FCAT 2.0

Florida Comprehensive Assessment Test®

Florida Statewide Assessments
2014 Technical Report

December 2014
Final Draft

Florida
EOC
Assessments

Acknowledgements

This technical report was produced on behalf of the Florida Department of Education. Requests for additional information concerning this technical report or the associated yearbook should be directed to Dr. Salih Binici at the Florida Department of Education (Salih.Binici@fldoe.org).

Major contributors to this technical report include the following staff from Pearson: Drs. Mark G. Robeck, Ahmet Turhan, Jie (Serena) Lin, Michael Clark, Robert Kirkpatrick, and the following staff from the Florida Department of Education: Dr. Salih Binici and Ms. Hülya Yürekli.

Chapter 1. Background

Technical Report and Yearbook

Increasing student achievement is a primary goal of any educational assessment program such as Florida's statewide assessment program, which includes the Florida Comprehensive Assessment Test 2.0 (FCAT 2.0) and the Florida End-of-Course (EOC) Assessments. This technical report and its associated yearbook have been produced in a way that can help educators understand the technical characteristics of the assessments used to measure student achievement.

In particular, this technical report provides information about the development and measurement characteristics of the Florida statewide assessments without overwhelming the reader with numerous tables that are often positioned within the explanation of the technical aspects of the program. To accomplish this goal, the report itself is organized into two parts:

- **Technical report: Chapters providing general information about the measurement process.** The text of the technical report outlines general information about the construction of the Florida statewide assessments, statistical analysis of the results, and the meaning of scores on these tests.

- **Yearbook: Yearly appendices providing the specific data related to a given year's test administration.** The appendices, organized as yearbooks, provide detailed statistics on the various assessments for a given academic year. Each year, a new yearbook will be added.

Florida's commitment to responsibly follow generally accepted professional standards is demonstrated by the inclusion of an annotated table of contents, which links the sections of this report to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999).

Overall, this report has been prepared for a broad audience including educators, parents, and the community at large. As such, the report constitutes a distillation of the information that is provided in greater detail in many technical documents (e.g., *Test Item Specifications*, *Test Construction Specifications*, *Field Test Sampling Specifications,* and *Calibration, Equating, and Scaling Specifications*) that guide the ongoing work of Pearson and the Florida Department of Education (FDOE). Nevertheless, certain sections of this report may require a special understanding of particular technical measurement issues and statistical procedures. Readers unfamiliar with such issues may contact FDOE, Office of Assessment, at (850) 245-0513, for further explanation. Additional information is also available at the FDOE website: http://www.fldoe.org/.

Historical Developments in Florida Statewide Testing

Since 1968, the Florida legislature has directed FDOE to develop assessment programs possessing technical qualities sufficient for having certain critical high-stakes educational decisions based on them. Over time, those decisions have included restricting student graduation and identifying schools needing improvement for not making adequate progress in maintaining student achievement. (See the *Assessment & Accountability Briefing Book* [FDOE, 2004, 2007] for more detailed chronologies of assessment and accountability in Florida.)

Within the current Florida statewide assessments program, today’s high-stakes decisions include the advancement of grade 3 and grade 10 students as well as the use of assessment scores in final course grades. In the rest of this chapter, some of the more recent developments in the Florida statewide assessments will be highlighted, including the transition from FCAT to FCAT 2.0 and the introduction of the EOC assessments. This brief background should establish the legislative and curricular framework for the technical analyses described in the remaining chapters of this report.

Recent Program Developments

Developments in 2007

The State Board of Education (SBE) accepted major revisions of the 1996 Sunshine State Standards for Reading/Language Arts and for Mathematics.

Developments in 2008

The 2008 legislation (Senate Bill [SB] 1908) authorizing the extension of the FCAT program and the introduction of EOC assessments was passed based on the experience gathered over four decades of high-stakes assessment implemented within Florida. One major goal of that legislation was to authorize the SBE to establish the Next Generation Sunshine State Standards (NGSSS) to replace the Sunshine State Standards. SB 1908 also removed the requirement that the statewide assessment program include norm-referenced tests.

The most transformative aspect of SB 1908 was that it allowed for the development and administration of EOC assessments, which were to be administered “within the last 2 weeks of a course.”

SB 1908 mandated revisions to what had been known as the FCAT Writing+ program, eliminating multiple-choice writing items beginning in 2009. Additional information about mandates for future changes in the writing program can be found at http://www.myfloridahouse.gov/Sections/Bills/billsdetail.aspx?BillId=38578.

Additionally, the SBE accepted a major revision of the 1996 Sunshine State Standards for Social Studies and for Science.

Beginning with the 2010–11 school year, SB 1908 required that the writing assessment be administered no earlier than the week of March 1 and that the comprehensive statewide assessments of any other subject be administered no earlier than the week of April 15.

Developments in 2009

In 2009, the revisions of the Sunshine State Standards approved by the SBE in 2007 and 2008 started to be referred to as the 2007 NGSSS and 2008 NGSSS, respectively.

Item development and review began for the FCAT 2.0 assessments in Reading/Language Arts and in Mathematics, based upon the 2007 NGSSS in those subject areas. Item development and review also began for the Algebra 1 EOC Assessment.

FDOE announced the need for budget reductions to the statewide assessment program’s services. All Summer FCAT Retake administrations were eliminated; performance tasks for FCAT Science were removed beginning with the 2010 administration; handscoring for FCAT Writing essays and the FCAT Reading and Mathematics performance tasks would be conducted by one reader rather than two readers, with 20 percent of all responses being scored by two readers for quality control; only one prompt would be administered at each grade level for FCAT Writing (previously two prompts were administered); and all annual FCAT support materials other than sample questions and answer key booklets would no longer be provided. It was also announced that performance tasks would not be part of the development of any of the new assessments aligned to the NGSSS.

Because of the switch from two readers to one reader for FCAT Writing, half-point scores, which were the average of the scores of both readers, were no longer possible.

A concordance study was conducted to ensure that the concordant scores used for graduation continued to be equivalent measures of the level of performance expected on the Grade 10 FCAT. As a result of this study, the required SAT Reading score was increased from 410 to 420, the required ACT Reading score was increased from 15 to 18, the required SAT Mathematics score was decreased from 370 to 340, and the required ACT Mathematics score of 15 remained the same. The new concordance score requirements were in effect for students scheduled to graduate in 2011 who had not already earned the previous passing scores by November 30, 2009.

Developments in 2010

Embedded field testing began for the FCAT 2.0 assessments in Reading/Language Arts and in Mathematics. A sample of students was drawn to take stand-alone field-test forms for the Algebra 1 EOC Assessment.

SB 4 amended the assessment window for the EOC assessments by striking the language “within the last 2 weeks” of the course and adding “during a 3-week period at the end” of the course.

Item development and review began for the FCAT 2.0 Science Assessment at grades 5 and 8, based upon the 2008 NGSSS for Science. Item development and review also began for the Geometry and Biology 1 EOC Assessments.

SB 4 also amended Section 1008.22 F.S. to require the implementation of EOC assessments at the high school level, which replace the FCAT Mathematics and Science assessments administered in grades 9 and 11, and to require the implementation of an EOC assessment in civics education at the middle school level. SB 4 mandated that the Algebra 1 EOC Assessment be administered beginning in the 2010–11 school year, the Geometry and Biology 1 EOC Assessments be administered beginning in the 2011–12 school year, the U.S. History EOC Assessment be administered beginning in the 2012–13 school year, and the Civics EOC Assessment be administered beginning in the 2013–14 school year. For the first year these EOC assessments are administered, the EOC results shall constitute 30 percent of a student’s course grade. With the exception of U.S. History, once standards are established for these EOC assessments, students must pass the assessment to earn course credit. SB 4 also authorized the Commissioner to establish an implementation schedule of EOC assessments in other subject areas, if feasible.

