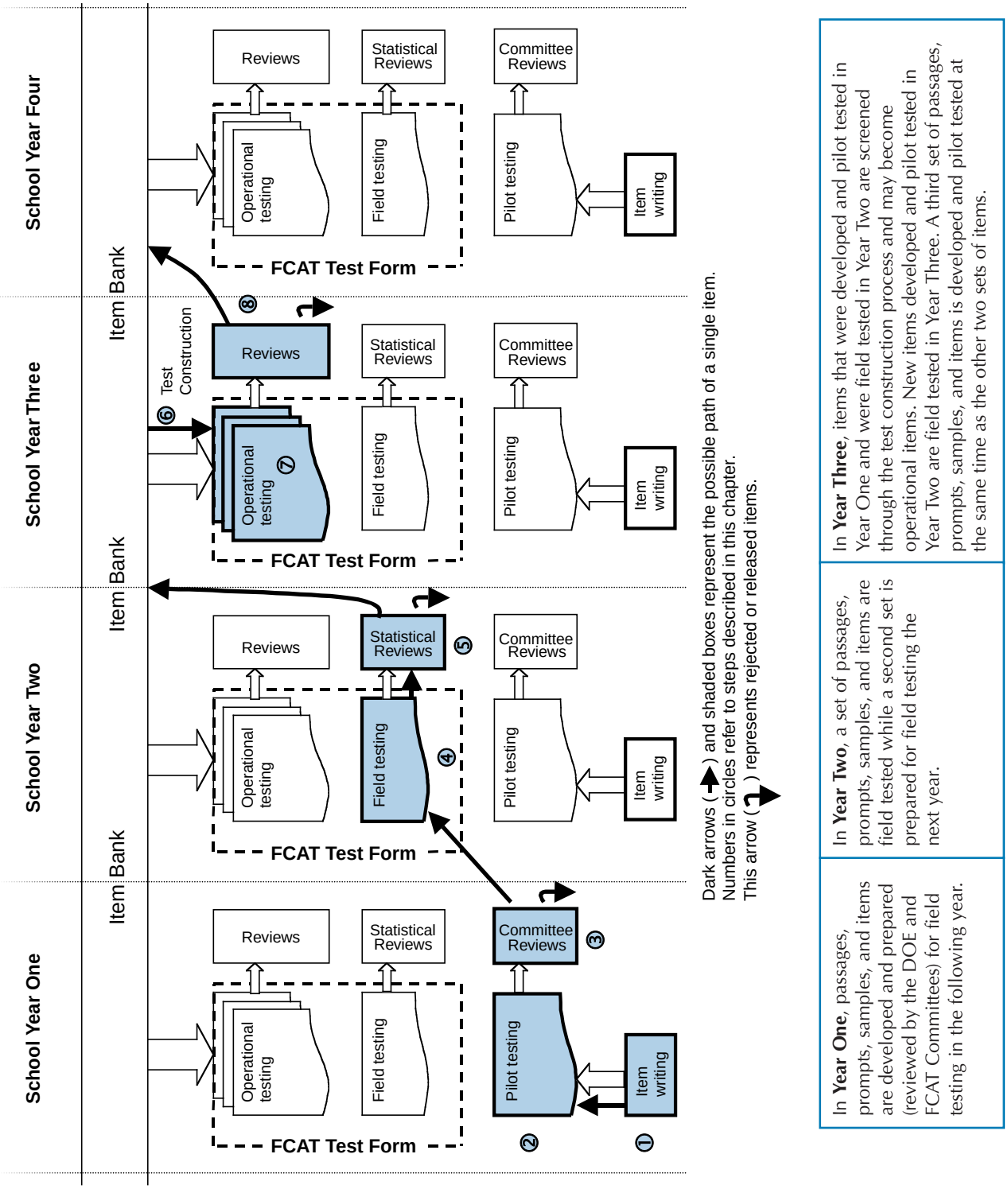
## 4.1 From Benchmark to Test Item: Developing an FCAT Item

There are eight stages in the development of an FCAT item, from item writing through inclusion on the FCAT as an operational item, to concluding with either release to the public or maintenance in the item bank for future use.

1. Item Writing
2. Pilot Testing
3. Committee Reviews
4. Field Testing
5. Statistical Review
6. Test Construction
7. Operational Testing
8. Item Release or Reuse

Each of the numbered stages above corresponds to a stage of item development shown in Figure 17 on the next page and to a section that follows.

## Figure 17: Development of an FCAT Item



### 1. Item Writing

For each subject and grade level, criteria for item development are specified by the DOE in *FCAT Test Item Specifications* (www.firn.edu/doe/sas/fcat/fcatis01.htm). The *Specifications* include the specific *Sunshine State Standards* benchmarks, the types of items used, guidelines for the relative balance of topics, item formats and complexity levels, plus general guidelines to minimize non-content influences, such as confusing wording or poor graphics.

The *Specifications* are developed by the DOE and are based on recommendations of the **Content Advisory Committees** in each of the four FCAT content areas. Each Content Advisory Committee is composed of 15–24 subject area specialists from schools, districts, and universities across Florida. These *Specifications* are revised periodically to provide new sample items, writing samples, and reading passages.

Each year, for all four FCAT subjects, the DOE, Florida educators, and the FCAT contractor agree on a list of benchmarks and item types for which items need to be written. This decision is based on a comparison of the benchmarks in the *Specifications* with items already in the item bank. Then teams of item writers use the *Specifications* to write new items for the designated benchmarks.

Item writers have varied and often specialized backgrounds and abilities, and have teaching experience. Each item writer's résumé is submitted to the DOE for approval. All item writers are required to attend a training session that includes a review of item specifications, cognitive complexity levels, good multiple-choice item characteristics, examples of good performance task items, scoring criteria, and an explanation of bias concerns. Each item writer is given multiple opportunities to draft and evaluate items during training. After training, item writers are assigned to write and submit items for review. Items are reviewed and edited several times before going on to the next stage of development.

### 2. Pilot Testing

After items have been written by the item writers and accepted by the DOE for use on the FCAT, they are compiled into pilot test booklets and administered to small groups of students outside Florida. The pilot tests are not intended for detailed statistical analysis, but rather to gain more general information about students' reactions to test items, clarity of items, and responses to performance tasks. Students are interviewed after the pilot test administration to identify any vocabulary that may be unfamiliar or confusing, graphics that may be unclear, or other concerns.

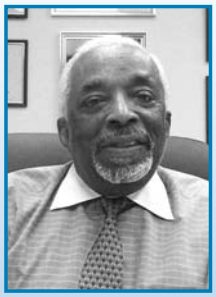
### 3. Committee Reviews

Pilot-tested items must be reviewed by several committees and the DOE before being approved for field testing with Florida students.

Items for all four subject areas are reviewed by **Bias Review Committees**, composed of educators from Florida school districts and universities. In addition to some returning members, new committee members are invited to participate each year on an ad hoc basis. They look for any items, prompts, samples, or passages that might provide an advantage or disadvantage (unrelated to an understanding of the content) to a student with certain personal characteristics, such as those related to gender, race, ethnicity, religion, socioeconomic status, disability, or geographic region.

Similar to the Bias Review Committees, the **Community Sensitivity Committees** are made up of Florida citizens associated with a variety of organizations and institutions. Membership is drawn from statewide religious organizations, parent organizations, community-based organizations, cultural groups, school boards, school district advisory councils, and business and industry from across the state. Reviewers are asked to consider whether the subject matter and language of test items, writing prompts, samples, or reading passages will be acceptable to students, their parents, and other members of Florida communities. Issues of sensitivity are distinct from bias because sensitivity issues do not necessarily affect student success on an item, whereas bias may. Examples of sensitive topics for Florida students may include wildfires, hurricanes, or other topics that may be considered offensive or too sensitive for students or that may distract students from the task at hand. The Community Sensitivity Committees meet once or twice a year.

After each committee meeting, a list of all members' comments is compiled and presented to the DOE for evaluation and inclusion in the materials used during the Item Content Review Committees that follow.



**Donald M. Foster**
C&C International Computers and Consultants, Inc.
Vice President and General Manager
Fort Lauderdale, Florida

**FCAT Committee Experience:** Community Sensitivity Committee; Standard Setting

**Related Experience:** Evaluate requests for and award scholarships to minority students; University of Miami School of Business Administration—Advanced Minority Executive Program; U.S. Commission on Minority Business Development; Girl Scouts of America, Board of Directors

**"I feel that education can be one of the solutions to poverty and that the time I have invested in FCAT committee work is a contribution toward that goal. My involvement in Minority Business issues for more than 20 years has provided me insight to the void that many of our students have in preparation for the business world. This preparation needs to start early in their educational lives as the process is long and arduous."**

**Item Content Review Committee** members are Florida educators, including teachers and administrators from the targeted grade levels and subject areas, and school and district specialists from the content areas. Committee members determine whether the passages, samples, and items are appropriate for the proposed grade levels. Committee members evaluate whether the items measure the benchmarks, evaluate the specified levels of cognitive complexity, are clearly worded, have only one correct answer (for multiple-choice items), and are of appropriate grade-level difficulty. Committee members also recommend approval, modification, or rejection of the passages, writing samples, or items presented by the DOE. There are four Item Content Review Committees, one for each FCAT subject with grade-level subcommittees, which usually meet in the fall. The committee members for all four content areas are invited to participate each year on an ad hoc basis. Another reading committee meets only to review potential reading passages. Additionally, FCAT Science items are reviewed by the *Science Expert Review Committee,* a panel of university-level and practicing research scientists. This review ensures the scientific accuracy of the test items.

Each fall, after the FCAT Writing+ pilot test, the **Prompt Review Committee** reviews the writing prompts and student responses to ensure that the prompts are clearly worded, are of appropriate difficulty and interest level, are unbiased, and will result in a full range of responses. Committee members are Florida educators.

Following committee reviews, the passages and items go through a final review. Approved items are ready to enter the field-testing stage.

### 4. Field Testing

Field-test passages and items are embedded among the operational items in FCAT Reading, FCAT Mathematics, and FCAT Science (and FCAT Writing+ beginning in 2006). On a test with 45–60 items, most test forms will contain six to nine field-test items. Field tests for FCAT Writing+ prompts are conducted on a separate date from operational testing.

Responses to field-test items do not count toward students' scores. Students' responses to these items yield statistics that further reveal the quality of the item. Based on the analyses of field-test data, items are either rejected or placed in the item bank for use as operational items on the FCAT. After being accepted into the item bank, but before being used as operational items, performance task items, writing prompts, and gridded-response items must undergo a further review. For more information about the statistical review, see the next page.

For performance task items and writing prompts, **Rangefinder Committees** examine a representative set of student responses from field tests to establish scoring guidelines. At least 1,000 student responses are reviewed and committee members identify student responses reflective of each specific point on the scoring rubric. The papers scored by the Rangefinder Committees are developed into materials for training teams of professional scorers. There are Rangefinder Committees for each tested subject area. The committees meet after administration of the field tests but prior to scoring of the field-tested performance task items and prompted essays. Members are Florida educators, including teachers from the targeted grade level and subject area, and school, district, and university specialists from the curriculum area. Before these items and prompts are used on a test to contribute to a student's score, the training materials will be reviewed by a Rangefinder Review Committee. See Section 6.2 for more information about this committee.

**Gridded-Response Adjudication Committees** review all responses to field-tested gridded-response items to determine whether all possible correct answers have been included in the scoring key. Based on their input, the DOE establishes rules for how each gridded-response item will be scored. The committees are comprised of Florida educators, including teachers from the targeted grade levels and subject areas and school and district curriculum specialists. The Gridded-Response Adjudication Committees for mathematics and for science meet after each spring administration before field-test gridded-response items are scored.

### 5. Statistical Review

After field-test items have been scored, information about each item is electronically filed in the FCAT item bank. This information includes an image of the item, the item statistics, and details about the item's location in the test book.

The statistical review of these items is conducted as an initial step of test construction. Prior to being selected for inclusion as an operational item on the FCAT, the field-test statistics for the item must satisfy quality criteria. See Section 4.4, Characteristics of FCAT Items, for more detailed information about these criteria.

### 6. Test Construction

Test construction is guided by a set of *Test Construction Specifications*, which are based on the *FCAT Test Item Specifications*, and other considerations such as statistical criteria. Because the *Test Construction Specifications* are used to develop a complete test for a single year, they include more detail about how benchmarks are addressed and about statistical characteristics of items and the final test. The *Test Construction Specifications* are revised annually to guide the construction of the FCAT. Because they contain very detailed information about the content of the FCAT, the *Test Construction Specifications* are protected by test security statutes and are not available to the public.

During the summer months, prior to each test administration, the DOE uses the *Test Construction Specifications* to carefully select items for use on the FCAT in the upcoming school year. A single set of operational items is selected to which either field-test or anchor items are added to create the test forms for each subject and grade. Next, the DOE approves the basic components of the test through a series of reviews resulting in a final version of the FCAT.

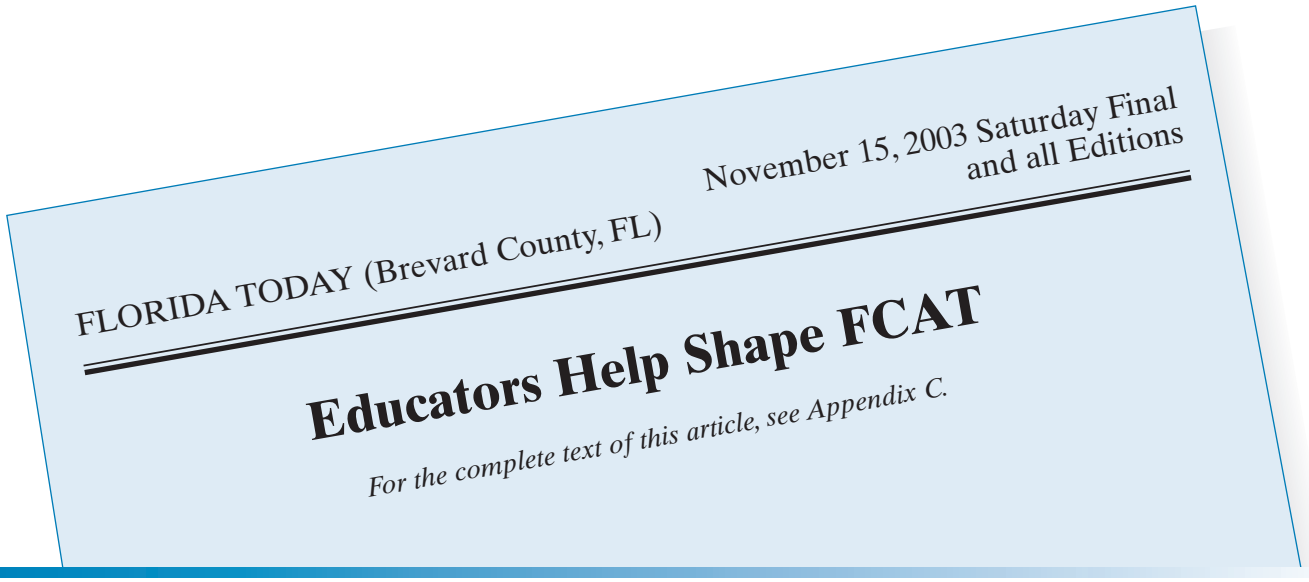
### 7. Operational Testing

Operational testing occurs when the FCAT is administered in all Florida public schools. FCAT Reading, FCAT Mathematics, and FCAT Science are all given in March, and FCAT Writing+ is given in February. Because of the multi-step item development process and the use of the item bank, operational items will have been written and reviewed at least two school years prior to appearing on the test.

During the scoring process, the DOE reviews statistical data from student performance on operational items, using many of the same statistical criteria as were used in the reviews of field-test items. Reviews ensure that both the items and the test as a whole meet established design and psychometric criteria, as the field-test results indicated they would.

### 8. Item Release or Reuse

After the tests are scored and the results are released to students and the public, some items are released in FCAT publications, so they will not appear on the FCAT again. Items not released to the public may be used again. Developing sufficient items to release entire tests to the public is very expensive, costing several million dollars; therefore, items are released using a phased release plan. Phased release means that not all test items are released in all content areas or grade levels at one time. For example, Grade 10 reading and mathematics items may not be released prior to the administration of Grade 10 retakes because it is possible that some test items will be used again on a future retake test form. Anchor and field-test items are not released.


