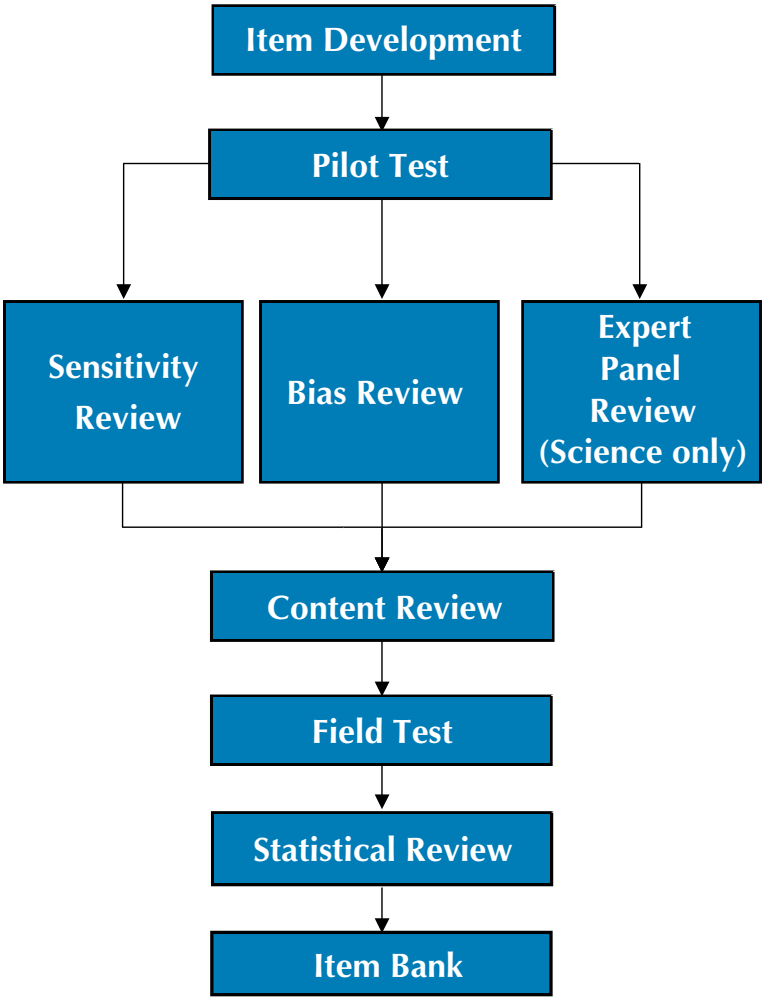
4.0 Test Development and Construction

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.

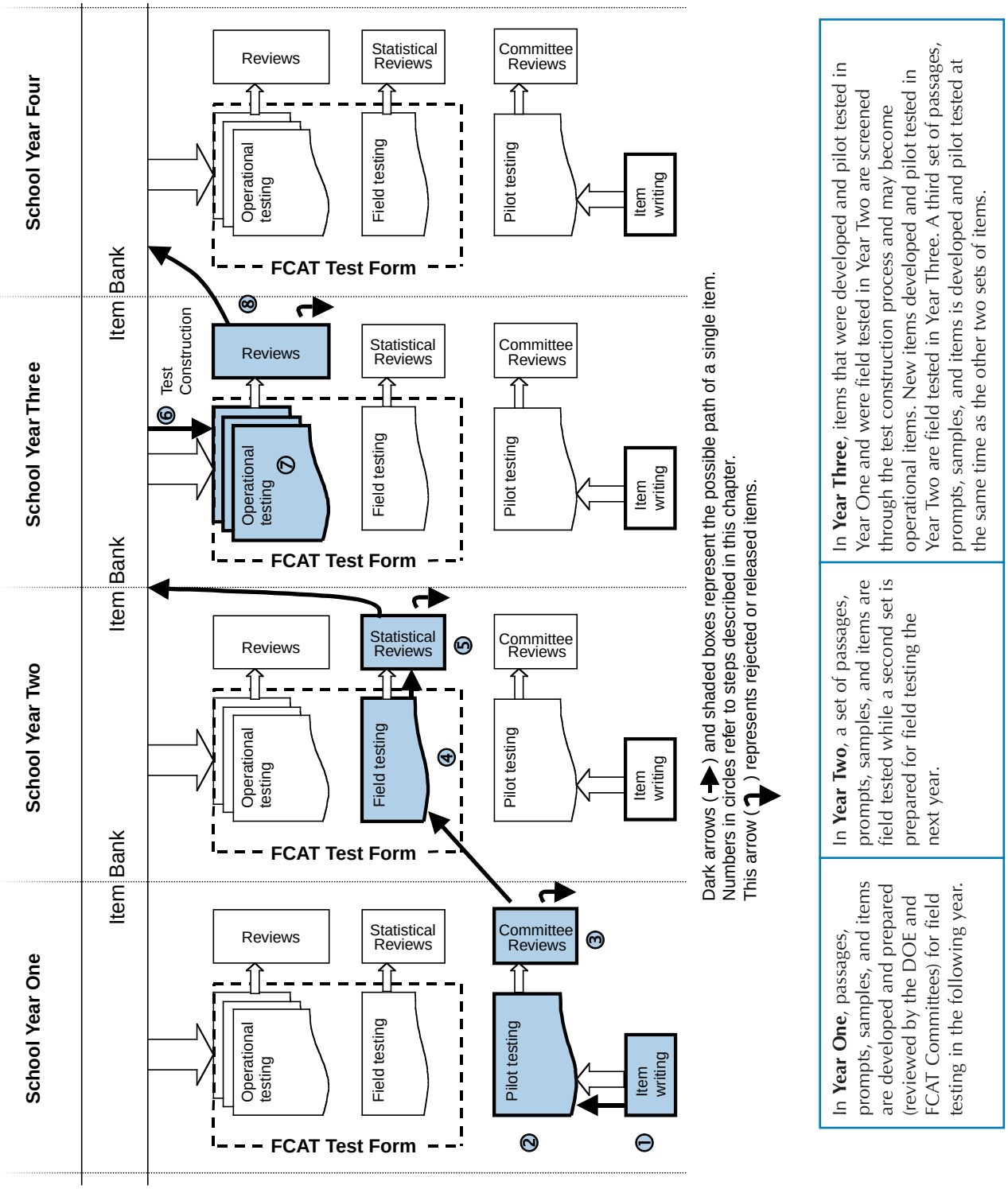
4.1 From Benchmark to Test Item: Developing an FCAT Item