Developments in 2011

The first live administration of FCAT 2.0 Reading and Mathematics and the Algebra 1 EOC Assessment occurred during the spring administration window. Standard-setting meetings for these grade/subject/course combinations occurred with educators in September 2011.

A sample of students was drawn to take stand-alone field-test forms for the Geometry and Biology 1 EOC Assessments. FCAT 2.0 Science field-test items were embedded in grades 5 and 8 FCAT Science forms in preparation for the first operational administration of FCAT 2.0 Science in those grades in 2011.

FDOE implemented a data forensics program beginning with the spring 2011 administration. The purpose of the program is to analyze data to identify highly unusual results. For the first year of implementation, student tests with extremely similar responses and schools with improbable levels of erasures were held by FDOE pending further investigation and appeals by school districts.

Developments in 2012

The first live administration of FCAT 2.0 Science assessments and the Geometry and Biology 1 EOC Assessments occurred during the spring administration window. Standard-setting meetings for these grade/subject/course combinations occurred with educators in September 2012. In addition, grades 6 and 10 FCAT 2.0 Reading were administered online the first time for all but a small percentage of students.

A sample of students taking United States History or United States History Honors was drawn to take stand-alone field-test forms for the U.S. History EOC Assessment.

For FCAT 2.0 Writing, in addition to the elements of focus, organization, support, and conventions described in the rubrics, the scoring decisions included expanded expectations regarding the following: (1) increased attention to the correct use of standard English conventions and (2) increased attention to the quality of details, requiring use of relevant, logical, and plausible support, rather than contrived statistical claims or unsubstantiated generalities.

Developments in 2013

The first live administration of the U.S. History EOC Assessment occurred during the spring administration window. Standard-setting meetings for this course occurred with educators in August 2013. In addition, grades 7 and 9 FCAT 2.0 Reading and Grade 5 Mathematics were administered online the first time for all but a small percentage of students.

A sample of students taking Civics was drawn to take stand-alone field-test forms for the Civics EOC Assessment. Also a sample of state representative students was administered the writing prompt field test.

There was a small policy change regarding grade 8 Mathematics test administration in which the districts were given the option to not administer the grade 8 Mathematics test if the students would be taking the Algebra 1 EOC test. As a result, approximately 7-8 % of the 8th grade students did not take the Mathematics test.

Developments in 2014

The first live administration of the Civics EOC Assessment occurred during the spring 2014 administration window. Standard-setting meetings for this course occurred with educators in July 2014. In addition, grade 8 FCAT 2.0 Reading and grade 6 Mathematics were administered online for the first time for all but a small percentage of students.

A major milestone was reached in 2014. FCAT 2.0 Reading and Mathematics and Algebra 1 and Geometry EOC assessments were discontinued for a new assessment program called the Florida Standards Assessment program. In 2015 only FCAT 2.0 Science (in grades 5 and 8), Biology 1, U.S. History, and Civics EOCs will be administered and managed by Pearson. No writing field testing will take place in 2015.

The policy change regarding the grade 8 Mathematics test administration, in which districts were given the option to not administer the grade 8 Mathematics test if the students would be taking the Algebra 1 EOC test, was continued. As a result, approximately 16–17% of the grade 8 students did not take the Mathematics test.

Participants in the Development and Analysis of the Florida Statewide Assessments

FDOE manages the Florida statewide assessment program with the assistance of several participants, including multiple offices within FDOE, Florida educators, a Technical Advisory Committee (TAC), and vendors. FDOE fulfills the diverse requirements of implementing Florida's statewide assessments while meeting or exceeding the guidelines established in the *Standards for Educational and Psychological Measurement.*

Florida Department of Education

Office of Assessment. The Office of Assessment oversees all aspects of Florida's statewide assessment program, including coordination with other FDOE offices, Florida public schools, and vendors.

Test Development Center. Funded by FDOE, the Test Development Center (TDC) works with Florida educators and vendors to develop test specifications and test content and to build test forms.

Florida Educators

Florida educators participate in most aspects of the conceptualization and development of Florida assessments. Educators participate in the development of the academic standards, the clarification of how these standards will be assessed, the test design, and the review of test questions and passages.

National Technical Advisory Committee

Though the exact number of meetings is not fixed, FDOE typically convenes a panel twice a year to discuss psychometric, test development, administrative, and policy issues of relevance to current and future Florida testing. This committee is composed of several nationally recognized assessment experts and highly experienced practitioners from multiple Florida school districts.

Assessment and Accountability Advisory Committee

The Assessment and Accountability Advisory Committee has approximately 15–20 members representing educators, school district personnel, and university faculty. The members of this committee advise FDOE about K–12 assessment and accountability policies. Their recommendations may relate to standards for state assessment achievement levels, school grading policies, and alternative assessments. This committee meets at least once a year.

Center for Advancement of Learning and Assessment—Florida State University

Members of the Center for Advancement of Learning and Assessment at Florida State University (CALA-FSU) conduct an independent, third party review of FCAT 2.0 and EOC assessment results.

Pearson

Pearson is the vendor responsible for developing test content, building test forms, conducting psychometric analyses, administering and scoring test forms, and reporting test results for the Florida assessments described in this report. Starting in 2010, Pearson became the primary party responsible for executing psychometric operations for the Florida statewide assessments.

Human Resources Research Organization

Human Resources Research Organization (HumRRO) has provided program evaluation to a wide variety of federal and state agencies as well as corporate and non-profit organizations and foundations. For the Florida statewide assessments, HumRRO conducts independent checks on the equating and linking activities and reports its findings directly to FDOE. HumRRO also provides consultative services to FDOE on psychometric matters.

Buros Institute of Mental Measurements

Buros Institute of Mental Measurements (Buros) provides professional assistance, expertise, and information to users of commercially published tests. For the 2014 Florida statewide assessments, Buros provided independent operational checks on the equating procedures of FCAT 2.0 and EOC assessments, on-site monitoring of writing handscoring activities, and the scanning and editing services provided by Pearson.

Caveon Test Security

Caveon Test Security analyzes data for the FCAT, FCAT 2.0, and EOC assessments using Caveon Data Forensics^{TM} to identify highly unusual test results for two primary groups: (1) students with extremely similar test scores; and (2) schools with improbable levels of similarity, gains, and/or erasures.

Chapter 2. Development

The test development phase of each Florida assessment includes a number of activities designed to produce high quality assessment instruments that accurately measure the achievement of students with respect to the knowledge and skills specified in Florida's academic standards. Those standards are intended to guide instruction for students throughout the state. Tests are developed according to the content outlined in the Next Generation Sunshine State Standards (NGSSS) at each grade level for each tested subject area.

The revision of the Sunshine State Standards for reading/language arts and mathematics began in 2006 and were adopted by the State Board of Education (SBE) in January and September of 2007, respectively. The revisions of the standards for science and social studies were made in 2007 and adopted by the SBE in February and December of 2008, respectively. The educational reform measure (SB 1908) became law in 2008, after which the reading/language arts and mathematics standards were renamed the 2007 NGSSS and the social studies and science standards were renamed the 2008 NGSSS. Florida's FCAT 2.0 and EOC assessments are developed based upon the 2007 NGSSS and 2008 NGSSS.