FLORIDA TODAY (Brevard County, FL) November 15, 2003 Saturday Final and all Editions

**Educators Help Shape FCAT**

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

## 4.2 Additional FCAT Committees

The **Assessment and Accountability Advisory Committee** is a standing committee that meets once a year and has 15–20 members representing school district and university personnel. They advise the DOE about K–12 assessment and accountability policies. Their recommendations relate to processes or actions needed with FCAT Achievement Levels, school grading policies, and alternative assessments.

The **Technical Advisory Committee (TAC)** is composed of 10–15 professionals with expertise in psychometrics and/or assessment. The members include Florida District Coordinators of Assessment, representatives from the FCAT Content Advisory Committees, Florida university faculty members, and representatives of universities and state agencies outside Florida. In addition, the psychometric advisors of the DOE's contractors participate in the committee meetings. Committee members assist the DOE by reviewing technical decisions and documents, and by providing advice regarding the approaches the DOE should use to analyze and report FCAT data. This committee meets once or twice a year.

**Laura B. Hassler, Ph.D.**
(Assessment; Data-driven instructional decision-making, reading, leadership and its relationship to student performance)
Educational Leadership and Policy Studies, Associate Professor; Learning Systems Institute, Director, Florida State University
Tallahassee, Florida

**FCAT Committee Experience:** Assessment & Accountability Advisory Committee; Community Sensitivity Committee; *Lessons Learned* Committee; Middle Grades Reform Task Force

**Related Experience:** Former middle school principal, high school assistant principal, and special education teacher

**"As a result of the insights gained in working with other Florida educators in the longitudinal analysis of student performance on FCAT Reading, Mathematics, and Writing (*Lessons Learned*, 2001) and in the review process, I strongly support the notion that FCAT results can provide powerful information for teachers and other school leaders to use in improving teaching and learning."**

The DOE regularly seeks the advice of district educators and business and community representatives to recommend achievement standards for the FCAT. **Standards Setting Committees** were used to recommend the FCAT Reading and FCAT Mathematics Achievement Levels currently in place and will be convened in the future to recommend Achievement Levels for FCAT Science and FCAT Writing+. Committees recommend Achievement Levels (sometimes referred to as performance standards or cut scores) after reviewing items that have different difficulty levels. Committee members evaluate what students must know to answer each item and which scores represent each level of performance or achievement. Selection of committee members is made from those familiar with the FCAT from prior committee participation and people who may be unfamiliar with FCAT but have an interest in the standards being established. Participants include teachers from the targeted grade level and subject area, school and district curriculum specialists, school and district administrators, university faculty from the discipline area, as well as business and community leaders.

The **FCAT Interpretive Products Advisory Committee** is composed of 8–10 professionals representing the many audiences for which FCAT interpretive products are prepared. It meets on an ad hoc basis to review FCAT publications and to provide input to the DOE for future FCAT materials. Interpretive products include publications such as the *FCAT Handbook; FCAT Test Item Specifications;* sample test materials for students and teachers; classroom posters; and reports to educators on the spring assessment *(Florida Writes!, Florida Reads!, Florida Solves!, Florida Inquires!,* and *Understanding FCAT Reports)* among other publications. FCAT interpretive materials are delivered to school districts in print, and many publications are also posted to the DOE web site in PDF format for the general public. Members of the FCAT Interpretive Products Advisory Committee represent Florida school districts as well as the private sector. These individuals are invited to bring experience related to exceptional student education, ESOL, vocational education, post-secondary education, parent involvement, publishing, and community relations.

**Lydia Navarro**
(Curriculum & Instruction; TESOL)
Teacher-on-Assignment,
School District of Volusia County
Deland, Florida

**FCAT Committee Experience:** Bias Review Committee, Sensitivity Committee

**Related Experience:** Florida Spanish Teachers Examination Scorer and Item Writer; FDOE Peer Review Training; TESOL International, Sunshine State, and North Eastern, Member

**"Through the FCAT Bias Review Committee I have gained insight to FDOE staff's effort to ensure FCAT items are free of bias and culturally sensitive to all students. Collaborative team work guarantees FCAT items assess the *Sunshine State Standards*. Constructive feedback from committee members is valued in the decision-making process when constructing future FCAT tests. This review process concurrently has helped me better understand the assessment process and meet our students' needs."**

The DOE also convenes **Special Ad Hoc Committees** on an as-needed basis. Various other groups of parents, teachers, school and district administrators, and others review different aspects of the testing program and advise the DOE on appropriate courses of action. These committees provide advice on issues such as score reporting and norm-referenced testing.

## 4.3 Test Construction

After committee reviews and field testing are completed, the process of selecting items to construct a test begins. The process of design and construction of each FCAT form targets important goals but is also constrained by the realities of cost and time. Since the purpose of the FCAT is to measure student achievement of *Sunshine State Standards* benchmarks, items must have clear connections to those benchmarks. To be of value, FCAT scores must accurately represent students' abilities, requiring not only a large enough sample of student work—in this case, a sufficient number of items—but also items providing specific types of information about student achievement. Constructing a test such as the FCAT requires using the science of psychometrics. For example, statistical analyses are used to verify the quality of the individual items and the validity of the test as a whole. In addition, the need for comparable results from year to year requires that the test design maintains consistent content and difficulty. The test should be appropriate for Florida's diverse student population and acceptable to all communities in Florida, while still providing an accurate assessment of the standards.

In order for the FCAT to serve its various functions within the limitations placed upon it, very clear criteria and quality control measures are established for designing both FCAT items and the test itself. The criteria and the quality control measures are partially based on the recommendations of the Technical Advisory Committee.

The next sections present descriptions of the desired characteristics of FCAT items and the entire test, as well as the measures taken to ensure them. Each section provides a general description of related characteristics, processes, and quality control measures. More detailed information on the statistical indicators and processes can be found in Appendix A.

## 4.4 Characteristics of FCAT Items

This section explains the various analyses performed on field-tested items in order to decide whether they will be used on the FCAT. The statistical analyses described in this section are performed both after the field test and again after each operational test to verify that the items performed as expected. Quality assurance methods used for these characteristics are summarized in Table 11 on page 57. Definitions for the terms referenced in Table 11 and throughout this section can be found at the end of the document in the Glossary and Appendix A.

### Content Validity – Connection to a Benchmark

All test items must address a specific *Sunshine State Standards* benchmark. Items are reviewed and evaluated for how well they address the benchmarks for which they were developed.

> ***Quality Assurance Measures***—Ensuring that items are written to specific benchmarks is the responsibility of item writers, Item Content Review Committees, and the DOE. In fact, content validity is not quantifiable by the statistical analyses routinely performed in FCAT data analysis; however, item writers are given clear instructions about writing items to assess specific benchmarks, and they are reviewed for direct connections to benchmarks at several points in the development process.

### Difficulty Level

Items that are very easy or very hard may provide useful information for some, but not all, students. For the majority of test takers, test items of moderate difficulty provide the most information. A moderately difficult item is not so easy that virtually all students answer it correctly, nor so difficult that virtually all students answer it incorrectly. These types of items provide the most useful information on student achievement at the aggregate school, district, or state levels.

> ***Quality Assurance Measures***—After items have been written, but before they have been field-tested, they are reviewed for grade-level difficulty and appropriateness by the DOE and the Item Content Review or Prompt Review Committees.
>
> After field testing, statistical analyses of student performance are used to verify that items are within an acceptable range of difficulty. One indicator of difficulty for all item

types is the *p*-value, an item's difficulty index expressed as the proportion of students who responded correctly (successfully) to an item. The *b*-parameter of the Item Characteristic Curve, the function used in Item Response Theory (IRT), is another indicator of item difficulty. If an item falls outside the range of acceptable values, it may be rejected from further use. (See more about IRT on pages 56 and 59–62.)

### Item Discrimination (Item-Test Correlation)

For an item to be useful on a test, there must be a positive correlation between students' success on an item and their success on the test as a whole. In other words, students who succeed on a given item should exhibit greater success on the test as a whole than students who do not succeed on that item. Similarly, students with relatively higher achievement on the test as a whole should exhibit greater success on any given item than students with relatively lower achievement. This relationship may seem obvious, since the test score is based on the scores of individual items; however, among items there will be variation in the strength of the relationship, with some items exhibiting only a minimal correlation. In rare cases, there may even be a negative correlation, meaning that students who succeed on an item exhibit lower levels of overall achievement on the test. Items with minimal or negative correlations with overall test success may be poorly worded, may have two correct answers, may not actually test what they are intended to test, or may assess something that is unrelated to what the other items test.

***Quality Assurance Measures***—Using detailed item development guidelines and field testing is intended to reduce the number of items with low or negative item-test correlations. These guidelines and the multi-step process of item development usually result in well-written items that assess what they are intended to assess and that are aligned with the overall content of the test. As verification, however, the item-total correlations are generated and reviewed after both field testing and operational testing. Appendix A describes the statistical indices used to analyze test data.

### Guessing

On a multiple-choice item with four choices, the likelihood of choosing the correct answer simply by guessing is about 25 percent. If the *distractors* (the incorrect alternative choices) are ineffective, and most students are able to easily eliminate one or more of them and then select their answer from the remaining choices by guessing, the likelihood of guessing the correct answer increases. Instead of a four-choice item, the item essentially becomes a three- or two-choice item. To minimize guessing on a multiple-choice item, item writers and reviewers are instructed to design items with plausible distractors, but only one correct answer.

***Quality Assurance Measures***—After field testing, test developers examine data for each item, including the percent of students choosing each possible response and the *c*-parameter of the Item Characteristic Curve, the function used in Item Response Theory (IRT). Items with unusually high guessing indices or high *c*-parameters are rejected. See more about IRT on pages 56 and 59–62.

## Freedom from Bias

An item is considered biased if it places a group or groups of students at a relative advantage or disadvantage due to characteristics, experiences, interests, or opportunities common to the group, that are unrelated to academic achievement.

***Quality Assurance Measures***—Instructions to item writers and reviewers call attention to the possibility of bias and include a checklist to ensure that items are free from bias. In the pilot test phase, test takers are interviewed about their reactions to items, providing test developers with reasons why a given item might be unexpectedly difficult or easy for a given group of students.

Two additional measures identify and eliminate potential bias. First, items are reviewed by the Bias Review Committees who note any potential bias and give their comments to item reviewers. In some cases, items are eliminated from further consideration at this point.

In addition to the thorough reviews by the Bias and Sensitivity Review Committees, gender and ethnic bias can also be identified in the statistical analysis of field and operational test data using a statistical technique called *differential item functioning* (DIF). Items with DIF exhibit differences in scores between males and females or between ethnic groups that are unique to the item and cannot be explained by differences between these groups in overall achievement. DIF statistics not only allow the DOE to identify potentially biased items, but also to understand the likely impact of the bias on student performance. Field-tested items can be rejected for future use as operational items based on these analyses.

**Egle Rodriguez**
(English for Speakers of Other Languages [ESOL]; Homeless and Migrant Education), Federal Programs Specialist, School District of Osceola County
Kissimmee, Florida

**FCAT Committee Experience:** Bias Review Committee

**Related Experience:** Teachers of English Speakers of Other Languages; Florida Association of State and Federal Educational Program Administrators

**"Having reviewed other state assessment tests, I can say that the FDOE has the most comprehensive and impeccable process for reviewing all content areas of the FCAT to ensure that ALL students in Florida have a fair and equal chance of demonstrating their knowledge and academic achievement."**

## Universal Design Principles

Applying universal design principles to the development of test questions results in assessments that are usable by the greatest number of students, including those with disabilities and non-native speakers of English. To support the goal of providing access to all students, the test maximizes readability, legibility, and compatibility with accommodations.

***Quality Assurance Measures***—The DOE trains both internal and external reviewers to write or revise items in such a way as to allow for the widest possible range of student participation. Item writers attend to the best practices suggested by universal design, including, but not limited to, reduction of wordiness; avoidance of ambiguity; selection of reader-friendly constructions and terminology; and application of consistently applied concept names and graphic conventions. Universal design principles are also used to make decisions about test layout and design, including, but not limited to, type size, line length, spacing, and graphics. The DOE and the test contractors use the *Test Production Specifications* to ensure that FCAT test documents meet established high-quality standards. The *Test Production Specifications* are not released to the public.

## Item Fit to the IRT Model

Data analyses conducted after field testing and after operational testing include *Item Response Theory* (IRT) analysis for each item. There are three parameters for each test item produced by the IRT analysis: the degree to which the item differentiates between students of different abilities (the *a*-parameter), the difficulty of the item (the *b*-parameter), and the likelihood of success by guessing (the *c*-parameter). These parameters are used to ensure that each item (and the test as a whole) fits established guidelines. They are also used to determine an overall test score for each student. For these item parameters to be useful and for student scores to accurately reflect knowledge of the content, each item's IRT function should fit the observed pattern of student responses.

***Quality Assurance Measures***—For each item, a statistic describing the quality of fit to the model is generated. This statistic is derived by estimating expected student performance on the item, and then comparing this estimate to actual student performance on the item. For FCAT data, there are established standards for fit values that indicate a good fit of the model. These standards are established in the *Test Construction Specifications*. More information can be found in the *FCAT Technical Report* on the DOE web site at: www.firn.edu/doe/sas/fcat/fcatpub2.htm.

**TABLE 11: CHARACTERISTICS OF FCAT ITEMS**

| Characteristic | Quality Assurance Methods |
|---|---|
| Content Validity | Item Content Review Committees<br>Percent choosing each answer choice<br>*Test Item Specifications* |
| Difficulty Level | Item Content Review Committees<br>Prompt Review Committee<br>Field test and operational test data analysis—*p*-values; IRT *b*-parameters |
| Item Discrimination (Item-Test Correlations) | *Test Construction Specifications*<br>Field test and operational test data analysis—Item-total correlations; IRT *a*-parameters |
| Guessing | *Test Construction Specifications*<br>Field test and operational test data analysis—IRT *c*-parameters |
| Freedom from Bias | *Test Construction Specifications*<br>Bias Review Committees<br>Pilot Test Results<br>Field test and operational test data analysis—Differential Item Functioning (DIF) analysis (Mantel-Haenszel statistic; Mantel statistic; SMD rating) |
| Adherence to Universal Design Principles | *Test Item Specifications* and *Test Production Specifications* |
| Item Fit to the IRT Model | *Test Construction Specifications*<br>Field test and operational test data analysis $Q_1$ ($Z_{Q1}$) |

## 4.5 Characteristics of the Test

This section describes the desired characteristics of the FCAT forms prepared annually, as shown in Table 12 (page 59). Each characteristic is followed by an explanation of the related quality assurance method.

### Content Coverage (Content Validity)

The FCAT measures student success on a specified set of *Sunshine State Standards* benchmarks with a balance of emphasis among them. It is important that the FCAT include items that collectively reflect the desired range of those benchmarks. Results from a test that does not sufficiently sample the set of benchmarks or the content domain will not provide an accurate measure of achievement in that subject area.

> ***Quality Assurance Measures***—Each year, test developers use the guidelines in the *Test Construction Specifications* to develop the FCAT. This document specifies the number of items on the FCAT to address each benchmark and the percentage distribution of items across content strands or clusters. The *Test Construction Specifications* help the DOE's test developers ensure that the FCAT reflects the range and balance of content specified in the set of benchmarks used to define the subject area.

## Test Difficulty

When all the items on a test are of the same level of difficulty, results tend to identify two groups of students: those who can correctly answer questions at the given difficulty level and those who cannot. It is more desirable that the items on a test address a range of knowledge of the content being assessed. When items represent a range of difficulty levels, it is much easier to identify students achieving at relatively higher levels (those who are able to correctly answer the most difficult items) and at relatively lower levels (those who are unable to correctly answer the easiest items). Generally speaking, a range of item difficulties allows creation of a scale of student achievement with useful information on students at all levels of achievement.

> ***Quality Assurance Measures***—Assuring the necessary range of item difficulties occurs mainly during test construction. In addition to selecting items for content coverage, test developers select items based on difficulty-related data gathered either from field tests or from operational use in previous years. The two indicators of item difficulty used in test construction (the items' *p*-values and IRT *b*-parameters) are the same as those used in item-level analysis. During test construction, test developers review both the *p*-values and *b*-parameters for all items to ensure distribution of item difficulties across all levels of achievement.