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.

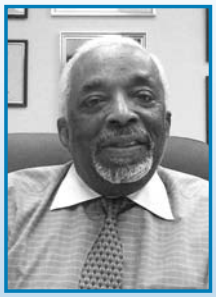
3. Committee Reviews

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.

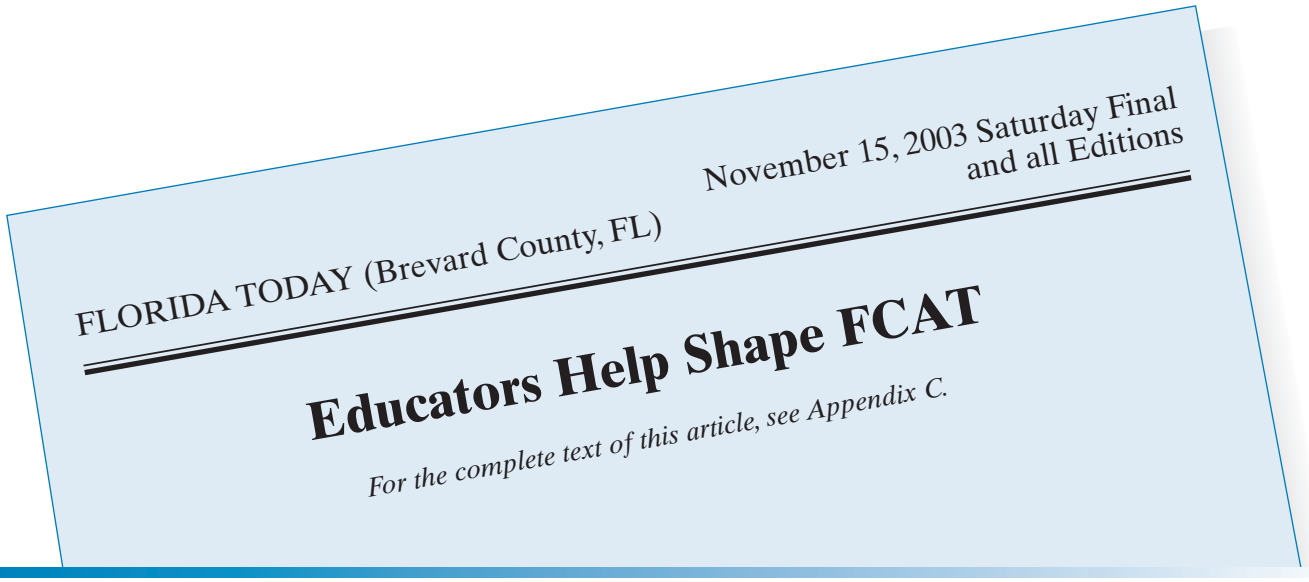
7. Operational Testing

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 Percent choosing each answer choice *Test Item Specifications*
Difficulty Level	Item Content Review Committees Prompt Review Committee Field test and operational test data analysis—*p*-values; IRT *b*-parameters
Item Discrimination (Item-Test Correlations)	*Test Construction Specifications* Field test and operational test data analysis—Item-total correlations; IRT *a*-parameters
Guessing	*Test Construction Specifications* Field test and operational test data analysis—IRT *c*-parameters
Freedom from Bias	*Test Construction Specifications* Bias Review Committees Pilot Test Results 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* 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.