Test Development Process

The following steps summarize the process FDOE uses to develop Florida's FCAT 2.0 and EOC assessments. These steps meet or exceed industry standards for developing large-scale, criterion-referenced assessments.

Development of Test Specifications. Committees of content specialists develop test specifications that outline the requirements of the test, such as eligible test content, item types and formats, and content limits and cognitive levels for items. Information about the content, level of expectation, and structure of the tests is based on judgments made by Florida educators. In addition, empirical information obtained from past Florida testing experience is used to inform decisions about topics such as the number of items needed to report reliable subscores. These specifications are published as a guide to the assessment program. Committees provide advice on test models and methods to align the tests with instruction.

Development of Items and Stimuli. Using the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) and test item specifications, Florida's Test Development Center (TDC) works with Pearson to develop items and stimuli. Please note that throughout the item development section, content created for Florida tests is broadly referred to as "items." When used in this broader sense, the term "items" also refers to other types of content, such as passages, art, and stimuli.

Item Review. All members of the assessment team review the developed items, discuss possible revisions, and make changes when necessary. Formal item review procedures include internal TDC review, Florida educator content review, bias review, and sensitivity review, which includes Florida educators and community leaders.

Field Testing. New items are field tested to empirically validate their effectiveness before they are included on an operational test. Field testing of items can occur in special studies or as part of regular administrations.

Test Construction. FDOE and TDC build operational tests according to blueprints and publicly available test specifications. The selection and sequencing of items and stimuli for operational tests are made using psychometric principles that are recognized in the field as best practices.

Test Specifications

Test specifications for Florida tests are developed through a process that includes educator committee review. See *Test Item Specifications*, which are available at: http://fcat.fldoe.org/fcat2/itemspecs.asp. Each document describes the process used for developing content appearing on Florida tests as well as content specifications for items and stimuli (such as reading passages) appearing on each test. In particular, each specification book includes sections on cognitive complexity, limits for particular benchmarks, as well as sample items, a glossary, and a sample item review judgment form. The test item specification books for reading, mathematics, science, and social studies also include clarifications of the benchmarks.

Item Development

Item Writers

Pearson identifies item writers who meet or exceed FDOE's criteria. Each prospective item writer must have three years of experience as a classroom teacher or two years experience as an item writer. As a matter of policy, FDOE rules prohibit teachers currently employed in Florida schools from participating in the item writing process, though they play a significant role in the review of the items once they have been developed. In addition, Pearson's item development team, prompted by TDC, has made a concerted effort to increase the size of the pool of qualified item writers in order to enhance the development of more items written from a fresh perspective.

Once a pool of prospective item writers is chosen by Pearson's development staff, a résumé of each item-writer candidate is submitted for FDOE review and approval.

Item Writer Training

At the beginning of the item development cycle, Pearson and TDC staff conduct training sessions for the approved item writers to provide information that is essential for new writers as well as information updates for veteran writers. The writer training includes a comprehensive overview of Florida's standards and benchmarks; Florida's cognitive level classification system; the requirements specified in the test item specifications document; sample formats, items, and prompts; bias and sensitivity considerations; and additional guidelines requisite for developing quality assessments.

The item writer training sessions for each content area are held at one of Pearson's facilities and are co-led by a Pearson content lead and a TDC content lead. Additional Pearson content specialists make up the full complement of training staff for each session.

The item writer training familiarizes the item writers with the layout and hierarchy of the NGSSS. For example, for each grade of FCAT 2.0, the mathematics standards include three "big" ideas and up to five supporting ideas. Big ideas are the main concepts; supporting ideas contain concepts that help students master the big ideas. Each big or supporting idea contains multiple benchmarks.

In the training sessions, the item writers learn about the content limits imposed by each grade level's standards. They are reminded that at particular grades there are restrictions concerning the use of certain item types (e.g., multiple-choice [MC], gridded-response [GR], and fill-in response [FR] items) and the circumstances under which they are to be used. Additionally, the item writers are advised that items requiring the use of rulers and/or calculators are restricted to particular grades and content area assessments (e.g., calculators in grade 7 mathematics and above).

During the course of the training, the content leads explain the importance of writing items that have different levels of cognitive complexity and how an item's complexity differs from its difficulty. Item writers must have mastery of these concepts if they are to meet the development targets of content staff with respect to the number of items needed for each benchmark at a particular level of cognitive complexity or estimated item difficulty.

Much of the remainder of the training comprises the practical guidelines that each writer can use to avoid introducing construct-irrelevant factors into test items. For example: Are the item stems concise? Is there parallel construction among the answer options? Is the readability of the item stem an obstacle to assessing a science or a mathematics benchmark? Does the item stem clue the answer for this question? Is the item free from economic, regional, cultural, and/or ethnic bias? These best practices are summarized in checklists that item writers are to use prior to submitting items to content staff.

During the training, there are breakout sessions where item writers work under the watchful eye of content specialists. In these sessions, the item writers practice writing and editing items to develop a sense of how the item writing guidelines are best applied. They also learn about the proper use of Pearson's item development software, including the steps for submitting associated art, graphs, and equations.

At the conclusion of the item writer training, the item writers are assessed in various ways on what they have learned, after which they are approved to develop items. These assessments may entail ungraded quizzes that are reviewed with the prospective item writers or a mock item review, where the content specialists highlight the common challenges that arise during the item review process.

Item Review

Pearson Review

After the item writers receive orders for new items, they write them to specification and submit them for review. That review entails the following steps:

1. Accept/Reject/Revise decision is made by the content specialist.
    - Accepted items are coded and edited for content by the content specialist.
    - “To be revised” items are sent to the item writer for revision.
2. Art is ordered on accepted items.
3. Art is reviewed by the content specialist.
4. Pearson copy editors edit the items until the items are approved by the content specialist.
    - Universal design guidelines are used for checking item content and art.
    - All items are reviewed by copy editors for adherence to Florida’s style guidelines.
5. Senior item review is conducted by the content lead at Pearson.
6. Approved items are marked for TDC review.

Throughout the period of item acquisition, the development staff may contact item writers, as needed, to coach them on item revisions. In these instances, the staff relates to the item writers the staff’s understanding, gathered in item committee reviews, about views of Florida educators and community members on the appropriateness of the items and passages for all Florida students.

A Note about Passages

For FCAT 2.0 Reading, Pearson’s staff members select potential passages from various literature sources in the public domain or from commissioned writers who create new passages. The passages that are within the FDOE guidelines are prepared for presentation to TDC reading content staff members. The staff members then determine which passages may be considered for passage review.

A Note about Prompts

The TDC content specialists, in coordination with FDOE, develop and review prompts for potential use on the FCAT 2.0 Writing assessment. The prompts are selected so that the subject matter is appropriate for students in grades 4, 8, and 10. In addition, prompts are reviewed for offensive or biased language relating to religion, gender, and racial or ethnic backgrounds. Prompts are field tested on approximately 3,000 students statewide. The student responses are then scored, and the statistical review of these items is conducted as an initial step in the operational prompt selection process. Before being selected for inclusion as an operational prompt on FCAT 2.0 Writing, the field-test statistics for the item must meet established prompt specifications and psychometric criteria. Following prompt selection, TDC and Pearson content specialists meet with Florida educators to score a sample of student responses from the field test that illustrate the range of quality allowed within each score point on the rubric. Responses are subsequently selected from the scored sample to create sets of materials used to train FCAT Writing scorers.