## Test Reliability

FCAT scores are estimates of students' levels of achievement. A reliable score provides an accurate estimate of a student's true achievement. As with any estimate, there is some error. On a reliable test, the amount of error will be small. When there are sufficient numbers of test items that reflect the intended content, are free from bias, are well-written, represent a range of difficulty, and have positive correlations to success on the test, the likelihood of the test being reliable will be high and the amount of error will be low.

***Quality Assurance Measures***—Virtually all of the steps in the test development process contribute in some way or another to minimize error and maximize the reliability of the FCAT. In the process of test construction, test developers review the statistical data for items and generate three indicators of overall test reliability: *standard error of measurement* (SEM), *marginal reliability*, and *Cronbach's alpha*. These statistics and measures are reviewed in light of established guidelines before final approval. SEM, test information curves, marginal reliability, Cronbach's alpha, and classification accuracy and consistency are all reviewed at test construction and after test administration.

### Test Fit to the IRT Model

The IRT model used in FCAT development and scoring is based on the idea that the content assessed has a single dimension. This *unidimensionality* represents consistency in the content assessed. A test that lacks unidimensionality may produce estimates of a student's achievement that are not as reliable as a test that assesses only a single dimension.

***Quality Assurance Measures***—Studies of the unidimensionality of the FCAT, conducted prior to the first operational test administration for each subject area, have confirmed that each test, as developed, fits the IRT model.

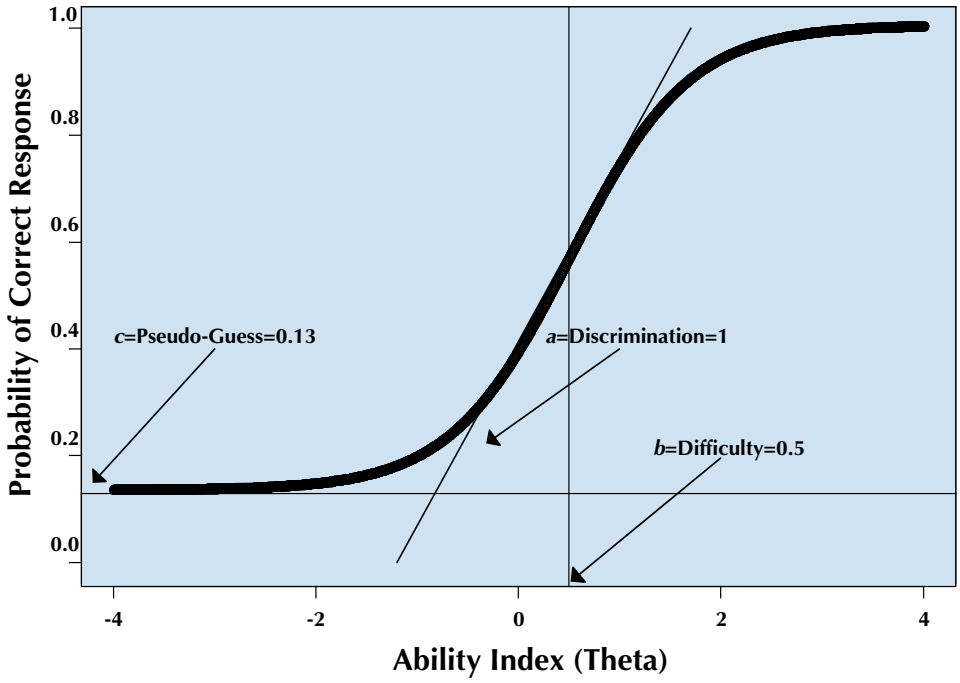
**TABLE 12: CHARACTERISTICS OF THE TEST**

| Characteristic | Quality Assurance Methods |
|---|---|
| Content Coverage (validity) | *Test Item Specifications* and *Test Construction Specifications*; test reviews |
| Test Difficulty (validity and reliability) | *Test Construction Specifications*—*p*-values; IRT *b*-parameters; test characteristic curves |
| Test Reliability | Field test and operational test data analysis—standard error of measurement (SEM); marginal reliability index; Cronbach's alpha; test SEM curves |
| Test Fit to the IRT Model | Test construct (e.g., mathematics) is found to be unidimensional. |

## IRT Framework

The purpose of this section is to provide a broad summary of the statistical model used to score the FCAT. Readers interested in more detailed information should consult the cited references as well as Appendix A. FCAT scoring is built on Item Response Theory (IRT). Essentially, IRT assumes that test-item responses by students are the result of underlying levels of knowledge and skills, known as ability, possessed by those students. Items that fit the IRT model will have lower probabilities of correct responses from low-achieving students and higher probabilities of correct responses from high-achieving students. This is reflected in the item characteristic curve, an example of which is depicted in Figure 18, for a single multiple-choice item.

### Figure 18: Item Characteristic Curve Example




*a* = a function of the slope at point of inflection of the item characteristic curve
*b* = theta value at point of inflection of the item characteristic curve
*c* = lowest probability value of item characteristic curve

In IRT analysis, a computer program creates a function for each item so that the resulting item characteristic curve most closely resembles the actual pattern of student responses. In this function, students' probability of success on an item corresponds to true levels of ability. The function incorporates three characteristics of the item: the *a*-, *b*-, and *c*-parameters. The *a*-parameter reflects the item's ability to distinguish between students above and below a given level; the *b*-parameter represents the relative difficulty of the item; and the *c*-parameter reflects the likelihood of low-achieving students guessing the correct answer of a multiple-choice item. During test construction, item parameters are carefully reviewed to determine if an item is suitable to become an operational item. The parameters are recalculated after operational use and then used to generate student scores.

- **The *a*-parameter reflects the item's ability to distinguish between students above and below a given level;**
- **the *b*-parameter represents the relative difficulty of the item; and**
- **the *c*-parameter reflects the likelihood of low-achieving students guessing the correct answer for a multiple-choice item.**

Items differ in their difficulty such that the position of the point of inflection of this curve (the vertical line on Figure 18, on the previous page) is higher or lower (to the right or to the left) along the theta (ability) scale. For example, the point of inflection of the item characteristic curve shown in Figure 18 is centered at one-half a standard deviation above the zero point. An efficient test is composed of items with characteristic curves similar to this example, but with varying difficulties (points of inflection) that are positioned along the entire theta, or ability, scale. The three-parameter logistic (3PL) model (Lord & Novick, 1968)[8] is used to analyze multiple-choice items, and the two-parameter partial credit (2PPC) model (Muraki, 1992)[9] is used to analyze performance tasks. Figure 18 depicts an item characteristic curve using the 3PL model.

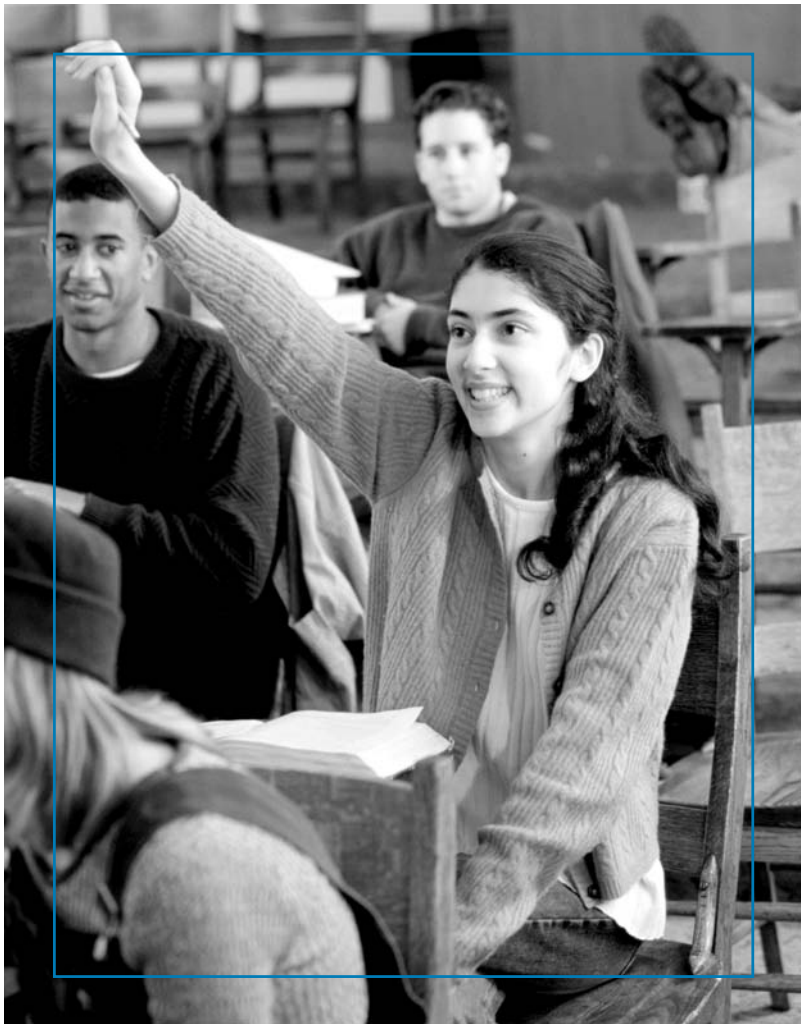
While IRT modeling of performance tasks is conceptually similar to that of multiple-choice items, performance tasks require a more complex mathematical treatment. In the end, however, modeling of a performance task includes the IRT parameters for each of the possible score points students can achieve on that performance task.

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

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

Gridded-response items receive a hybrid treatment. Initially, item parameters are computed using a two-parameter logistic (2PL) model, and then converted to the 2PPC for subsequent processing.

IRT item parameters for all items on a test provide the means for determining scores of individual students. Because the item parameters represent response probabilities, each student's achievement is assigned as the score most likely to correspond to that student's responses.[10] Using the sophisticated IRT model is advantageous for large-scale testing programs, such as the FCAT, because it helps create a stable scoring system when items included on the tests change from one year to the next.



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

# 5.0 ADMINISTERING THE FCAT

After the test has been designed, items have been field-tested and approved, and test forms have been printed, the next step is to administer the test to students. More than four million test booklets are distributed to more than 3,000 schools, whose staff must then give the test to the students and return those booklets for scoring—all within a three-week window (a one-week window for FCAT Writing+). Not only must this process be completed in a timely manner, it must also be conducted in such a way to ensure comparable testing conditions in every school. In addition, secure handling of all test documents must be maintained at all times. It is only through a standardized and secure administration process that the FCAT can provide an accurate representation of student achievement. It is this standardization that makes comparisons across schools and years possible. Because Florida educators have first-hand experience with test administration, their input and feedback is sought throughout this process so that test administration can be improved each year.

In the fall, the DOE convenes the **Annual Statewide Meeting for Florida District Coordinators of Assessment** to provide them with information about assessment and accountability issues for the upcoming school year. Agenda topics usually include information about test development, administration, scoring, and interpretive materials. In addition to the annual meeting, the DOE often solicits feedback from this group about specific issues via formal and informal surveys, focus groups, or ad hoc meetings.

After each spring's test administration, another meeting, the **Annual Debrief on the Assessment**, is held with a representative group of District Coordinators of Assessment to discuss issues related to the recent test administration. During this meeting, these district representatives provide the DOE and the FCAT contractors with information about aspects of the test administration that went well and areas that need improvement. The DOE is able to consider and act on this input in planning the next FCAT administration.

The DOE also solicits information about the test administration from educators in the classroom and at the school-level via comment forms. Each test administrator is given the opportunity to provide input on specific survey questions as well as communicate his or her own opinions about the test administration process. The DOE compiles and reviews all of these surveys and comments in time to plan for the next test administration. In the past, the DOE has acted on these suggestions and comments, e.g., the length of individual test sessions has been changed based on the comments received from those who administered the test at the classroom level. While this is not a formal FCAT committee, the input from Florida educators at this level is a critical part of the process.

***Quality Assurance Measures***—Detailed information relating to test administration is provided in the *FCAT Test Administration Manuals*. The manuals provide all the administration requirements for teachers who administer the test, School Coordinators who organize the administration in their schools, and District Coordinators of Assessment who coordinate the FCAT program for their districts.

## 5.1 Administration Process and Personnel

The DOE prints, ships, and retrieves FCAT materials with the assistance of a contractor. The contractor prints, distributes, and assists with scoring the FCAT test materials. FCAT test materials include the test books and answer documents, forms, training materials, comment sheets for the various district and school personnel, and preprinted labels with student names and other information to be affixed on individual answer documents. After testing, the contractor serves as the point of return for all materials.

Test materials follow a three-level chain of distribution (district, school, and testing session) as described below.

***District level***—The district designates one of its employees as the *District Coordinator of Assessment* to act as the point of contact between the DOE, the contractor, and the schools.

***School level***—The school designates an employee, typically a school administrator or guidance counselor, as the *School Coordinator of Assessment* to act as the point of contact between the district and the school.

***Testing session***—*Test Administrators* supervise testing sessions. Test Administrators must be employees of the school district and are usually classroom teachers. They must remain in the testing room at all times. Test Administrators may be assisted by *proctors.* Proctors are recommended at all times, but are required when the number of students in the testing room exceeds 29. School personnel may be involved in the handling of secure documents; non-school personnel or students may not serve in this capacity.

**Owen A. Roberts, Ph.D.**
(Educational Measurement; Science and Mathematics Specialist)
Executive Director of Accountability and Assessment
School Board of St. Lucie County
Fort Pierce, Florida

**FCAT Committee Experience:** Technical Advisory Committee; Bias Review Committee; Science Content Advisory; Science Performance Review; FCAT RFP Committee (DOE); Mathematics Content Advisory; Mathematics Item Review

**Related Experience:** Florida Educational Research Council (FERC), President; Florida Educational Research Association (FERA), Executive Board Member; JASON PROJECT teacher argonaut, NSTA science standards development team, NSTA Outstanding Teacher in Secondary Science award, participant in scientific expedition with Dr. Robert Ballard.

**"My involvement with the FCAT since 1998 has convinced me of the thorough and comprehensive approach taken in its design and development. I have served on most sub-committees, ranging from the item review to the technical review committee, and have found the FCAT to be a valid test based upon high standards."**

The District and School Coordinators of Assessment are responsible for receiving and verifying materials and sending them to the next person in the distribution chain. District Coordinators provide training to School Coordinators regarding administration procedures and are available during testing to answer questions. School Coordinators have similar responsibilities in relation to Test Administrators and proctors. All District and School Coordinators as well as Test Administrators receive a copy of the *FCAT Test Administration Manual* prior to each test administration.

***Quality Assurance Measures***—Each year following the spring administration, a debriefing meeting is held with DOE staff and a representative group of District Coordinators. This meeting provides the districts an opportunity to give the DOE feedback related to the administration of the test—what worked well and what

did not. The DOE works closely with District Coordinators to ensure that each FCAT administration runs as smoothly as possible.

## 5.2 Students Tested

In general, all students enrolled in the tested grade levels (3–11) should participate and must take the test appropriate for the grade level in which they are enrolled. *Limited English Proficient (LEP)* students are required to participate unless the student has received services in an LEP program operated in accordance with an approved district LEP Plan for one year or less, AND a majority of the student's LEP committee determines that an exemption is appropriate. An Exceptional Student Education (ESE) student may be exempted from the FCAT if he or she has a current *Individual Educational Plan* (IEP) and meets the following criteria according to Rule 6A-6.0331, Florida's Administrative Code (FAC):

- the student's cognitive ability prevents the student from completing required coursework and achieving the *Sunshine State Standards*, even with appropriate and allowable course modifications; and
- the student requires extensive direct instruction to accomplish the application and transfer of skills and competencies needed for domestic community living, leisure, and vocational activities.

If an ESE student is exempted, the IEP must document why the assessment is not appropriate and what alternative assessment will be used.

Some students outside the public school system and outside Grades 3–11 may also take the FCAT, such as students seeking a high school diploma, who have not yet passed the Grade 10 FCAT Reading and/or Grade 10 FCAT Mathematics; students enrolled in an adult high school credit program; and home-educated students. Private school students receiving an Opportunity Scholarship must take the test and McKay Scholarship recipients may choose to take the test.