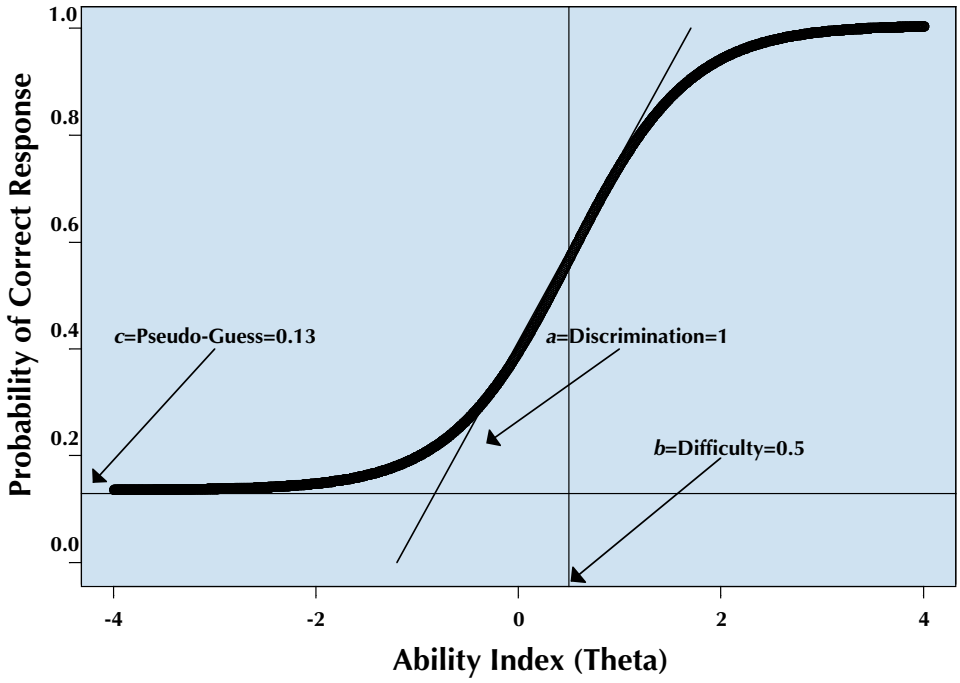
TABLE 12: CHARACTERISTICS OF THE TEST

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.

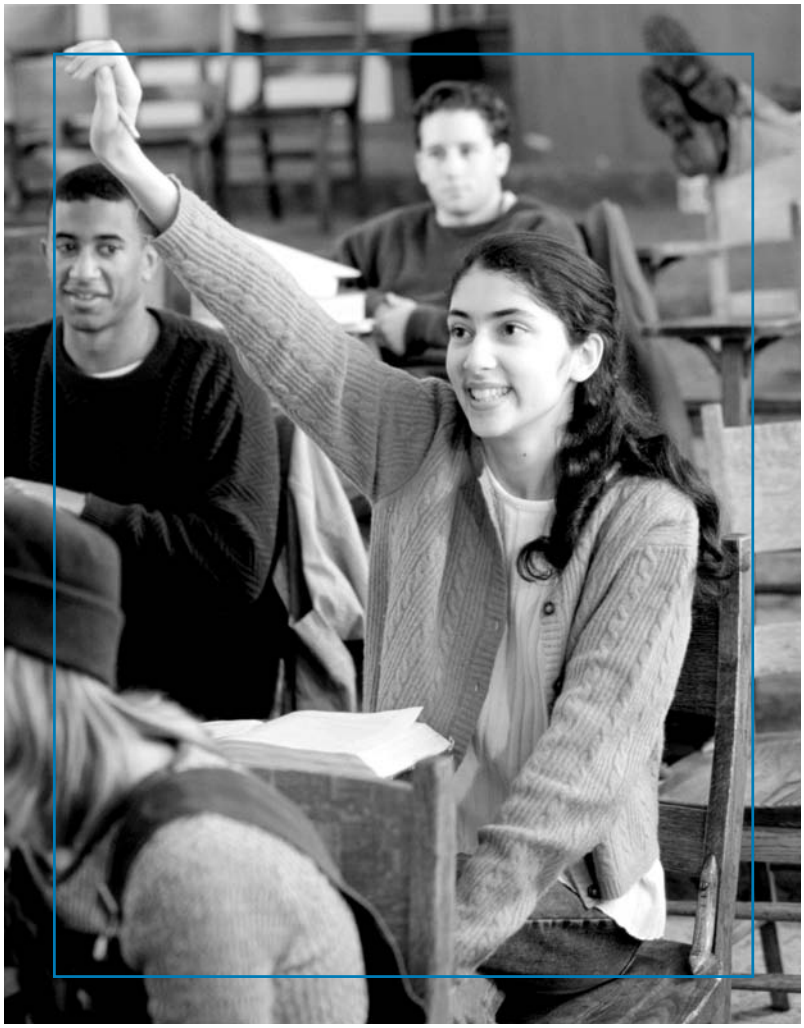
While IRT modeling of performance tasks is conceptually similar to that of multiple-choice items, performance tasks require a more complex mathematical treatment. In the end, however, modeling of a performance task includes the IRT parameters for each of the possible score points students can achieve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