TDC Review

At each step during the development of items for the Florida statewide assessments, TDC requires that item information be recorded on a template. The information included on the item template provides descriptive elements included in FDOE's item banking system. TDC requires that the complete item template, source information, a copy of the source document, and reading passage copyright status be provided before proposed items are reviewed. Specialized information is also required for the different item types. Multiple-choice test items must include the item stem and stimulus, plausible answer choices, and correct answer. Mathematics and science multiple-choice items also require plausible distractors and a descriptive rationale for each distractor. Gridded/fill-in response test items must include the item stem, a set of correct answers/numerical ranges, type of grid to be used, and guidelines for evaluating correct responses, e.g., rounding criteria for evaluating student responses. For graphics, any color requirements must be noted in the template.

Upon receipt of the items, the TDC content specialist will review them and mark them in one of the following ways:

- Accept (A)
- Accept with metadata change(s) (AM)
- Accept with a designated revision (AR)
- Revise and resubmit (RR)
- Reject (R)
- Move grade (in cases where the item is rejected for the grade submitted but accepted for another grade in which it fits the content specifications) (MG)

Once items are approved by the TDC content specialist, Pearson's content specialist prepares the TDC-approved items for review by Florida educators. Items approved for review by Florida educators are assembled into item review books and prepared for secure shipping to the site of the item review meetings in Florida.

Florida Educator Reviews

FDOE and TDC, in conjunction with Pearson staff, conduct numerous meetings each year with Florida educators. The purpose of an educator meeting is to receive feedback on the quality, accuracy, alignment, and appropriateness of the passages, prompts, and test items that are developed annually for Florida statewide assessments. Item review and content advisory committees are composed of Florida educators. The bias and community sensitivity review committees and the science and social studies expert review committees are composed of educators, university professors, and other Florida citizens selected by the TDC staff. The meetings are held at various venues throughout Florida. Typically, the social studies expert reviewers are part of the item review committees. For science, the expert reviewers typically do not attend the item review meetings but instead attend a one- or two-day expert review meeting.

Item Bias and Community Sensitivity Reviews

Bias committee members are represented by both genders and are from a variety of multicultural backgrounds. Bias committee members are asked to consider and identify passages or items that in some way inadvertently demonstrate bias (e.g., racial, ethnic, gender, geographic, etc.). After a brief training session conducted by a Pearson and an FDOE representative, committee members evaluate items and passages and provide feedback if they believe any of the content may be biased. These data inform the decisions made about the suitability of the passages and items for placement on future tests. The bias committee may also review items that received statistical flags for group comparisons during field testing. The committee fills out a sheet stating if and when flagged items can be used in test construction. Typically, the determination is “do not use,” though when other aspects of the item statistics are favorable, the committee can articulate the conditions for its use, if deemed necessary by the TDC and FDOE staff. This bias review serves as an extra measure, ensuring that the items do not contain bias and are appropriate for use on future tests.

Community sensitivity committee members, representing various communities throughout Florida, are asked to consider passages and items for issues of sensitivity with respect to the wide range of cultural, regional, philosophical, and religious backgrounds of students throughout Florida. After a brief training session conducted by a TDC representative, committee members evaluate items and passages and provide feedback if they believe any of the content may raise sensitivity issues. These data inform the decisions made about the suitability of passages and items for placement on future tests.

During community sensitivity reviews, multiple copies of each item review book are available to allow the simultaneous review of any given book by up to five reviewers. Pearson and TDC pre-assign the books to each committee member. Each book of items and passages must receive a minimum number of reads, which is determined by TDC based on the volume of material and the number of committee members. The reading assignments for each reviewer are organized so that each set is reviewed by a demographically representative sample of the committee members. Pearson and TDC staff monitor the committee’s progress throughout each meeting day. As the meeting progresses, adjustments may be made to assignments to account for the speed at which individual committee members read the material.

Committee members sign out their review books, review the content according to the training, and identify concerns on the response area of review feedback forms. Feedback forms with comments are transcribed throughout the day into an electronic file by TDC staff. Upon completion of the review assignments, reviewers sign in and return their review books. Reviewers also complete an affidavit indicating which sets of passages and items they reviewed.

Item/Passage Review

The purpose of the item/passage review committee meeting is to review and evaluate newly developed test items/passages to ensure that they align to the NGSSS, are grade appropriate, and have accurate content.

For each item review meeting, a member of Pearson's staff is present to keep an electronic record of decisions made and to document any changes requested to item stems, options, art, or metadata. This electronic record is reconciled daily with the written record kept by the TDC staff member in charge of facilitating the meeting. Items may be marked as accepted (A), accepted with metadata changes (AM), accepted with revisions (AR), revise and resubmit (RR), rejected (R), or move grade (MG). A complete description of each metric type is available in the Item Development Plan (IDP) document developed for each particular assessment.

Committee members review the items/passages and discuss their feedback with the TDC facilitator and other committee members. Typically, each committee member is assigned a specific task from a rating form, which is used to document information about each item.

Expert Review Meeting

After the science item review meetings each year, an expert review meeting is held where science experts, representing Florida universities and science research institutions, review science items approved by content review committees. The purpose of this additional review is to confirm scientific accuracy and appropriateness. After a brief training session conducted by a TDC representative, committee members evaluate items provided to them in binders. Expert reviewers for social studies are embedded in the social studies item review meeting, so a separate social studies expert review meeting is not held.

To ensure that the items in each binder are reviewed by two different members, Pearson uses a sign-out/sign-in sheet to track the review. Any inaccuracies or concerns identified by committee members are noted directly on the item and compiled by TDC representatives. Items that have inaccuracies or concerns are later addressed and corrected by Pearson and TDC staff, or the item is rejected. An electronic file of reviewer comments is compiled, organized, and reviewed by TDC staff and shared with Pearson staff prior to the first round of test composition.

Field Testing

Once a newly constructed item has survived committee reviews, it is field tested. For example, in a particular grade's reading administration there might be 40 different forms containing the same operational test items. However, each form would also contain one or more unique field-test reading passages and corresponding unique field-test items. The field-test items do not count toward an individual student's score.

Embedded Field Testing

Newly developed items and prompts are embedded into the FCAT 2.0 and EOC forms for the purpose of field testing. The positioning of these embedded field-test items and prompts is carefully considered by Pearson and FDOE psychometricians. These positions affect not only the layout of the concurrent core assessment form but also the relative positioning of these items and passages, if they survive data review, on future live core forms.

Approximately 5,000 students are sampled for each field-test form. Field-test forms for FCAT 2.0 and EOC assessments are spiraled at the student level throughout the state along with the linking forms. This spiraling method helps to minimize the sampling error (Phillips, 2011).

Stand-alone Field Testing

Stand-alone field testing is implemented whenever a new testing program is being established, as it is in this case for the EOC assessments. The Algebra 1 EOC Assessment had an online stand-alone field test in 2010, and the Biology 1 and Geometry EOC Assessments had online stand-alone field tests in 2011. These stand-alone field tests were conducted in preparation for three new online EOC assessments that became operational in 2011 and 2012, respectively. The U.S. History EOC Assessment had an online stand-alone field test in 2012 prior to going operational in 2013. The Civics EOC Assessment's online stand-alone field test is set for 2013 prior to going operational online in 2014.