## 5.3 Testing Conditions and Special Accommodations

Generally, ESE or LEP students who receive special accommodations in their classroom instruction are entitled to similar accommodations on the FCAT as long as the validity or reliability of the test is not compromised. Accommodations enable all students to demonstrate their level of achievement without altering the knowledge or skill being tested by providing students a test format or situation that addresses the nature of their disability.

The IEP or 504 (Section 504 of the Rehabilitation Act of 1973) plan indicates which accommodations students should receive on the FCAT, and the school reports to the district the names of students who received such accommodations. Accommodations fall into five categories: presentation, response, scheduling, setting, and assistive devices. On the next page are examples of some of the accommodations that might be granted in these categories. For more information, refer to the *FCAT Test Administration Manual* distributed to districts and schools prior to each test administration.

- ***Presentation:*** Students may be administered sessions of the test through the use of large print or braille versions of the test; devices to magnify the test; or signed or oral presentation of the test directions, writing prompts, and mathematics items (but not reading passages or reading test questions).
- ***Response:*** All responses must be in English. Students may respond to test questions orally, by signing, by typing, by using a machine to write in braille, or by writing in the test book or on separate paper.
- ***Scheduling:*** Students may be allowed flexible scheduling of their testing through the division of normal testing sessions into two or more smaller sessions with breaks in between, and through extended time for any session on the test.
- ***Setting:*** Students may be administered the test individually or in small groups with a Test Administrator or proctor, or in a specially designed classroom to accommodate special lighting or equipment needs with a test administrator or proctor present. LEP students may be offered the opportunity to be tested in a separate room with the ESOL or heritage language teacher acting as test administrator.
- ***Assistive Devices:*** Students may use assistive devices that are typically used in classroom instruction (such as auditory amplification devices) and technology for writing assessments or extended-response items without accessing spelling or grammar-checking applications.

***Quality Assurance Measures***—The *FCAT Test Administration Manuals* (Chapter 9.0) provide very specific guidelines related to the assessment of special populations. School-level staff may refer to the information in the manual in order to provide the appropriate and allowable testing accommodations to these students.

## 5.4 Security Measures

The Florida Test Security Statute (Section 1008.24), enacted by the Florida State Board of Education, established security guidelines that must be followed and prohibits activities that could threaten the integrity of the test. Examples of prohibited activities include:

- giving examinees access to test questions prior to testing;
- copying, reproducing, or using in any manner inconsistent with test security rules all or any portion of any secure test book;
- coaching examinees during testing or altering or interfering with examinees' responses in any way;
- making answer keys available to students;
- failing to follow security rules for distribution and return of secure test materials as directed, or failing to account for all secure test materials before, during, and after testing;

- failing to follow test administration directions specified in the *Test Administration Manual;* or
- participating in, directing, aiding, counseling, or encouraging any of the acts prohibited in this list.

The requirements also necessitate that all materials be kept in secure, locked storage prior to and after administration of any test and between testing sessions. They also prohibit anyone, including District or School Coordinators, from opening test books before the designated testing times. To enforce this rule, test books are sealed, and the seals must only be broken by the students at the beginning of the test session as directed in the scripts for administering the test. FCAT security measures also prohibit anyone from unsealing and reviewing unused test books after testing is completed. After students complete all testing sessions in a subject and return their testing materials to the Test Administrators, their test books are not opened until they reach the scoring site.

The DOE encourages districts to ask any person who handles test materials (including District and School Coordinators of Assessment, Test Administrators, proctors, and test assistants) to sign a security agreement (included as Appendix B) stating that he or she was made aware of these regulations and procedures and that he or she agrees to follow them.

The State Board of Education Rule also requires that each secure test book or answer document has a unique security number. These numbers appear on the front or back of test books and answer documents. Each time test books and answer documents pass between people in the chain of administration (e.g., from the test administrator to the students), the numbers must be checked to make sure that all material is appropriately accounted for. Any missing documents or other potential breaches of security are reported to the DOE via the School and District Coordinators of Assessment. If a School Coordinator cannot find the documents or if a security breach is suspected, the District Coordinator is notified immediately. The District Coordinator notifies the Office of Assessment and School Performance at the DOE.

The Board of Education Rule also authorizes officials from the DOE to conduct unannounced observations of any test administration site to ensure the testing procedures are being correctly followed. The statute requires local districts to cooperate in the investigation of a security breach or testing irregularity.

> ***Quality Assurance Measures***—The DOE routinely conducts analyses of FCAT data prior to their release in order to ensure that the results accurately reflect student performance. When anomalies are identified in the data, districts are required to conduct security investigations in order to determine the validity of the scores. In addition, a missing materials report is produced after each administration that identifies any secure test document that was not returned by the district. Districts are then required to conduct investigations to locate any missing materials and report their findings to the DOE.

# 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.

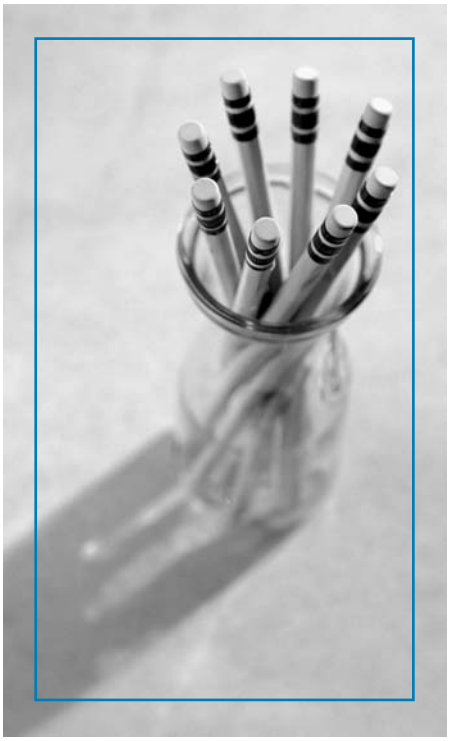
The anchor papers and item-specific criteria are developed initially by Florida educators serving on Rangefinder Committees (see page 46 and Appendix D for more information) and then reviewed and refined by other Florida educators on **Rangefinder Review Committees.** After performance task items are 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.

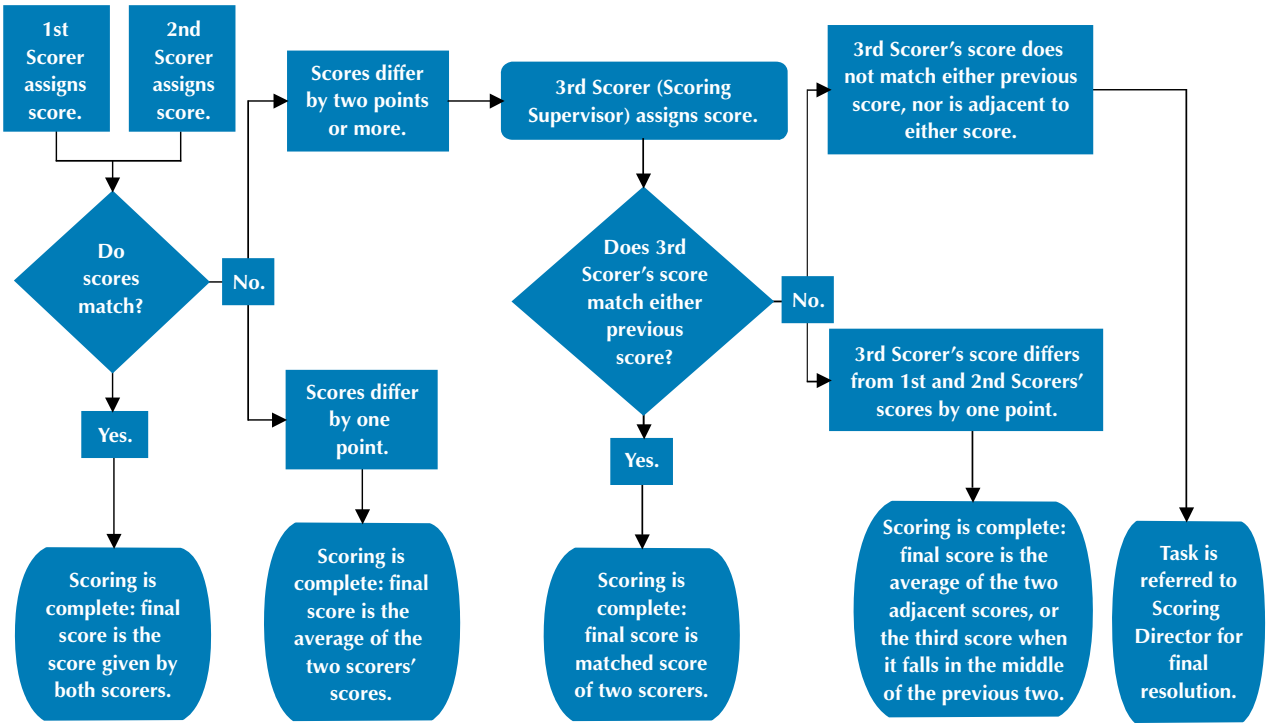
Those selected as professional scorers work in teams of 10–15 members with each team having a Scoring Supervisor. Each team specializes in a set of two to three performance task items, known as *rater item blocks* (RIBs) (for reading, mathematics, or science), or in a single writing prompt. A Scoring 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.

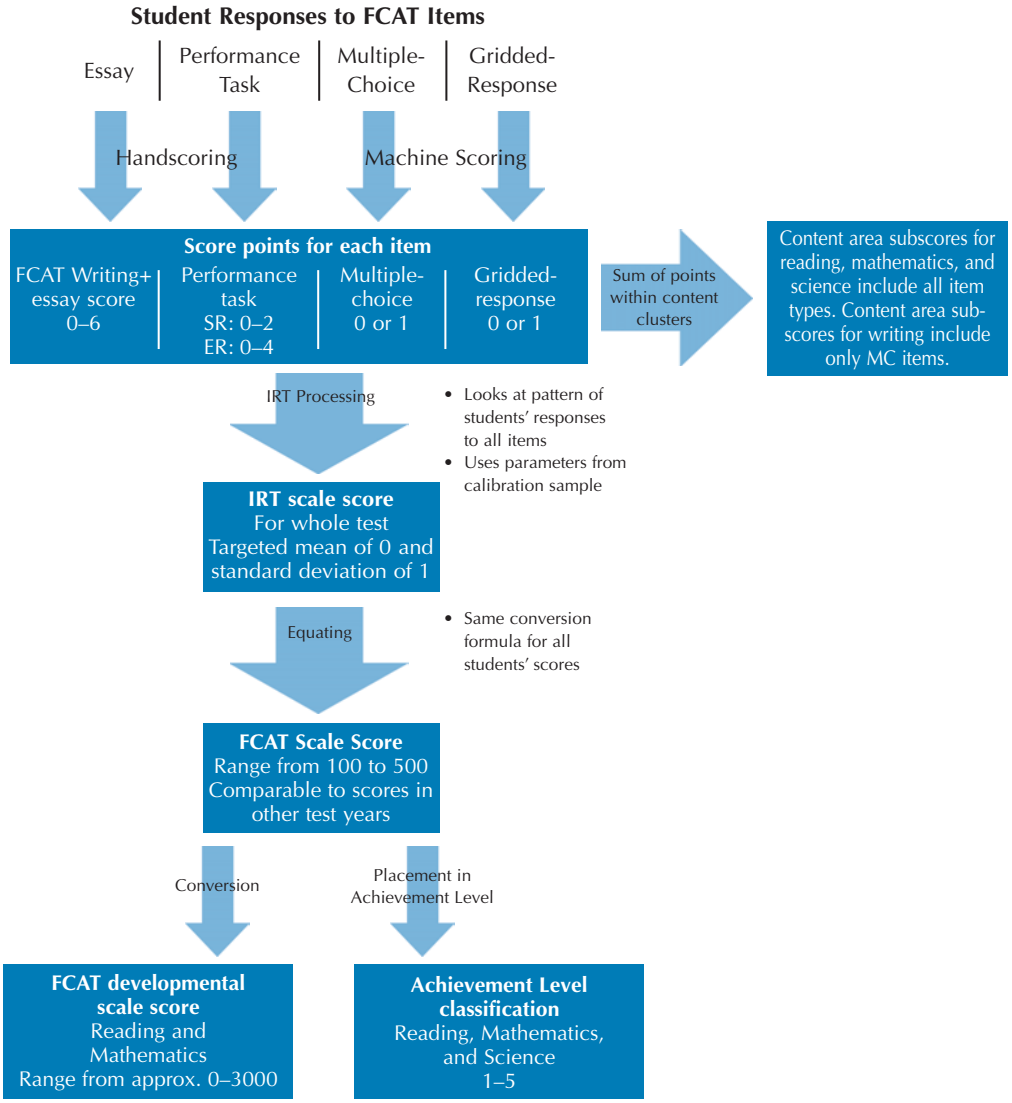
- ***Daily Review of Training Materials***—At the beginning of each scoring session, team members spend at least 15 minutes reviewing their training materials and scoring guidelines, including anchor papers and item-specific criteria.

- ***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.

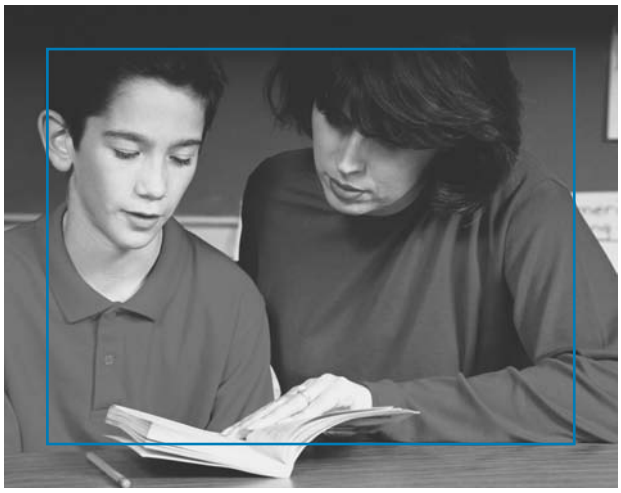
Another key feature of the IRT model is that ability and item difficulty exist on a single dimension so that students with low scores[12] will generally succeed on less difficult items, students with moderate scores will typically succeed on items with low to moderate difficulty, and students with high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

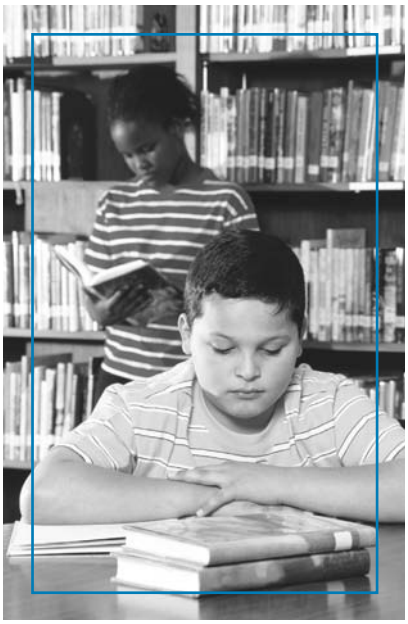
[13] Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Measurement*, 7, 201–210.

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<br>IRT *b*-parameters, Test Characteristic Curves (TCC) |
| Item-Test & Item-Strand Correlations | Item-total correlations, biserial correlations,<br>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<br>(Mantel-Haenszel statistic, Mantel statistic, SMD rating) |
| Reliability | Test information curves, SEM curves<br>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,<br>and by-cut-point accuracy and consistency |

# 7.0 REPORTING FCAT RESULTS

Each spring, reports containing FCAT results are sent to four major audiences: students and their parents, school administrators, district administrators, and state-level administrators and policy makers. The DOE also makes results available to the general public on the FCAT web site. Educators seeking a thorough understanding of FCAT reports should review the publication *Understanding FCAT Reports*.[14] This document is issued each May and can also be found on the DOE web site (http://fcat.fldoe.org/).