Writing uses a stand-alone format because it is not practical to field test new prompts in an embedded field-test format. The time necessary for a student to respond to any given writing prompt makes the stand-alone (one prompt per student) format more conducive to gathering accurate and reliable field-test data.

The following seven steps are followed during the selection of field-test samples.

1. The EOC and FCAT 2.0 Writing sampling uses the most recent relevant State Student Results files, which FDOE provides to Pearson. These data are combined with the current enrollment information for targeted EOC courses. However, the timing of the FCAT 2.0 Writing field test does not allow for the inclusion of current enrollment data.

2. The target N-count for each form is 4,000 students for each EOC online stand-alone field test and 3,000 for each FCAT 2.0 Writing stand-alone field-test prompt.
3. The EOC online field-test forms and the FCAT 2.0 Writing prompts are spiraled within selected schools. For FCAT 2.0 Writing, one sample is selected for each participating grade.
4. The five Florida district regions are combined into three regions (northern, central, and southern) for the sake of the sampling plan.
5. A stratified proportional sampling method is used with region and school size as the stratification variables. In each region, for EOC assessments, schools are rank-ordered based on the number of students taking the course (eligible students) and divided into four groups, each with the same number of schools. For Writing, the schools are rank-ordered and divided into groups using the Writing State Student Results files from the most recent administration.
6. The school is used as the sampling unit.
7. The ethnic and gender distributions of students, the content curriculum group (i.e., standard curriculum, Exceptional Student Education [ESE], and English Language Learner [ELL]), and their scale scores in the related subject are used to evaluate the representativeness of the selected samples. For FCAT 2.0 Writing, only standard curriculum students are sampled, but all students within a school take the test and everyone is hand-scored.

Item Statistics

Subsequent to the field testing of items on the Florida statewide assessments, the items are reviewed using statistics based on classical test theory (CTT), item response theory (IRT), and differential item functioning (DIF). The following is a brief description of these statistics.

Classical Statistics

Classical Item Difficulty. The difficulty of an item is commonly expressed as a *p*-value, which is the mean score on an item. For items in the multiple-choice, gridded-response, and fill-in response categories, where the maximum item score is 1, the mean score is equivalent to the proportion of examinees answering the item correctly.

Classical Item Discrimination. The correlation between scores on an item and scores on the entire test is typically used as an item discrimination index. The principle underlying this statistic is that a student's performance on any single item should be positively related to the student's overall performance on the test. That is, an item "discriminates" well when a student who does well on the entire test also does well on the given item, and vice versa. On the other hand, an item discriminates poorly if a student who does well on the entire test performs at random on the given item.

For Florida assessments, three types of item-total correlations are computed. First, the item-total correlations are computed as Pearson correlations. For the MC, GR, and FR items, these correlations are equivalent to point-biserial correlations between the dichotomous variable (right and wrong) and the total score.

Second, to eliminate the influence of the item score on the total test score, a *corrected* point-biserial correlation is calculated. For a designated item, the corrected point-biserial is the correlation between the item score and the total test score with that item removed.

Last, the biserial item-total correlation may be understood as an estimate of the correlation that would have been obtained if the dichotomous item had actually been a normally distributed continuous measure.

The relationship between point-biserial and biserial correlations is:

$$r_{bis} = \frac{\sqrt{p(1-p)}}{Y} r_{pbis} \qquad (2.1)$$

where *Y* is the y ordinate of the standard normal curve at the *z*-score associated with the *p*-value for this item. Because the y ordinate on a normal curve is always less than $\sqrt{p(1-p)}$, biserial correlations will generally be larger than the corresponding point-biserial correlations. In fact, if the total score on the test is not normally distributed, the biserial correlation can nonsensically exceed 1 (Cohen & Cohen, 1975).

In test construction, the corrected point-biserial correlations were used. Preferably, the correlation of any item used on a test should be greater than 0.25. It is also recommended that the distractor correlations should not be positive for MC items.

Item Response Theory Statistics: Difficulty, Discrimination, and Guessing

Florida’s FCAT 2.0 and EOC assessments use IRT for scaling and equating. Two different IRT models are used: a three-parameter logistic (3PL; Lord & Novick, 1968) for MC items and a two-parameter logistic (2PL; Lord & Novick, 1968) for GR items in FCAT 2.0 and FR items in the EOC. Several statistics, known as parameters, are used to describe the characteristics of items in IRT. (Details about the 2PL and 3PL can be found in “Chapter 6. Scaling.”)

IRT Difficulty. IRT item difficulty using the 2PL model is defined as the level of ability (i.e., knowledge and skills) where the probability of a correct response is 0.5. For the 3PL model, the item difficulty is defined as the level of ability where the probability of a correct response is halfway between IRT lower asymptote (see definition below for details) and 1.0. Selected items should have scaled *b* parameters in a specific range; however, it is recommended that they be concentrated around performance-level cut points because an item tends to bring the most information (i.e., the least measurement error) at scale score points around its *b* parameter. Items with high *a*-parameters are preferred.

IRT Discrimination. The item discrimination in IRT is defined as the item's ability to differentiate between lower and higher performing examinees and is represented by the *a* parameter, which is a function of the slope of the item characteristic curve. The preferred ranges for this parameter are discussed in "Guidelines for Item Selection" later in the chapter.

IRT Lower Asymptote. The lower asymptote parameter for MC items is conceptualized in IRT as the probability of examinees with extremely low ability levels getting a correct answer and is represented by the *c* parameter. This statistic is sometimes referred to as the pseudo-guessing parameter as examinees with low ability levels might only be able to respond correctly by guessing. However, it is feasible that for some benchmarks instructional emphases can lead to widespread mastery. Items from these benchmarks may result in lower asymptotes that are relatively high.

Summaries of classical and IRT statistics for FCAT 2.0 and EOC assessments can be found in the summary statistics reports of the yearbook.

Differential Item Functioning (DIF) Statistics

DIF statistics are used to identify items that result in differential performance from students of comparable ability who belong to one of two contrasting groups (either the focal group or the reference group). For Florida's statewide assessments, three comparisons are made for each item, as listed below. The focal group is given first in Table 2-1.

Table 2-1. DIF Analysis Groups for Florida Statewide Assessments

Focal Group		Reference Group
African-Americans (AA)	vs.	Whites (W)
Hispanics (H)	vs.	Whites (W)
Females (F)	vs.	Males (M)

In the DIF analyses, the total raw score on the core items is used as the ability-matching variable. Following Dorans and Holland (1993), Mantel-Haenszel Delta DIF statistics are computed:

$$MHD = -2.35\ln(\hat{\theta}_{MH}), \qquad (2.2)$$

where $\ln(\hat{\theta}_{MH})$ is the log-odds ratio. The odds ratio, $\hat{\theta}_{MH}$, is computed with the following formula:

$$\hat{\theta}_{MH} = \frac{\sum_j \left[\frac{A_j D_j}{T_j} \right]}{\sum_j \left[\frac{B_j C_j}{T_j} \right]}, \qquad (2.3)$$

where *j* indicates the number of matching criterion for a given item and the counts of correct (A_j or C_j) and incorrect (B_j or D_j) responses for the focal and reference comparison groups, respectively.

For the MC, GR, or FR items, the Mantel-Haenszel Delta DIF statistics are computed to classify test items as one of three levels of DIF: negligible DIF (A), moderate DIF (B), and large DIF (C) for each focal/reference comparison. An item is flagged if it exhibits the B-level or the C-level of DIF, using the ETS classification rules (Dorans & Holland, 1993):

Level A: MH delta (MHD) not significantly different from 0 (based on alpha=.05) or |MHD| < 1.0.
Level C: |MHD| ≥ 1.5 and significantly different from 1.0.
Level B: All other cases not encompassed by Level A or Level C.

During test construction of Florida's statewide assessments, these MHD flagging rules are used to analyze DIF in the following specific ways:

- Only items flagged A should be selected for core and anchor items.
- Any items flagged B are reviewed by TDC and Pearson content leads to determine whether there is any performance difference between the comparison groups.
- Any items flagged C should rarely be selected for operational or anchor item purposes. Those rare instances would include cases in which both the content balance could only be met with the use of that flagged item and other psychometric indices of that item are satisfactory.

Beginning in 2010, new items flagged with a B are referred to the bias committee described above for additional bias review. Items determined to be flawed by this committee have been marked as "do not use" in the item bank. Summaries of DIF statistics for FCAT 2.0 and EOC assessments can be found in the summary statistics reports of the yearbook.

Item Bank

The item bank for Florida's statewide assessments is a database that contains test item images and the accompanying artwork. The item bank includes all the items developed for Florida, whether field tested or not. This system allows test items to be readily available to the FDOE for test construction and reference and to Pearson for test booklet design and printing.

In addition, a statistical item bank is maintained for the storage of item data such as unique item number (UIN), grade level, subject, benchmark/instructional target measured, dates the item has been administered, and item statistics. The statistical item bank also warehouses information obtained during the data review committee meetings indicating whether a test item is acceptable for use, acceptable with reservations, or not acceptable at all. During the test construction process, FDOE and Pearson use the item statistics to calculate and adjust for differential test difficulty and to check and adjust the test forms for content coverage and balance. The item bank is also used for reviewing or printing individual item statistics as needed.

Test Construction

Form Design

FCAT 2.0 and EOC assessment operational test forms include core items, anchor items, and embedded field-test items. Anchor items are included for the purposes of linking test forms together. The anchor items serve one of two purposes on Florida's statewide assessments: external anchoring or internal anchoring. External anchor items are used for statistical analyses; however, they do not directly contribute to a particular student's raw score. Internal anchor items *do* contribute to a student's raw score and, after scaling, to his or her final scale score.

Guidelines for Item Selection

Each FCAT 2.0 or EOC assessment form is constructed by strictly following the content and psychometric guidelines, which are defined by the content and psychometric teams from Pearson and Florida.

Content-level considerations during item selection

- Items are chosen so that a given test meets the provided test blueprint,
- Items are chosen to balance content domains and standards,
- Keys of 1–4 are similarly represented on the entire test,
- Selected items address a variety of topics within an objective (no clones),
- Identified key is correct,
- Each item has only one correct response,
- No duplicate operational items appear within the form,
- Identified item content classification is correct,
- Reporting categories are represented accurately,
- Reporting categories and benchmarks reflected on the item cards match test blueprint,
- No clueing/clangs exist among items,
- Items are free from typographical, spelling, punctuation, or grammatical errors,
- Items meet style specifications (with respect to bolding, italics, etc.),
- No “do-not-use”(DNU) items are selected for the forms,

- The correct version of the item is used, and
- If there are multiple forms, the forms are comparable to each other, and each form has the same reporting category distributions.

There are numerous psychometric requirements listed in detail in the test construction specifications for each Florida assessment. Below are a few of the psychometric considerations related to item statistics and DIF.

Psychometric considerations

- A reasonably wide range of item difficulties (as specified below) are represented,
- *p*-values for MC items are reasonable and most are between 0.25 and 0.90,
- *p*-values for GR/FR items are reasonable and most are between 0.11 and 0.90,
- Corrected point-biserial correlation of each item is preferably equal to or greater than 0.25,
- No items with negative corrected point-biserial values are selected,
- *a* parameters for all items are reasonable and most are between 0.50 and 3.50,
- *b* parameters for all items are reasonable and most are between -2.0 and 3.0,
- *c* parameters for MC items are smaller than 0.40,
- There are only a few items with model fit flags, and
- Very few items on the test have DIF “B” flags and none DIF “C” flags from field-testing analysis, preferably.

Application of Item Response Theory

Test construction for Florida’s FCAT 2.0 and EOC assessments utilizes *test characteristic curves* (TCCs), *test information curves* (TICs), and *conditional standard error of measurement* (CSEM) curves for both anchor and core forms. For each test constructed for a given grade and subject, a set of targets is developed in the form of target TCCs, TICs, and CSEM curves. The TCC, TIC, and CSEM curves of the newly built form are always compared to the target curves. If the curves are not aligned, the newly constructed form will be modified until the alignment is satisfactory. For that reason, it may be necessary to carry out this procedure several times for any given form.

Test Characteristic Curves (TCCs)

A useful feature of the IRT is that the TCC (see Figure 2-1 for an example) can be constructed as the sum of item characteristic curves (ICCs) for the items included in the test. The TCC can be used to determine examinees’ raw scores (or percent-correct scores) that are expected at given ability levels. When two tests are developed to measure the same ability, their scores can be conveniently equated through the use of TCCs (the true score equating method). The target TCCs help to build a test form that is similar to the historical ones so that the equating can be performed smoothly. Thus, it is desirable to have a prospective form as similar to the target as possible.

Identification of Target Test Characteristic Curves

In the first operational year for each FCAT 2.0 and EOC assessment, there is no statistical information available for calculating the target TCC. In this case, a form is constructed by the content team, which uses the test specifications to guide its use of the items that have been qualified by field testing. However, in the second year, the previous year's TCC is used as the target TCC. Starting in the third year of the program, the test construction target is taken from the conditional (on theta) average of the post-equated curves from previous years. Once the TCC for a constructed form is calculated, it is plotted and verified for reasonableness.

The calculation of the target TCC also allows for the setting of its upper and lower bounds. Based on the difference that matters (DTM; Dorans & Feigenbaum, 1994), which is approximately 0.5 raw score point, a delta term is identified for each theta point. Next, the delta term specific to each theta point is used to construct the upper and lower bounds around the target TCC. These upper and lower bounds are useful in that they provide the test construction team the acceptable range within which to construct the test.

Figure 2-1 presents a hypothetical example of the TCCs for illustration purposes. The figure includes the test TCC (Test), the target TCC (Target), and the upper bound TCC (Upper_B) and lower bound TCC (Lower_B). The terms cut 2, cut 3, cut 4, and cut 5 represent the four vertical lines separating five achievement levels on the theta scale.