Depending on the FCAT subject, reports include the following scores:

- ***FCAT Score*** (reading and mathematics)—This is the developmental scale score (ranging from 0 to about 3000 points). For science, it is the 100–500 scale score. For writing, it is currently the rubric-based score (ranging from unscorable to 6) assigned by the scorers. In 2006, FCAT Writing+ scores will be on the 100–500 scale. (The methods for awarding all of these scores are described in Chapter 6.0, Scoring the Test.)
- ***Achievement Level*** (reading and mathematics)—This is the Achievement Level (from 1 to 5) into which the student's FCAT Score falls. Achievement Levels will be reported for science beginning in 2006, and for writing beginning in 2007. (Achievement Levels are also described in more detail in the previous chapter.)
- ***Content scores*** (reading, mathematics, science, and for writing beginning in 2006)—For the content clusters (subcategories) in each subject, content scores are reported as the actual number of points earned out of the number of points possible.
- ***Scores from previous year(s)*** (reading and mathematics)—FCAT Scores from previous year(s) are presented alongside the current year's score as an FCAT Score history.

[14] The DOE also provides student-, school-, district-, and state-level NRT reading and mathematics reports. While these reports are not addressed in this *Handbook,* information and samples of the reports are available in *Understanding FCAT Reports.*

- ***Performance task scores*** (reading, mathematics, and science)—Performance task scores include points earned on a selected (released) task, the total points possible for all tasks, and other information about the selected (released) task.

In addition to scores, reports contain the following information:

- A range of scores within which the student's "true score" is likely to fall if the student were to take the same test again, and again, and again. This is derived from the student's scale score and the estimated SEM at that score level. This range describes the scale score using the following statement (or one similar): "This score shows your achievement the day you were tested. If you were to take this test again, it is likely that your FCAT Score would be between _ _ _ and _ _ _."
- A written message about the student's performance. Grade 10 students also receive a message that indicates whether they have met the passing scores in reading and mathematics for high school graduation.

## 7.1 Promotion and Graduation Requirements

Reporting results directly to students and parents is critical in helping them understand if students have met state requirements for promotion or graduation. While statewide promotion and graduation requirements are explained here, districts and individual schools may also have separate promotion or graduation requirements that must be met. Anyone not familiar with local requirements should check with district or school administrators for more information.

FCAT scores are reported in five Achievement Levels. If a student's Achievement Level improves from one year to the next, that student has clearly made progress. A student also makes progress by scoring in the same Achievement Level for two years in a row. This is because the content assessed at the higher grade is more difficult. Students who score at Levels 3, 4, or 5 are performing at or above expectations and meet the requirements of the *Sunshine State Standards.* Students who score in Level 1 and Level 2 are performing below expectations and need additional instruction in the content assessed at his or her grade. The schools and districts have guidelines for making decisions about promoting students who score at the lowest level (Level 1). For more information about FCAT Achievement Levels, go to the DOE web site: http://www.firn.edu/doe/sas/fcat/fcatachv.htm.

Florida statutes specifically mention retention for students in Grade 3 who have not demonstrated sufficient reading skills, that is, those students who have scored in Level 1. These students must be provided additional instruction before being promoted to Grade 4. Each school board has some latitude in how it implements this requirement; however, school personnel must develop an individual academic improvement plan for each student who scores in Level 1. The district Pupil Progression Plan (available at local schools and school district offices) specifically outlines the grade-level promotional requirements.

Passing scores on the Grade 10 FCAT Reading and Mathematics are required for high school graduation. In 2001, the State Board of Education adopted administrative rule 6A-1.09422 that specified passing scores on the Reading and Mathematics *Sunshine State Standards* portion of the Grade 10 FCAT. The Board acted on recommendations from the Commissioner of Education that were based on input from the education community as well as from groups of interested citizens. As a result, students who expect to graduate from high school must earn passing scores on the FCAT Reading and FCAT Mathematics at Grade 10.

As determined by the State Board of Education, the current[15] Grade 10 passing scores are as follows:

- FCAT *SSS* Reading Test 1926 (scale score of 300) or above
- FCAT *SSS* Mathematics Test 1889 (scale score of 300) or above

Performance on the FCAT is not the sole criteria in determining eligibility for graduation. Florida Statute 1003.43 is very specific in that no student can receive a standard high school diploma from a Florida public school unless that student has met all academic requirements. Students must take required courses, earn the requisite number of credits, maintain a grade point average of 2.0, and pass Grade 10 FCAT Reading and FCAT Mathematics before graduating.

If students do not earn passing scores on the FCAT the first time they take the test, they have additional opportunities. The Grade 10 FCAT is administered in fall, spring, and summer to 11th and 12th graders who have not yet passed the tests. Most students in Grades 10 through 12 have six opportunities to pass the FCAT. For more information on passing scores, see the DOE web site: http://www.firn.edu/doe/sas/fcat/fcatpass.htm.

## 7.2 Reports for Students and Parents

The number and type of reports for students and their parents vary depending on the grade level and the subjects tested. The major reports include:

- *Sunshine State Standards Reading and Mathematics Student and Parent Report*—Provides results for reading and mathematics, a letter to parents and guardians, and general information about the FCAT program. Reading and mathematics results are provided separately within the same report. Tables and charts display student achievement level results, content scores, and an FCAT Score history. Information is presented in English, Spanish, and Haitian Creole.
- *Sunshine State Standards Science Student and Parent Report*—Provides results for science by content score and FCAT Score. Information is presented in English, Spanish, and Haitian Creole.

[15] There are alternate passing scores for students in certain categories: seniors (Grade 12 or Grade 13) who took the Grade 10 FCAT in March 2003; students who took the Grade 10 FCAT for the first time in 2001; students who took the Grade 10 FCAT for the first time prior to 2001; and students who were in the ninth grade in school year 1999–2000.

- *Writing Student Report*—Includes the rubric-based total score and descriptions of the scoring rubric, scoring process, and the topic the student was given.
- *Reading Sunshine State Standards Performance Task Student Report*—Includes the combined score points for all performance tasks, one selected short-response task, a scanned image of the student's actual response to the task, and the total points earned for that response.
- *Mathematics Sunshine State Standards Performance Task Student Report*—Identical in format to the performance task report for reading.
- *Science Sunshine State Standards Performance Task Student Report*—Identical in format to the performance task report for reading.
- *Sunshine State Standards Reading and Mathematics Retake Tests Student Report*—Provides results for students who took the FCAT Reading and/or the FCAT Mathematics retake test(s).
- *Norm-Referenced Test Student Report*—Provides results for FCAT Reading NRT and FCAT Mathematics NRT on the same report.

**TABLE 15: REPORTS SENT TO STUDENTS AND PARENTS BY GRADE**

| Grade | Reading and Mathematics Report | Science Report | Writing+ Report | Performance Task Reports | | |
|---|---|---|---|---|---|---|
| | | | | Reading | Mathematics | Science |
| 3 | ✓ | | | | | |
| 4 | ✓ | | ✓ | ✓ | | |
| 5 | ✓ | ✓ | | | ✓ | ✓ |
| 6 | ✓ | | | | | |
| 7 | ✓ | | | | | |
| 8 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 9 | ✓ | | | | | |
| 10 | ✓ | | ✓ | ✓ | ✓ | |
| 11 | | ✓ | | | | ✓ |
| Retakes | ✓ | | | | | |

Refer to *Understanding FCAT Reports* for samples of reports, including the *Sunshine State Standards Reading and Mathematics Student and Parent Report*.

## 7.3 Reports for School, District, and State Administrators

Reports for school, district, and state administrators include some of the same types of results as the reports for students and parents, but they also include summary data for all students in a school, district, state, grade level, or demographic group. School principals receive school-level reports for their own school, and district-level reports for their district. District superintendents and district assessment offices receive school-level reports for all schools in their district, district-level reports for their own district, and state-level reports. The DOE receives copies of the summary reports for all schools and all districts in the state, in addition to state-level reports.

The following reports are generated separately for FCAT Reading, FCAT Mathematics, and FCAT Science:

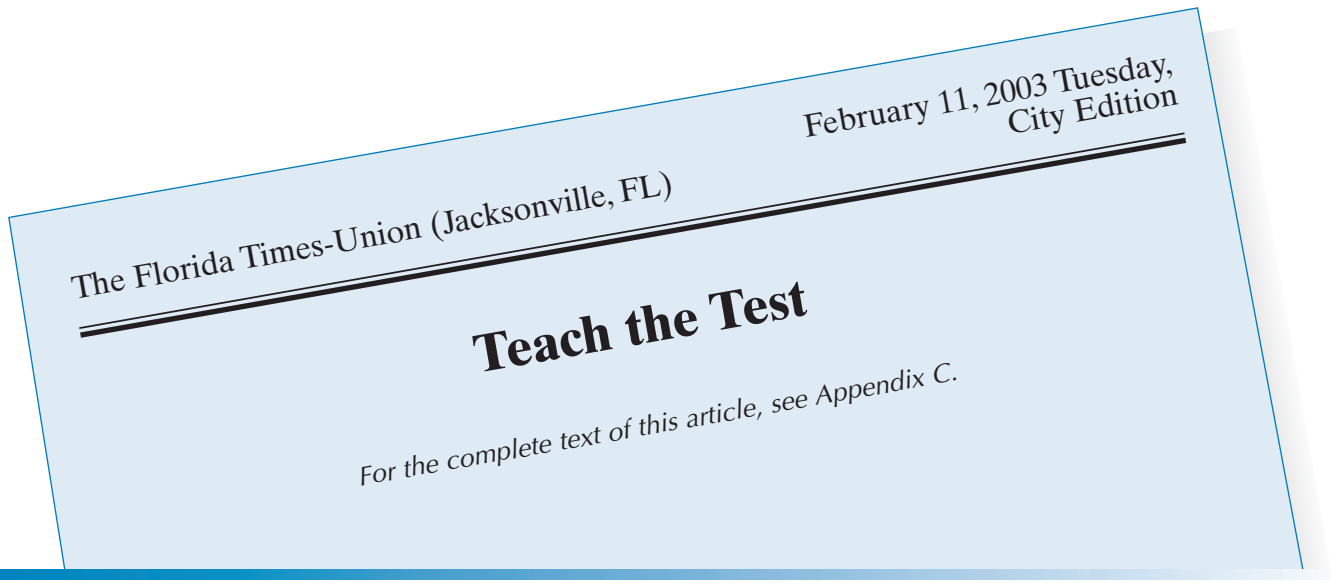
- *School Reports of Students*—student-level data for all students in a school
- *District Reports of Schools*—school-level data for each school in the district
- *District Reports of Scores*—district-level data for all grades tested
- *State Reports of Districts*—district-level data for all districts in the state
- *State Reports of Scores*—state-level data for all grades tested

For FCAT Writing, reports for school-, district-, and state-level administrators include mean scores and the percentages of students receiving different score points on the rubric scale. Data are presented separately for each type of writing (i.e., expository, persuasive, or narrative) and for all types of writing combined. FCAT Writing reports include:

- *Writing School Listing of Achievement*—student-level data consisting of an alphabetical list of students tested and the scores they received
- *FCAT Writing School Results*—school-level data for a single school
- *FCAT Writing District Results*—district-level data for a single district
- *FCAT Writing State Results*—state-level data
- *District Report of School Means and Score Point Distributions*—school-level data for all schools in a district
- *State Report of District Means and Score Point Distributions*—district-level data for all districts in the state

## Demographic Reports

Reports with demographic data are the same as the school-, district-, and state-level reports for all four subjects tested. (FCAT Reading and FCAT Mathematics are presented on the same report.) In these reports, data are disaggregated for racial and ethnic categories, gender, and other special categories, including standard curriculum, limited English proficient (LEP), migrant, Section 504, free or reduced lunch, not free or reduced lunch, exceptional student education classifications (ESE), total ESE other than gifted, not ESE plus gifted, and students not matched to the enrollment file.


The Florida Times-Union (Jacksonville, FL)

February 11, 2003 Tuesday, City Edition

**Teach the Test**

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

# 8.0 GLOSSARY

Terms in **boldface type** appear within the glossary as a separate entry.

**Achievement Levels**—Five categories of achievement that represent the success students demonstrate with the ***Sunshine State Standards*** content assessed on the FCAT. Achievement Levels are established using the input of classroom teachers, curriculum specialists, education administrators, and other interested citizens. These professionals helped the DOE identify the score ranges for each Achievement Level. The Achievement Levels are helpful in interpreting what a student's **scale score** represents.

**Anchor Items**—A common set of **items** on tests administered in two different years used to develop comparable **scale scores**. The student performance on the anchor items is the source for the data used in the statistical **equating** procedures.

**Anchor Papers**—Student responses that demonstrate typical performance for each score point in the **rubric**. As they score each student's response to an **item**, scorers compare the response to the anchor papers to determine the number of points that student has earned. Also called **rangefinders**.

**Backreading**—Method used to ensure adherence to scoring guidelines for **performance task items** and essays. Scoring officials evaluate appropriateness of scores assigned by scorers.

**Benchmark**—A statement within the ***Sunshine State Standards*** that describes what students at a certain grade level should know and be able to do. More detailed than a **strand** or **standard**.

**Bias**—Advantage or disadvantage conferred upon groups of students because of certain personal characteristics (such as gender, race, ethnicity, religion, socioeconomic status, disability, or geographic region) unrelated to an understanding of the content.

**Calibration Sample**—Carefully selected group of students representative of all students statewide whose response data are used to generate **Item Response Theory (IRT)** parameters used in **operational** test scoring.

**Census Test**—The assessment of all eligible students at a particular grade level in a specific **content area**. This term is used to distinguish an assessment of all students from an assessment of only a sample of students.

**Cloze**—Text with blanks inserted where a word or words need to be added. After reading the cloze sample, students choose the answer that correctly completes the sentence. FCAT Writing+ cloze samples contain high-interest material in a relatively short format that can be more literary or technical in nature than the text in the other sample types. Cloze samples are not presented as representative of student-generated work. On a test form, each cloze sample contains three to four numbered blanks used to measure the student's knowledge of spelling or usage conventions.

**Cluster (content cluster)**—A grouping of related **benchmarks** from the ***Sunshine State Standards***. Clusters are currently used to summarize and report achievement for FCAT Reading and FCAT Science and, beginning in 2006, will be used to summarize and report achievement for FCAT Writing+.

**Cognitive Complexity**—System used to classify FCAT **items** according to the complexity of the steps and processes they require students to use.

**Content Area**—The information or skills contained in an area of study. The content areas (or subject areas) assessed on the FCAT are reading, mathematics, science, and writing.

**Content Area Subscores (content cluster** scores, **content area** scores)—The number of points earned by a student in each **cluster** or **strand** of the ***Sunshine State Standards*** portion of the FCAT. Content subscores are reported for **clusters** in FCAT Reading and FCAT Science. In FCAT Mathematics, content subscores are reported for **strands**. Beginning in 2006, FCAT Writing+ content subscores will be reported for **clusters** and for the essay. Computed before **IRT** processing and **equating**.

**Content-Sampled Benchmarks**—**Benchmarks** assessed periodically (as opposed to annually) by FCAT Science.

**Criterion-Referenced Test (CRT)**—An assessment where an individual's performance is compared to a specific learning objective or performance standard and not to the performance of other students. Criterion-referenced tests show how well students performed on specific goals or **standards** rather than just telling how their performance compares to a norm group of students nationally or locally. The FCAT, a CRT, is based on the ***Sunshine State Standards*** and measures student progress toward meeting these **standards**.

**Cut Point Scores**—FCAT **scale scores** or FCAT **developmental scores** that mark the boundaries between different **Achievement Levels**.

**Developmental Scale Score (DSS)**—A type of **scale score** used to determine a student's annual progress from grade to grade. Calculated by converting a student's **scale score** (100–500) to a scale from 0 to about 3000 that is used for Grades 3–11.

**Equating**—A process 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. Students are tested in two different years with tests that have a common set of **items** called **anchor items** as well as different **items**. The **anchor items** and how students perform on them from year to year are used in the statistical equating procedures. Equating scores ensures that the same standard of achievement is used each year so the progress of students and schools can be evaluated fairly, i.e., Grade 8 scores in 2004 are comparable to Grade 8 scores in 1998.

**Exceptional Student Education (ESE)**—Special educational services that are provided to eligible students, e.g., visually impaired or hearing impaired. These services are required by federal law and are provided to Florida students according to the State Board of Education Rule 6A-6.0331, FAC. Also known as Students With Disabilities (SWD).