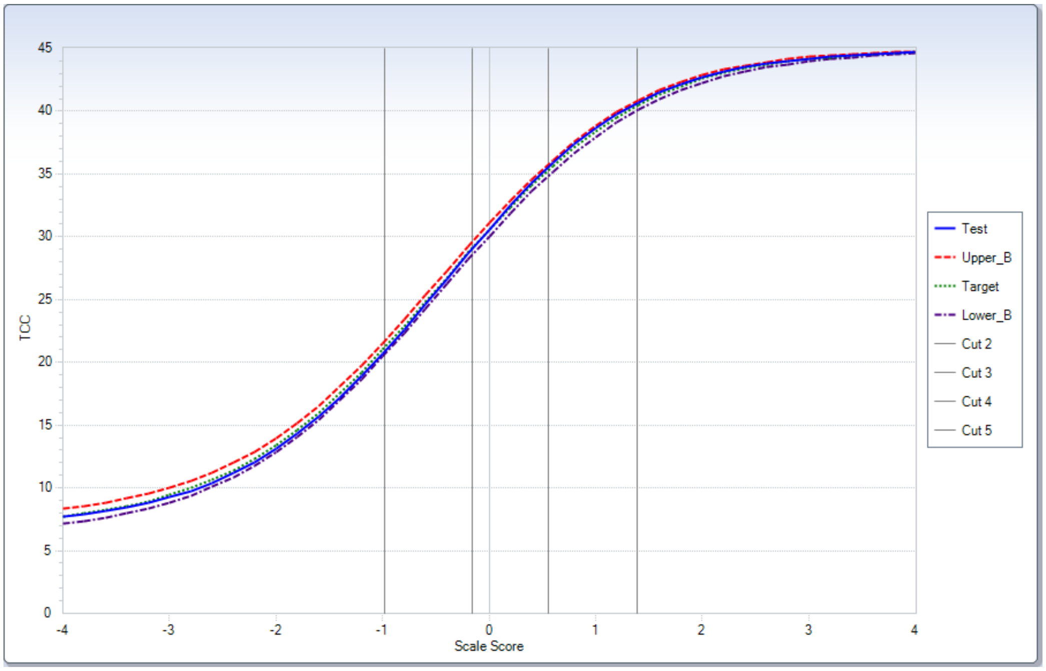

Figure 2-1. Example of Test Characteristic Curves with Hypothetical Cut Points

Test Information Curves (TICs)

One of the most useful features of IRT is the concept of item information, which can be interpreted as the contribution of an item towards the reliability of measurement, conditional on the level of the measured construct (e.g., the level of knowledge and skills in mathematics or reading). The item information curves (IICs) for the items included in a test can be summed up in the TIC, which is conceptually analogous to the reliability of measurement conditional on the scale score level. The test information is calculated using the following formula:

$$I(\theta) = \sum_{i=1}^{n} \frac{P_i^{'}(\theta)^2}{P_i(\theta)Q_i(\theta)} \ , \qquad (2.4)$$

where $P_i(\theta)$ is the probability of an examinee responding correctly to item *i* given an ability of θ, $Q_i(\theta) = 1 - P_i(\theta)$ and $P_i^{'}(\theta)$ is the first derivative of $P_i(\theta)$.

During 2014 FCAT 2.0 test construction, TICs were used for monitoring test information and guiding the selection of items. The goal of test developers was to match the TIC for a newly pulled form to that of the target form. The test developers selected the items that make a new form's TIC similar to the target TIC to a satisfactory degree. There should be as much test information as possible in the scale score area and at the points at which important decisions will be made (i.e., cut-points between achievement levels).

Identification of Target Test Information Curves

The same methods used to find the target TCCs are used to identify target TICs. Test information curves from previous years are added together conditioning on theta and divided by the number of years to the form the target TIC. Next, 10% of the TIC is added to and subtracted from the target TIC at each theta point to construct the upper and lower boundaries (see Figure 2-2 for an example). The four vertical lines in this figure represent hypothetical cut points separating five achievement levels.

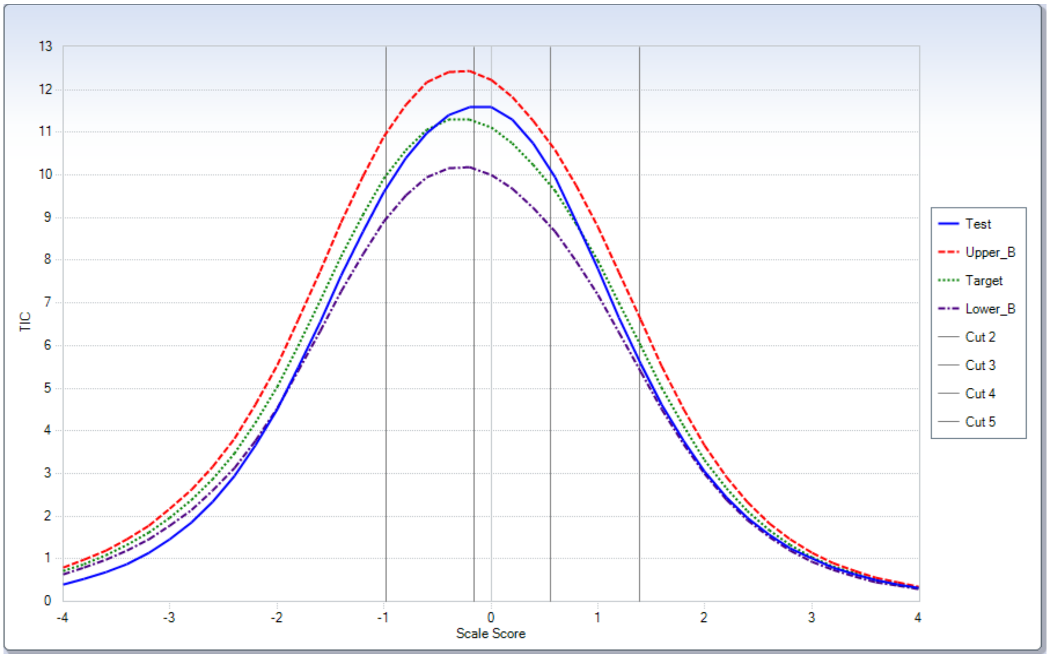

Figure 2-2. Example of Test Information Curves with Hypothetical Cut Points

Conditional Standard Error of Measurement (CSEM) Curve

The conditional standard error of measurement curve shows how much error of measurement is expected at different scale score levels. The CSEM for a given θ can be estimated by using the following formula:

$$CSEM(\theta) = \frac{1}{\sqrt{I(\theta)}}. \qquad (2.5)$$

Because the CSEM is computed as a reciprocal of the square root of the test information curve, the estimated scale scores in the middle of the scale appear to be more reliable than the scores at the low and high ends. Although the ideal test would be one in which the amount of error in a reported score is low regardless of a student's scale score level, the amount of measurement error at extreme score points is typically higher than that at intermediate score levels because a given form would not have enough items at the extreme score points to have a reliable estimate of the extreme scores.

Identification of Target Conditional Standard Error of Measurement Curves

The target CSEM is computed as a reciprocal of the square root of the target test information curve; therefore, the FCAT 2.0 test construction target CSEMs are computed as a reciprocal of the FCAT 2.0 test construction target TICs (see Figure 2-3). Upper and lower bounds for CSEMs are computed by following the same logic. As described previously, the constructed test CSEM should trace the same or a very similar line as the target test CSEM. The upper and lower bounds define how much a constructed test

CSEM curve may move from the target CSEM. The four vertical lines in Figure 2-3 represent hypothetical cut points separating five achievement levels.