**Expository Writing**—Writing that gives information, explains why or how, clarifies a process, or defines a concept. In FCAT Writing+, students in Grades 4, 8, and 10 are assigned **prompts** intended to elicit expository writing.

**Extended-Response Item (ER)**—See **Performance Tasks**.

**FCAT Score**—For FCAT Reading and FCAT Mathematics, the **Developmental Scale Score**. For FCAT Science, the **scale score**.

**Field-Test Item**—**Item** included on the FCAT for **item** development purposes only. Student response data are reviewed to determine whether a field-test **item** would be a useful **operational item**. Does *not* count toward student scores.

**Gridded-Response Item (GR)**—Test **items** that require students to solve a problem for which the answer is numerical. Answers must be written and bubbled into a number grid. The gridded-response **item** format is used in FCAT Mathematics (Grades 5–10) and FCAT Science (Grades 8 and 11).

**Holistic Scoring**—A method of scoring written work that considers the overall quality of the entire work. Scores are assigned to student work using a pre-defined **rubric**.

**Individual Education Plan (IEP)**—Describes special education services provided as part of **Exceptional Student Education**. Also specifies the testing accommodations a student needs for classroom instruction and assessments.

**Item**—Any test question or task for which a separate score is awarded.

**Item Bank**—Database of **field-test** and **operational items**. **Items** are selected from it each year to construct the FCAT.

**Item Response Theory (IRT)**—Statistical model for student responses to test **items**. Based on the idea that the likelihood of student success on an **item** is the result of the student's true level of ability and three characteristics of the **item**: ability of the **item** to differentiate between students at different **Achievement Levels** (the *a*-parameter), difficulty of the **item** (the *b*-parameter), and the effectiveness of guessing (the *c*-parameter, for **multiple-choice items** only). Used solely in FCAT **item** and test development and as the basis of generating **scale scores**.

**Limited English Proficient (LEP)**—Special education services for non-native speakers of English. LEP students, also known as English Language Learners (ELL), are permitted testing accommodations when taking the FCAT.

**Linking**—Method used to create **developmental scale score**. A small sample of identical **items** are given to students in adjacent grades.

**Mode of Writing**—Characteristics of written work that reveal the purpose of the writing. The essay portion of FCAT Writing+ assesses three modes of writing: **narrative**, **expository**, and **persuasive**.

**Multiple-Choice Items (MC)**—**Items** that present students with several options from which to choose. FCAT Reading, Mathematics, Science, and Writing+ multiple-choice **items** have four choices, only one of which is correct. Writing+ has some three-option multiple-choice **items**.

**Narrative Writing**—Writing that tells a story based on a real or imagined event. In FCAT Writing+, only students in Grade 4 are assigned a **prompt** intended to result in narrative writing.

**Norm-Referenced Test (NRT)**—A test designed to compare the performance of one group of students to a national sample of students, known as the "norm" group. The NRT portion of the FCAT includes both the Reading Comprehension and Mathematics Problem Solving subtests from the *Stanford 10* test published by Harcourt Assessment, Inc.

**Operational Items**—**Items** that count toward a student's score. Most **items** on the FCAT are operational **items**.

**Pattern Scoring**—A method of calculating a test score based on comparison of students' overall patterns of success on **items**. Pattern scoring shows inconsistencies in student responses (i.e., lack of success on an **item** with the same level of difficulty as other **items** with which the student had success).

**Performance Tasks**—**Items** that require students to provide either a short or extended written response. For example, short-response (SR) tasks may ask students to describe a character in a story, write a mathematical equation, or explain a scientific concept. Examples of extended-response (ER) tasks may include comparing two characters, constructing a graph, or describing the steps in an experiment.

**Persuasive Writing**—Writing that attempts to convince the reader that an opinion is valid or that the reader should take a specific action. In FCAT Writing+, students in Grades 8 and 10 are assigned **prompts** intended to result in persuasive writing.

**Pilot Test**—An assessment of a sample of students for the purpose of gaining general information about students' reactions to test **items**. Statistical analysis is not the focus of this initial tryout of **items**.

**Plan-Based Items**—A writing plan provides a prewriting structure and is based on a topic that is within the purview of students at the specified grade level. Possible graphic organizers may include charts, webs, diagrams, and outlines. In FCAT Writing+ tests, students answer questions about strengths and weaknesses of the writing plan.

**Prompt**—The topic a student is given on which to write an essay in FCAT Writing+. The **prompt** has two parts: the *writing situation* (presents and clarifies the topic) and the *directions for writing* (guides the student to think about the topic and suggests an approach that may help the student begin writing).

**Rangefinders**—Student responses to **prompts** (FCAT Writing+) or **performance tasks** (FCAT Reading, Mathematics, or Science) used to illustrate score points on the **rubric**. Rangefinding is the process of identifying these student responses. Also called **anchor papers**.

**Raw Score**—A score that reports the number of points a student earned on each test **item**, **cluster/strand**, or the entire test. Students earn one raw score point for each correctly answered **multiple-choice item** and **gridded-response item**, and up to four raw score points on **performance tasks**. Raw scores are reported as **content subscores**.

**Released Item**—A test question that has been released to the general public.

**Reliability**—Desired characteristic of a test. Achieved when measurement error is minimized and the test score is close to the **true score**.

**Retake**—Alternate Grade 10 reading or mathematics test given to those who do not achieve the passing score required for high school graduation.

**Rubric**—Scoring guidelines or criteria used to evaluate all FCAT **performance tasks** and essays. Describes what characterizes responses at each score point.

**Sample-Based Items**—A writing sample is an example of draft writing. Writing samples may be draft stories, reports, or articles that contain some mistakes. FCAT Writing+ **items** based on writing samples ask about the strengths and weaknesses of the sample.

**Scale Score**—Score used to report student results for the entire test in FCAT Reading, Mathematics, and Science. Scale scores on the FCAT range from 100 to 500 at each grade level. The scale score is the result of **IRT** processing and **equating**.

**Section 504**—Special classification of students as defined in Section 504 of the Rehabilitation Act of 1973. Testing accommodations are permitted for students who meet the Section 504 criteria.

**Short-Response Item (SR)**—See **Performance Tasks**.

**Stand-Alone Items**—Provide a succinct context for measuring the student's knowledge of the conventions of capitalization, punctuation, and sentence structure to address the breadth of the FCAT Writing+ editing **benchmark**.

**Standard**—In the ***Sunshine State Standards***, a statement of what students should know and be able to do. More specific than a **strand** and not as specific as a **benchmark**.

**Standard Error of Measurement (SEM)**—A whole-test **reliability** indicator that is calculated using data from the entire tested population. For example, 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 difference in the resulting test scores is called the standard error of measurement.

**Strands**—The broad divisions of **content areas** in the ***Sunshine State Standards***. For example, in the Language Arts ***Sunshine State Standards***, there are seven **strands** (Reading, Writing, Listening, Viewing, Speaking, Language, and Literature).

***Sunshine State Standards (SSS)***—Florida's curriculum framework that provides guidelines for what students should know and be able to do in each subject at each grade. Describes learning expectations at increasingly more detailed levels: **strands**, **standards**, and **benchmarks.** The purpose of the FCAT is to measure the *Sunshine State Standards* **benchmarks**. All FCAT **items** are based on specific **benchmarks**.

**Test Form**—A unique set of **items** consisting of a common core of **operational items** and a smaller number of either **field-test** or **anchor items**. FCAT Reading, Mathematics, and Science all use multiple test forms. Students with different test forms face exactly the same **operational items** but different **field-test** or **anchor items**.

**True Score**—FCAT seeks to measure a student's "true" achievement or true score on the content assessed. By definition, a student's test score is composed of two parts—the true score and the **standard error of measurement** associated with the test.

**Validity**—Desired characteristic of a test. Achieved when the test actually measures what it is intended to measure.

# 9.0 GUIDE TO RELATED RESOURCES

The following is a list of major topics related to the FCAT. A detailed description of each reference is provided in the first instance it is mentioned and thereafter only the title is listed. Other important information related to education in Florida can be found at the DOE web site at: www.fldoe.org.

| Topic | |
|---|---|
| **Accommodations** | *FCAT Test Administration Manual*—Describes procedures for FCAT administration, including roles and responsibilities of District Coordinators of Assessment, School Coordinators of Assessment, and Test Administrators; scripts for test administration; accommodations; and security measures. Three manuals are published and distributed each year: one for FCAT Writing+, one for FCAT Reading, Mathematics, and Science combined, and a third for FCAT Reading and Mathematics retakes.<br><br>For further information about accommodations, refer to the following technical assistance documents: "Planning FCAT Accommodations for Students with Disabilities" (product # 309603) and "Descriptions of FCAT Accommodations" (product #311930). These documents are available in print from the Bureau of Instructional Support and Student Services Clearinghouse/Information Center at (850) 245-0477. These and many other documents may be downloaded from the following web site: http://www.firn.edu/doe/commhome/fcatasd.htm |
| **Accountability for Schools** | *Assessment & Accountability Briefing Book*—Provides a summary of the FCAT program, including frequently asked questions, content assessed by FCAT, school accountability, FCAT results, and the history of the program. http://www.firn.edu/doe/sas/fcat/fcatpub1.htm<br><br>DOE Office of Evaluation and Reporting web site—Includes information on school grades and school accountability: http://www.firn.edu/doe/evaluation/home0018.htm |

## Content

*FCAT Test Item Specifications*—Guidelines for item writers and reviewers, including the *Sunshine State Standards* benchmarks to be tested, response formats for items associated with each benchmark, and other considerations for developing quality items. Separate documents for the individual grades or grade blocks in reading, mathematics, and science. *Item Specifications* for FCAT Writing+ are currently under development.
http://www.firn.edu/doe/sas/fcat/fcatis01.htm

*Keys to FCAT*—Contains information for parents and students preparing for FCAT Reading, Writing+, Mathematics, and Science. Distributed each January to district offices and available in English, Spanish, and Haitian Creole. Separate publications for Grades 3–5, 6–8, and 9–11.
http://www.firn.edu/doe/sas/fcat/fcatkeys.htm


## General Information

*About the FCAT* (online brochure)—Provides a summary for all FCAT subjects and grades. Available in English, Spanish, and Haitian Creole.
http://www.firn.edu/doe/sas/fcat/fcatpub3.htm

*FCAT Myths vs. Facts* (brochure)—Addresses common concerns about FCAT with relevant facts. Available in English and Spanish.
http://www.firn.edu/doe/sas/fcat/fcatpub1.htm

*FCAT Posters* (elementary, middle, and high school)—Focus on test-taking strategies and are available at district assessment offices. A poster for high school students identifies the scores necessary for passing the Grade 10 FCAT Reading and Grade 10 FCAT Mathematics and reminds students that they have multiple opportunities to retake the test.

*Frequently Asked Questions About the FCAT* (brochure)—Provides answers to frequently asked questions about the FCAT program.
http://www.firn.edu/doe/sas/fcat/fcatpub1.htm

| | |
|---|---|
| **History** | *Assessment & Accountability Briefing Book*<br><br>FDOE History of Statewide Assessment Program (HSAP) web site: www.firn.edu/doe/sas/hsaphome.htm |
| **Item Development** | *FCAT Test Item Specifications* |
| **Item Response Theory (IRT)** | *FCAT Technical Report*—Provides detailed information on item and test scoring and the statistical methods used to verify the quality of the items and test. The *2000* and *2002 Technical Reports* are available on the FCAT web site at http://www.firn.edu/doe/sas/fcat/fcatpub2.htm. For other years, portions not dealing with specific test items are available upon request from the DOE. |
| **No Child Left Behind** | The Florida Department of Education's No Child Left Behind web site: www.fldoe.org/NCLB<br><br>U.S. Department of Education's No Child Left Behind web site: www.ed.gov/nclb |
| **Norm-Referenced Tests** | DOE norm-referenced tests web site: www.firn.edu/doe/sas/nrthome.htm<br><br>FCAT NRT scores web site, including historic information for *Stanford 9 Reading Comprehension and Mathematics Problem Solving* tests from 2000–2004 administrations: http://www.firn.edu/doe/sas/fcat/nrinfopg.htm |
| **NRT SAT 10** | Information about the FCAT NRT (*SAT 10*): http://www.firn.edu/doe/sas/fcat/fcatpub2.htm |
| **Preparing Students for the FCAT** | The *FCAT Explorer* interactive web site is designed to help children strengthen the critical skills that are outlined in the *Sunshine State Standards* and tested on the FCAT. The *FCAT Explorer* features skills practice for both reading and math and includes passage and question topics from several areas of the *Sunshine State Standards*, such as social studies, science, and the arts. The *FCAT Explorer* web site: http://www.fcatexplorer.org/<br><br>*Keys to FCAT* |

*Sample Test Materials*—Produced and distributed each fall for teachers to use with students. The student booklet contains different kinds of FCAT questions and hints for answering them. The teacher's answer key provides the correct answer and an explanation for the correct answer and also indicates which *Sunshine State Standards* benchmark is being assessed by each question. Available for FCAT Reading, Mathematics, Science, and Writing+.
http://www.firn.edu/doe/sas/fcat/fcatsmpl.htm

*What every teacher should know about FCAT*—Provides suggestions for all subject-area teachers to use in helping their students be successful on the FCAT.
http://www.firn.edu/doe/sas/fcat/fcatpub2.htm

**Reporting and Results**

FCAT Developmental Scores web site:
http://www.firn.edu/doe/sas/fcat/fcat_score/index.htm

FCAT Results web site:
http://fcat.fldoe.org/

FCAT Scores and Reports web site:
http://www.firn.edu/doe/sas/fcat/fcatscor.htm

*Florida Inquires! Report on the [test administration year] FCAT Science Released Items*—Guide to the scoring of the FCAT Science performance tasks displayed on the Grades 5, 8, and 11 student reports. Distributed to districts each May.

*Florida Reads! Report on the [test administration year] FCAT Reading Released Items*—Guide to the scoring of the FCAT Reading performance tasks displayed on the Grades 4, 8, and 10 student reports. Distributed to districts each May.

*Florida Solves! Report on the [test administration year] FCAT Mathematics Released Items*—Guide to the scoring of the FCAT Mathematics performance tasks displayed on the Grades 5, 8, and 10 student reports. Distributed to districts each May.

*Florida Writes! Reports on the [test administration year] FCAT Writing Assessment*—For educators involved in teaching, planning, and evaluating curriculum in the Florida public schools. Separate publications for Grades 4, 8, and 10 describe the content and application of FCAT Writing+ prompted essay and offer suggestions for activities that may be helpful in preparing students for the assessment. Distributed to districts each May.


*Lessons Learned*—FCAT, *Sunshine State Standards and Instructional Implications*—Provides an analysis of previous years' FCAT results and contains analyses of state FCAT Reading, Writing, and Mathematics data through 2000. Intended to assist educators in interpreting and understanding their local FCAT scores in order to help improve classroom instruction.
http://www.firn.edu/doe/sas/fcat/fclesn02.htm

Results spreadsheets are archived at the web site below, providing school, district, and state means for each grade tested by year.
http://www.firn.edu/doe/sas/fcat/fclesn02.htm

*Understanding FCAT Reports*—Provides information about the FCAT student, district, and school reports for the most recent test administration. Samples of reports, explanations about the reports, and a glossary of technical terms are included. Distribution to districts is scheduled to coincide with the delivery of student reports each May.
http://fcat.fldoe.org/

| | |
|---|---|
| **Sample Items** | *FCAT Explorer* interactive web site: http://www.fcatexplorer.org/ |
| | *FCAT Test Item Specifications* |
| | *Keys to FCAT* |
| | *Sample Test Materials* |
| **Scoring Procedures and Methodology** | *FCAT Performance Task Scoring—Practice for Educators* (publications and software)—Designed to help teachers learn to score FCAT Reading, Writing, and Mathematics performance tasks at Grades 4, 5, 8, and 10. A *Trainer's Guide* includes instructions for using the scoring publications and software in teacher education seminars and workshops. The publications mirror the scorer training experiences by presenting samples of student work for teachers to score. |
| | *FCAT Scoring Rubrics* http://www.firn.edu/doe/sas/fcat/rubrcpag.htm |
| | *Florida Inquires! Report on the [test administration year] FCAT Science Released Items* |
| | *Florida Reads! Report on the [test administration year] FCAT Reading Released Items* |
| | *Florida Solves! Report on the [test administration year] FCAT Mathematics Released Items* |
| | *Florida Writes! Reports on the [test administration year] FCAT Writing+ Assessment* |
| **Security Measures** | *FCAT Test Administration Manual* |
| ***Sunshine State Standards*** | DOE *Sunshine State Standards* web site: http://www.firn.edu/doe/menu/sss.htm |
| | *FCAT Test Item Specifications* |
| **Test Administration Procedures** | *FCAT Test Administration Manual* |

# APPENDIX A: STATISTICAL INDICATORS USED IN TEST DATA ANALYSIS

After field testing, during the test construction process, and after operational testing, a series of statistical analyses are performed on FCAT items and the test as a whole to ensure that established criteria for items and test forms have been or will be met. The purpose of the review is to determine whether individual items can be used in the future as operational items. During test construction, data are reviewed for individual items and proposed test forms. After operational testing, data are generated from a sample of students representative of all students tested (the calibration sample) to generate the parameters necessary for scoring (IRT processing) and to determine whether any items require special treatment in the scoring process. Additional measures are generated after scoring to verify the reliability of the test and the accuracy and consistency of the Achievement Level classifications.

It is important to remember that items not meeting established criteria may be rejected for use as operational items or excluded from calculation of student scores. These instances are rare because the processes of item development and test construction are carefully guided and include many quality control measures.

The following information on the various indicators is more detailed than that presented in the body of this publication. For even more detailed information, including selected data for a given year, refer to the *FCAT Technical Report*. (The *FCAT Technical Reports* are available on the FCAT web site: http://www.firn.edu/doe/sas/fcat/fcatpub2.htm.)

**TABLE 16: STATISTICAL ANALYSES FOR TEST DATA AND INDICATORS**

| Purpose | Indicator |
|---|---|
| Describe item difficulty | *p*-values, IRT *b*-parameters |
| Compare likelihood of success on item with likelihood of success on test | Item-total correlations, IRT *a*-parameters |
| Estimate gain from guessing | IRT *c*-parameters |
| Measure item fit to IRT model | $Q_1$ ($Z_{Q1}$) statistics |
| Measure test fit to IRT model (unidimensionality of achievement scale) | $Q_3$ statistics |
| Identify bias | Differential Item Functioning (DIF) analysis (Mantel-Haenszel statistic, Mantel statistic, SMD rating) |
| Measure reliability | Standard error of measurement (conditional SEM), Marginal reliability index, Cronbach's alpha |
| Verify Achievement Level classification accuracy and consistency | Indices of accuracy and consistency: overall, conditional-on-level, cut point |

## Indicator Definitions

**Differential Item Functioning (DIF)**—Indicates differences in scores between males and females and between ethnic groups that are unique to the item and cannot be explained by differences between these groups in overall achievement. Test developers use two types of measures of DIF, the *Mantel-Haenszel statistic* (and a variation of it, the *Mantel statistic*, used for performance task items) and *standardized mean differences* (SMDs). To derive both types of measures, all students are divided into groups with similar total test scores. Within these groups, scores for each individual item are compared between males and females and between ethnic groups (i.e., African American, Caucasian, and Latin American). If an item is not biased, then these comparisons should yield no difference in performance because the individuals being compared are already at the same level of overall achievement. On the other hand, if an item is biased against a particular gender or ethnic group, there will be a difference in performance on that item, a difference that is inconsistent with overall test performance. The Mantel-Haenszel statistic (and the Mantel statistic) indicates whether there are any statistically significant differences in performance; the SMDs indicate the magnitudes of these differences.

**IRT *a*-parameter**—Represents the degree to which the item differentiates between test takers with different abilities.

**IRT *b*-parameter**—Interpreted similarly to *p*-values, indicates where the item slope is centered on the ability scale.

**IRT *c*-parameter**—Estimates the gains from guessing by comparing student success on any given item with the pattern of success on all the other items. A high *c*-parameter results when student success on the item is inconsistently high in comparison to success on other items of similar or lesser difficulty.

**Item-Total Correlations**—Measures the correlation between the score on an item and the total score for all items (raw score). Reported for individual items and as a single summary statistic for all items within a content cluster, and for all items on the test as a whole. Examples of item-total correlations are the point-biserial correlation, the biserial correlation, and the Pearson product moment correlation.

***p*-value**—A measure of student success on an item, equivalent to the mean score on the item divided by the total score points available for it. For multiple-choice and gridded-response items, this is the same as the percentage of students answering the item correctly.

**$Q_1$ Statistic**—Uses an item's IRT function to estimate students' expected performances on the item and then compares the estimates to students' actual performances. Low values indicate little difference and good fit of the test data to the IRT model. The $Z_{Q1}$, an adjustment of the $Q_1$ statistic, is used for FCAT analysis purposes.

**$Q_3$ Statistic**—Uses the IRT parameter estimates to generate item scores for students based on overall achievement data and then compares the estimate to actual student performance. These differences, the residuals, represent the influence on performance of factors other than the true ability. They are then compared for all possible pairs of items on the test. If differences in performance between items in a pair are due solely to differences in item difficulty, and thus to no other factors, there will be little correlation between each pair of residuals, and $Q_3$ will be low.

## Reliability Measures

**Standard Error of Measurement (SEM), Marginal Reliability Index, Cronbach's Alpha**—In statistical terms, reliability is a ratio of the variation in true achievement (that the test seeks to estimate) to variation in observed test scores, which are subject to error. If the error is minimal, the ratio will be close to 1, and the test can be said to be reliable. The review of FCAT statistical characteristics is based on three indicators of reliability: *conditional standard error of measurement*, *marginal reliability*, and *Cronbach's alpha*. The SEM describes the error associated with different levels of overall achievement. SEMs for the complete range of scores are often represented graphically as *conditional standard error curves* to illustrate where the error is lowest. Typically, the error is lowest in the middle of the achievement spectrum because there are more items associated with this level of achievement than at the extremes. *Marginal reliability* is a measure of the overall reliability of the test based on the average SEM for all students. *Cronbach's alpha* is a traditional measure of test reliability in which the degree of error is assumed to be the same at all levels of student achievement.

**Achievement Level Classification Consistency and Accuracy**—*Consistency* of classification is the agreement between classifications based on two equally difficult forms of the test. *Accuracy* of classification is the degree to which actual classifications agree with those that would be made on the basis of students' true abilities, if they could be known. Three types of accuracy and consistency indices are estimated for the FCAT tests: *overall*, *conditional-on-level*, and by *cut point*. To describe consistency, these indices examine the agreement between actual performance and performance on a statistically modeled alternate and parallel test form. To describe accuracy, they examine agreement between actual performance and a statistically constructed true score. *Overall* indices show the classification agreement grouped across all Achievement Levels; indices *conditional-on-level* outline the agreement at a selected Achievement Level; and indices by *cut point score* show the agreement around a single Achievement Level cut point.

# APPENDIX B: SECURITY AGREEMENT

## TEST SECURITY AND NON-DISCLOSURE AGREEMENT
## FLORIDA DEPARTMENT OF EDUCATION

**2004**

Florida State Board of Education Rule 6A-10.042, FAC, was developed to meet the requirements of the test security statute, Section 1008.24, FS, and applies to everyone involved in the administration, handling, scoring, and reporting of a statewide assessment test. The rule prohibits activities that may threaten the integrity of the test. Prohibited activities include:

- revealing or copying test items;
- revealing student responses to test items;
- changing or otherwise interfering with student responses; and
- causing individual, school, district, or state achievement to be inaccurately measured or reported.

I, _____, affirm that:

I have received and am responsible for reading and complying with the Florida test security statute, Section 1008.24, FS, and State Board of Education test security rule, Rule 6A-10.042, FAC. I understand that persons violating the law may be guilty of a first-degree misdemeanor, punishable by a fine of not more than $1,000 or imprisonment of not more than 90 days, or both.

I further affirm that I know that during the process of reviewing test items, I will have access to secure testing materials. I agree to the following:

- I shall not reveal, copy, reproduce, or use in any manner inconsistent with test security rules any secure information, secure testing materials, or portions of any secure testing materials; and
- I understand that all secure testing materials and secure information to which I have access are and shall remain the exclusive property of the State of Florida.
- I will NOT remove any secure testing materials or information from the review site.
- I acknowledge that the intellectual property rights subsisting in the materials related to these assessments are the property of the Florida Department of Education.

I further affirm that I understand that I may **NOT** use or share any secure test material or secure information gained from my involvement in reviewing the test items and the assessments.

By virtue of the foregoing, I am on notice that any actions by me that are contrary to the foregoing affirmations and acknowledgements will subject me to possible legal action by the Florida Department of Education to protect its interest in its intellectual property rights and the integrity and security of the assessments.

_____ _____

Signature Date

_____ _____

Witnessed by Date

## Test Security Requirements, Statutes, and Rule

Chapter 1008.24 of Florida Statutes and Florida State Board of Education Rule 6A-10.042 establish the requirement that Florida Department of Education tests are to be maintained in a secure manner during development, administration, and scoring in order to preserve the integrity of the tests. When not in use, all test materials are to be kept in secure, locked storage. Individuals who have access to secure test materials are not to copy or otherwise reproduce test questions or reveal test questions verbally or in writing. Persons who are involved in administering or proctoring the test or preparing examinees for the tests are not to participate in, direct, aid, counsel, assist in, or encourage any activity which could result in the inaccurate measurement or reporting of the examinees' achievement. Examinees' answers to questions are not to be interfered with in any way by persons administering or scoring the tests. Persons violating test security requirements are guilty of a first-degree misdemeanor, punishable by a fine of not more than $1,000.00 or imprisonment for not more than 90 days, or both.

The security requirements and penalties established by the rule and statute must be provided by the contractor to each person who has access to tests or test questions during the development, printing, administration, or scoring of the tests. A copy of the Statute and Rule begins on the next page.

## Florida Test Security Statute

1008.24 Test Security

(1) It is unlawful for anyone knowingly and willfully to violate test security rules adopted by the State Board of Education for mandatory tests administered by or through the State Board of Education to students, educators, or applicants for certification or administered by school districts pursuant to §1008.22, or, with respect to any such tests, knowingly and willfully to:

(a) Give examinees access to test questions prior to testing;

(b) Copy, reproduce, or use in any manner inconsistent with test security rules all or any portion of any secure test booklet;

(c) Coach examinees during testing or to alter or interfere with examinees' responses in any way;

(d) Make answer keys available to examinees;

(e) Fail to follow security rules for distribution and return of secure test materials as directed, or fail to account for all secure test materials before, during, and after testing;

(f) Fail to follow test administration directions specified in the test administration manuals; or

(g) Participate in, direct, aid, counsel, assist in, or encourage any of the acts prohibited in this section.

(2) Any person who violates this section commits a misdemeanor of the first degree, punishable as provided in § 775.082 or § 775.083.

(3) A district superintendent of schools, a president of a community college, a president of a university, or a president of a private postsecondary institution shall cooperate with the commissioner of Education in any investigation concerning the administration of a test administered pursuant to state statute or rule.

History § 370, ch. 2002-387.

## Rule 6A-10.042, FAC Maintenance of Test Security

(1) Tests implemented in accordance with the requirements of Sections 229.053(2)(d), 229.57, 231.087, 231.0861(3), 231.17, 233.011, 239.301(10), 240.107(8), and 240.117, Florida Statutes, shall be maintained and administered in a secure manner such that the integrity of the test shall be preserved.

(a) Test questions shall be preserved in a secure manner by individuals who are developing and validating the tests. Such individuals shall not reveal in any manner, verbally or in writing, the test questions under development.

(b) Tests or individual test questions shall not be revealed, copied, or otherwise reproduced by persons who are involved in the administration, proctoring, or scoring of any test.

(c) Examinees shall not be assisted in answering test questions by any means by persons administering or proctoring the administration of any test.

(d) Examinees' answers to questions shall not be interfered with in any way by persons administering, proctoring, or scoring the examinations.

(e) Examinees shall not be given answer keys by any person.

(f) Persons who are involved in administering or proctoring the tests or persons who teach or otherwise prepare examinees for the tests shall not participate in, direct, aid, counsel, assist in, or encourage any activity which could result in the inaccurate measurement or reporting of the examinees' achievement.

(g) Each person who has access to tests or test questions during the development, printing, administration, or scoring of the test shall be informed of specifications for maintaining test security, the provisions in statute and rule governing test security, and a description of the penalties for breaches of test security.

(h) During each test administration, school district and institutional test administration coordinators and contractors employing test administrators and proctors shall ensure that required testing procedures are being followed at all test administration sites. Officials from the Department are authorized to conduct unannounced observations of test administration procedures at any test administration site to ensure that testing procedures are being correctly followed.

(2) Test materials, including all test booklets and other materials containing secure test questions, answer keys, and student responses, shall be kept secure and precisely accounted for in accordance with the procedures specified in the examination program administration manuals and other communications provided by the Department. Such procedures shall include but are not limited to the following:

(a) All test materials shall be kept in secure, locked storage prior to and after administration of any test.

(b) All test materials shall be precisely accounted for and written documentation kept by test administrators and proctors for each point at which test materials are distributed and returned.

(c) Any discrepancies noted in the number or serial number of testing materials received from contractors shall be reported to the Department by designated institutional or school district personnel prior to the administration of the test.

(d) In the event that test materials are determined to be missing while in the possession of an institution or school district, designated institutional or school district personnel shall investigate the cause of the discrepancy and provide the Department with a report of the investigation within thirty (30) calendar days of the initiation of the investigation. At a minimum, the report shall include the nature of the situation, the time and place of occurrence, and the names of persons involved in or witness to the occurrence. Officials from the Department are authorized to conduct additional investigations.

(e) In those cases where the responsibility for secure destruction of certain test materials is assigned by the Department to designated institutional or school district personnel, the responsible institutional or school district representative shall certify in writing that such destruction was accomplished in a secure manner.

(f) In those cases where test materials are permitted by the Department to be maintained in an institution or school district, the test materials shall be maintained in a secure manner as specified in the instructions provided by the Department. Access to the materials shall be limited to the individuals and purposes specified by the Department.

(3) In those situations where an employee of the educational institution, school district, or contractor, or an employee of the Department suspects a student of cheating on a test or suspects other violations of the provisions of this rule, a report shall be made to the Department or test support contractor, as specified in the test administration procedures, within ten (10) calendar days. The report shall include a description of the incident, the names of the persons involved in or witness to the incident, and other information as appropriate. Officials from the Department are authorized to conduct additional investigations.

(4) Violations of test security provisions shall be subject to penalties provided in statute and State Board Rules.

Specific Authority 120.53(1)(b), 1008.24, 229.053(1) FS. Law Implemented 120.53(1)(b), 228.301, 229.053(2)(d), 229.57, 231.087, 231.0861, 231.17, 233.011, 239.301, 240.107, 240.117 FS. History-New 7-5-87, Amended 10-26-94

# APPENDIX C: FCAT NEWSPAPER ARTICLES

## Study Praises FCAT as Indicator of Learning

By Leslie Postal, staff writer
*The Orlando Sentinel,* Orlando, Florida

As thousands of Florida students start taking FCAT today, a new study finds the test a reliable gauge of academic performance and refutes the complaint that it encourages teaching to the test over real learning.

The Florida Comprehensive Assessment Test, the linchpin in the state's education reform efforts, comes with high-stakes—the FCAT is used to grade public schools, help determine third-grade promotions and decide whether high school seniors earn a diploma.

This year's FCAT season begins today, as students sit for writing exams. Reading and math tests and a new science exam will be given in the coming weeks. Students in third through 10th grades take the tests.

FCAT critics argue that high-stakes testing creates a "narrow" focus that doesn't lead to "real learning." But the improved scores Florida students have shown in recent years appear to reflect genuine learning, according to a study by the Manhattan Institute of Policy Research.

That's because Florida students have also improved on the Stanford 9, a national test given at the same time as FCAT. The national test comes with no consequences, so "it's like an audit" of FCAT, said Jay Greene, a senior fellow with the institute and the study's lead author.

Similar results on the two tests, Greene said, suggest improvement on FCAT has come from real learning and has not been distorted by the pressure and consequences tied up with the exams. If teachers are teaching to the FCAT, he added, they are teaching what most everyone wants students to learn.

"If the test is well designed, and it's administered properly, the teaching to the test might be a good thing," he added. "It means students are learning to read and to do math."

Greene of Davie said his study should put to rest that one major complaint about FCAT. But it also raised some questions about testing programs elsewhere.

The institute's study backed up what Florida education officials have said about FCAT — but contradicted two recent studies by Arizona State University researchers.

Audrey Amrein, one of the Arizona researchers, contends high-stakes tests have become an "easy way to control what's going on in the schools," though they haven't led to real progress.

Her studies compared results on high-stakes state tests with results on national tests, such as the SAT, Advanced Placement exams and the National Assessment of Education Progress. They found states with high-stakes tests had made little gains, or had posted declines, on these national exams. Their results for Florida, however, were not clear-cut.

Greene criticized those studies, saying it didn't make sense to compare scores on tests only college-bound students take with scores on yearly state tests almost everyone takes.

His study looked at high-stakes and low-stakes tests given to the same groups of students at the same time. The study looked at Florida and Virginia and at seven school districts including Boston and Chicago.

In both states and all the districts, the standardized tests were an accurate measure of student performance. But only in Florida did the high-stakes tests also accurately measure a school's role in student achievement from year to year.

# Educators Help Shape FCAT

## Three locals attend sessions to ensure test's effectiveness

By Kimberly C. Moore, staff writer
*Florida Today,* Brevard County, Florida

MELBOURNE — Psssst. Want to know a secret?

Three local educators are involved in the super-secret, ultra-secure process of helping to develop the Florida Comprehensive Assessment Test. It's the same test Gov. Jeb Bush jokingly refers to as "dreaded and hated" but seriously touts as a way to save Florida students from years of broken dreams.

Edgewood Jr./Sr. High school teacher David Bradford, Madison Middle School administrator Fielding Hossley and Apollo Elementary School teacher Marie Clifford recently participated in sessions to rewrite, rework or revisit FCAT questions on the reading, writing and math tests.

Bradford joked that like a CIA operative, he could tell you what the test questions are — but then he'd have to kill you.

"The security was very impressive," Bradford said. "We were sworn to secrecy and read the state law. There were no armed guards, but every scrap of paper was picked up. We signed in and out, and the conference rooms were locked."

Inside those locked rooms were groups of six panelists consisting of teachers "on the front lines" in Florida classrooms. They read every question, sometimes a dozen times, to ensure it was appropriate for the grade level being tested, free from slang or slurs and not offensive. They reviewed the answers.

They took the test. In the end, they threw out some questions and answers and reworded others.

Their mission was to ensure that questions adhered to Sunshine State Standards, the benchmark of what students are supposed to be learning at different grade levels.

"We don't want the questions to be the test," Hossley said. "We want to determine whether the student understood the reading task that was being assessed."

Bradford and Hossley said a priority was making sure each word in each question and answer was appropriate for the grade level tested.

"For instance, a question might ask, "What is your opinion of the way the author described Johnny?'" Hossley said. "If this was a third-grade test, the student might not understand the word 'opinion.'"

Both say they appreciate the FCAT and how the test aims to pinpoint students' academic weaknesses. The results can assist teachers in helping students strengthen skills.

"I was an outspoken critic of high-stakes testing, and I came home with a whole different view of it," Bradford said. "This test is fair, it's unbiased and it's sensitive to what's happening in our own state and our communities."

Hossley said he agrees.

"I think it's a good test and a good measure of what we're supposed to be teaching students," Hossley said. "Although the political side of it, I have some concerns with."

**Editor's Note:**
David Bradford has served on the FCAT Reading Item Review Committee.

Marie Clifford has served on the FCAT Science Item Review and Science Performance Review Committees.

Fielding Hossley has served on the following FCAT committees: Writing Content Advisory, Writing Rangefinder, Reading Item Review, and Reading Standards Setting.

## FCAT Gets High Marks in Measuring Achievement

By Steve Harrison, staff writer
*The Miami Herald,* Miami, Florida

Florida's FCAT accurately measures student achievement and cannot be manipulated by "teaching to the test," according to a study released Monday by a Davie-based education think tank.

The Manhattan Institute wanted to determine whether the pressure placed on schools to perform on high-stakes examinations such as the Florida Comprehensive Assessment Test prompts teachers to spend too much time teaching skills that don't improve a child's education.

Under the No Child Left Behind Act of 2001, every state must have a high-stakes test to receive federal funding. Florida's test is considered the most aggressive in the nation.

The Manhattan Institute, a conservative-leaning organization that has criticized and supported Gov. Jeb Bush's education record, released a study last year that said Florida had the lowest graduation rate in the nation.

But the institute had high praise Monday for Florida, calling the FCAT the best high-stakes test of the nine the firm studied, including those used by Chicago, Boston and Toledo and the state of Virginia.

"If schools are 'teaching to the test,' they are doing so in a way that conveys useful general knowledge as measured by nationally respected low-stakes tests," the report said.

As its reference point, the institute used a national test called the SAT-9.

In Florida, the SAT-9 is used to compare students with their peers nationwide.

The authors wanted to see if there was a correlation between a school's rise or decline on the FCAT and a similar rise or decline on the SAT-9.

For example, if a high-stakes test like the FCAT emphasized spelling, teachers might focus on improving students' spelling performance but neglect reading and mathematics, which is measured by the SAT-9.

Florida's FCAT results closely matched the SAT-9.

"I've always believed the test is a good measure, a valid measure," said Anne Dilgen, the Broward public school district's director of student assessment. "It's testing what we are trying to teach."

The Florida Education Association has raised objections about the FCAT, not because the group doesn't support the test, said Tony Welch, association spokesman. The union is concerned about how the test is used, he said.

"Most everyone says the FCAT is a good test. That's uniform," Welch said. "The objection is that everything boils down to this one test. It might not catch all the improvement a student has done over the year. There are very good students who won't perform well."

## Teach the Test

*The Florida Times-Union,* Jacksonville, Florida

Although it is little appreciated by those who think the public schools do not need to have any accountability, the Florida Comprehensive Assessment Test is a useful tool, according to a think tank that looked at the facts.

A report released today by the Manhattan Institute shows that the FCAT accurately measures student proficiency. While some are concerned that FCAT results do not measure real learning, the study finds that the FCAT is a reliable measure of student achievement.

Institute scholars Jay P. Greene, Marcus A. Winters and Greg Forster compared the FCAT results with another test, the Stanford-9, and found the scores track closely.

One of the chief criticisms of the FCAT is that teachers "teach the test" and are not actually learning anything except the answers to the test.

Because it has repercussions, the FCAT is called a "high stakes" test. The education establishment continually rails about high-stakes testing as being an unfair burden on teachers and students.

But the Stanford test is not used for accountability and there is no incentive for teachers to teach that test. Presumably, students would test lower.

The researchers examined test results at all schools in Florida and in eight other school systems nationwide. They concluded high-stakes testing is an accurate measure of student proficiency.

"Our findings suggest that if Florida teachers are focusing exclusively on FCAT material, as some claim, then in doing so they are teaching skills that are generally useful rather than useful only to pass a single standardized test" the report said.

"By forcing teachers to alter their curricula and teaching techniques in order to get their students to pass the FCAT, Florida has forced them to better prepare their students for life outside the classroom walls. The evidence suggests that the FCAT has effectively communicated to teachers and schools what general knowledge they must teach, and provided them with incentives to ensure that students acquire that knowledge."

At least teaching a test is teaching something.

# Appendix D: Participation in FCAT Committees

The development and implementation of the FCAT have been shaped by the active involvement of thousands of Florida educators serving on FCAT committees. Since 1996, educators have guided the development of the *Sunshine State Standards*, the determination of which benchmarks to assess and how to assess them on the FCAT, and how essays as well as other performance tasks should be scored. All FCAT test items have been reviewed and accepted by committees of Florida educators. The DOE maintains open communication with Florida educators regarding how the FCAT and the various associated processes and activities might be improved. Educators may be nominated to FCAT committees by their District Superintendent, district-level administrators, or by peers serving on FCAT committees.

## Standing Committees

### *Rotating membership*

**Assessment and Accountability Advisory Committee**—This committee has 15–20 members representing educators, school district personnel, and university faculty. They advise the DOE about K–12 assessment and accountability policies. Their recommendations may relate to standards for FCAT Achievement Levels, school grading policies, and alternative assessments. The committee meets once a year.

**Reading Content Advisory Committee**—This committee is composed of 15–20 reading and/or language arts professionals from schools, school districts, and universities. They advise the DOE about the scope of the reading assessment. Their recommendations may include which benchmarks should be assessed on FCAT Reading, the item types recommended for each benchmark, the types of reading materials to be used, the range of difficulty for passages to be used on the FCAT, and the number of benchmarks, passages, and items to be assessed per grade level. This committee meets once or twice a year.

**Writing Content Advisory Committee**—This committee is composed of 15–20 writing or language arts professionals from schools, school districts, and universities. They advise the DOE about the scope of the writing assessment, including the benchmarks that should be assessed and the item types recommended for each benchmark. In years prior to 2000, this committee was constituted as separate grade-level committees and was used to advise the DOE about the implementation of the Florida Writing Assessment Program. In 2000–2001, the title FCAT Writing was used and their discussions were broadened to include an expanded assessment of writing assessment topics (FCAT Writing+). This committee meets once or twice a year.

**Mathematics Content Advisory Committee**—This committee is composed of 15–20 mathematics professionals from schools, school districts, and universities. They advise the DOE about the scope of the mathematics assessment. Their recommendation may relate to the benchmarks that should be assessed on FCAT Mathematics and the item types recommended for each benchmark. This committee meets once or twice a year.

**Science Content Advisory Committee**—This committee is composed of 15–20 science professionals from schools, school districts, and universities. They advise the DOE about the scope of the science assessment. Their recommendations may relate to the benchmarks that should be assessed on FCAT Science and the item types recommended for each benchmark. This committee meets once or twice a year.

**Interpretive Products Advisory Committee**—This committee is composed of 8–10 professionals that represent the many audiences for which FCAT materials are prepared. Members from Florida school districts and the private sector bring experience related to exceptional student education, ESOL, vocational education, post-secondary education, and parent involvement. This committee meets no more than once a year to review FCAT publications and provide input to the DOE for future products.

**Technical Advisory Committee**—This committee is composed of 10–15 professionals with expertise in psychometrics. The members include Florida district test directors, representatives from the FCAT Content Advisory Committees, Florida university faculty members, and representatives of universities and state agencies outside of Florida. In addition, the psychometric advisors of the DOE's contractors participate in the meetings of this committee. Committee members assist the DOE by reviewing technical decisions and documents, and by providing advice regarding the approaches the DOE should use to analyze and report FCAT data. This committee meets once or twice a year.

## Annual and Ad Hoc Committees

**Prompt Review Committee**—This committee reviews the prompts and student responses from the FCAT Writing pilot test. The review ensures that prompts selected for the FCAT employ clear wording, are of appropriate difficulty and interest level, and are unbiased. The purpose of the committee is to select prompts for the FCAT Writing Field Test. Participants include language arts teachers from the targeted grade level, and school and district curriculum specialists. This committee meets each year in the fall after the pilot test.

**Community Sensitivity Committee**—Florida citizens associated with a variety of organizations and institutions review all passages, prompts, and items for issues of potential concern to members of the community at large. This review ensures that the primary purpose of assessing achievement is not undermined by inadvertently including in the test any material that may be deemed inappropriate by parents and other citizens. Reviewers are asked to consider whether the subject matter and language of each reading passage, writing prompt, or test item will be acceptable to Florida students, their parents, and other members of Florida communities. Participants in these committees include representatives of statewide religious organizations, parent organizations, community-based organizations, and cultural groups (e.g., Hispanic or Native American), school boards, school district advisory council members, and leaders in business and industry from across the state. Each Community Sensitivity Committee meets once a year.

**Bias Review Committee**—Groups of Florida educators representative of Florida's regional, racial/ethnic, and cultural diversity review passages, prompts, and items for potential bias. Reviewers look for the following types of bias: gender, racial/ethnic, linguistic, religious, geographic, and socioeconomic. A test item (or prompt or passage) is considered biased if characteristics of the item, unrelated to the skill being measured, result in an unfair advantage or disadvantage for a particular group of students. (In addition to this professional judgment model, differential item functioning [DIF statistic] is examined for all FCAT items.) Participants in these committees include representatives of Florida school districts, universities, and statewide organizations that serve the various groups that are potentially affected by the types of bias described, such as Title I, ESOL, and EEO. Every attempt is made by the DOE to represent the various groups potentially affected by bias at a level at or above their representation in the general population. Each Bias Review Committee meets once a year.

**Item Content Review Committee**—Content reviews are conducted for reading passages and reading, mathematics, science, and writing test items to determine whether the passages and items are appropriate for the grade level for which each is proposed. In addition, participants are asked to evaluate whether the items measure the benchmark, are clearly worded, have one and only one correct answer, or are of appropriate difficulty. Participants include teachers from the targeted grade level and subject area, and school and district curriculum specialists. The Item Content Review Committees usually meet during late fall each year.

**Science Expert Review Committee**—Due to the theory-based nature of the content area, all potential science test items undergo an extra level of scrutiny. Participants in this committee review newly developed science test items to ensure the accuracy and currency of the science content. Participants include practicing scientists from the private sector and university-level science researchers and faculty. The Science Expert Review Committee usually meets during late fall each year.

**Rangefinder Committee**—After performance items (short- and extended-response) and writing prompts are field tested on the FCAT, scoring of a representative set of student responses for each item/prompt is conducted to establish guidelines for the handscoring of all students' responses. Participants establish the range of responses that represent each score point of the rubric for each item or prompt. As a result of these meetings, training materials for handscorers are assembled. Participants include teachers from the targeted grade level and subject area, and school and district curriculum specialists. Participants will have served on other FCAT committees, such as Item Content Review Committee, prior to serving on a Rangefinder Committee. The Rangefinder Committees meet after spring testing and prior to handscoring of field-test performance items.

**Rangefinder Review Committee**—After performance items and writing prompts are selected for use on the FCAT, a scoring and review of a representative set of student responses is conducted to establish guidelines for the handscoring of all responses. Participants discuss and verify the range of student responses that represent each score point of the rubric for each item or prompt. As a result of these meetings, training materials for handscorers are reviewed and, if necessary, revised. Participants include teachers from the targeted grade level and subject area, and school and district curriculum specialists. Participants will have served on other FCAT committees, such as the Rangefinder Committee, prior to serving on a Rangefinder Review Committee. The Rangefinder Review Committees meet in the late fall.

**Gridded-Response Adjudication Committee**—A review of all field-test responses to gridded-response questions is conducted to determine whether all possible correct answers have been included in the scoring key. The various responses are examined and judged as either incorrect or correct. Committee members are asked to evaluate the possibility of finding the answer through an alternate process and determine if resulting answers are acceptable. Based on their advice, the DOE establishes rules for how each gridded-response item will be scored. Participants include teachers from the targeted grade level and subject area, and school and district curriculum specialists.

**Standards Setting Committees**—From time to time the DOE seeks the advice of district educators and business/community representatives to recommend Achievement Level standards for the FCAT. For example, committees were used to recommend the Achievement Levels for FCAT Reading and Mathematics currently in place. For these committees, members are selected from persons familiar with the FCAT from prior committee participation and persons who may be unfamiliar with the FCAT but have an interest in the standards being established. Participants include teachers from the targeted grade level and subject area, school and district curriculum specialists, school and district administrators, university faculty from the discipline area, and business and community leaders.

**Special Ad Hoc Committees**—On occasion, groups of parents, teachers, school/district administrators, and others are convened to review various aspects of the testing program and to advise the DOE on appropriate courses of action. These committees provide advice on such issues as score reporting, norm-referenced tests, and interpretive products.


**FLORIDA DEPARTMENT OF EDUCATION**
**www.fldoe.org**

1 2 3 4 5 6 7 8 9 10 11 12 A B C D E

ISBN 999826919-9

9 789998 269194

Assessment and School Performance
Florida Department of Education
Tallahassee, Florida