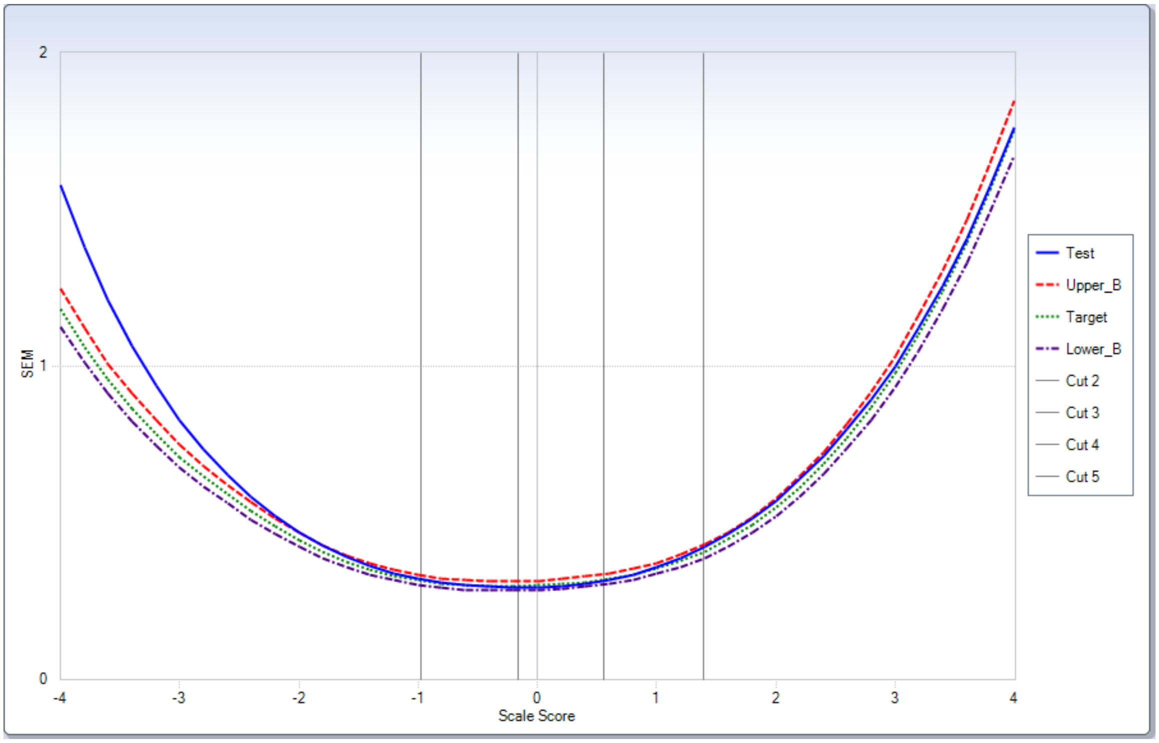

Figure 2-3. Example of Curves for Conditional Standard Error of Measurement with Hypothetical Cut Points

Quality Control in Test Construction

I. Pearson's Content Teams

Pearson's content teams are responsible for pulling the initial forms and subsequent revisions to those forms. Specifically, these teams: (1) select the initial set of operational and anchor items or pre-anchor items, (2) revise the forms to incorporate feedback, (3) provide the forms to Pearson's psychometric team for review, (4) prepare the test construction reports for TDC and FDOE reviews, (5) export the test map from the item banking system and modify it as necessary after FDOE's approval, and (6) forward the test map to the Pearson's test map team.

II. Pearson's Psychometric Team

Pearson's psychometric team is represented by at least one psychometrician for each content area. Each psychometrician is responsible for reviewing the statistical properties of the forms.

III. TDC Content Specialists and Content Leads

TDC content specialists collaborate with Pearson content specialists to revise and select items. Both parties select items with respect to the statistical guidelines and the Florida's FCAT 2.0 and EOC assessment item specification guidelines. If TDC content specialists have a content concern, they communicate with either Pearson's content team or the TDC content lead. If content specialists have statistical concerns, they discuss them with the Pearson psychometrician representing that content area.

TDC content leads review the test forms (core and anchor) and either provide approval or feedback to Pearson content specialists, who will revise the forms in response to TDC feedback. All of the test construction reports are made ready for TDC content lead review so that if the form is approved, it can be immediately forwarded to FDOE psychometrics. TDC content leads ensure that all of the required outputs, reports, and cover sheets that go to FDOE psychometrics are with the test forms. Additionally, the TDC content leads provide justification for any items that have questionable statistical properties (e.g., an item has a DIF flag of B, but there is no identifiable bias).

IV. FDOE Psychometric Team

The FDOE psychometric team conducts a reasonableness check on the item parameters after they are loaded to the item banking system. They review the plots of *p*-value and item difficulty and the item total correlation and item discrimination. FDOE psychometricians also randomly check the actual core item parameters with the original approved parameters from the calibration and equating activities.

The FDOE psychometric team evaluates the differences between the target and pre-equating curves. Additionally, they study the statistical properties of the constructed forms. The proposed forms might be returned to Pearson's content staff for revision based on the feedback that the FDOE psychometrics team provides. Pearson's content and psychometric teams study the feedback and revise the test if alternative items are available for replacement. When the proposed form is accepted by the FDOE psychometrics team, it is forwarded to the FDOE and TDC leadership teams for final review.

V. FDOE/TDC Leadership Teams

Proposed forms are reviewed by Florida's leadership team to determine the suitability of the proposed sets of items for Florida students. FDOE and TDC leadership look at the diversity of the topics, the projected level of difficulty, and the challenge to the examinees. The leadership team approves proposed forms or returns them to be revised by Pearson's content team.

Chapter 3. Administration

Unless specified for a particular program, all information provided in this chapter is applicable to FCAT 2.0 and the Florida End-of-Course (EOC) Assessments.

Eligibility

Public school students are required to participate in the statewide assessment program.

Private school students do not participate in the statewide assessments because it exists to meet federal and state assessment accountability requirements for Florida public schools; however, public school students attending private school through the use of a school choice scholarship, such as the McKay Scholarship, may participate in the statewide assessment program.

Home Education Program students may participate in the appropriate grade-level statewide assessments if they are used as the chosen measure of yearly progress. At the beginning of each school year, parents must notify the district testing office of their intention to use the statewide assessment program as that year's annual measure of their Home Education Program student's progress.

A small number of ***students with disabilities*** who have an individual educational plan (IEP) may not be required to participate in the statewide assessment program. Only those students who meet the exclusion criteria set forth in State Board of Education Rule 6A-1.0943, Florida Administrative Code, Statewide Assessment for Students with Disabilities can be excluded from participating in the statewide assessment program. Specifically, students whose demonstrated cognitive ability prevents them from completing the required coursework and achieving the state standards and who require extensive direct instruction to accomplish and transfer skills and competencies needed for domestic, community living, leisure, and vocational activities may be excluded from taking the statewide assessments. If a student meets the requirements described above and the individual educational plan (IEP) team determines that it is not appropriate for the student to participate, the student is required to take the Florida Alternate Assessment. A student for whom the Next Generation Sunshine State Standards (NGSSS) are inappropriate will work toward a special diploma. There is also a special exemption from participation in the statewide assessments due to extraordinary circumstances. For additional information about the statewide assessment program as it pertains to students with disabilities, please contact the Bureau of Exceptional Students and Educational Services.

English Language Learners (ELLs) are expected to participate in statewide assessments; however, in accordance with State Board Rule 6A-6.0902, Florida Administrative Code, Requirements for Identification, Eligibility Programmatic and Annual Assessments of English Language Learners, if an ELL has been receiving services in an English for

Speakers of Other Languages (ESOL) program operated in accordance with an approved district ELL plan for one year or less AND a majority of the student's ELL committee determines that it is appropriate, the ELL may be exempt from the FCAT 2.0 Reading administration. Exempt ELL students must participate in the Comprehensive English Language Learning Assessment (CELLA). Additionally, all ELLs are expected to participate in the FCAT 2.0 Mathematics, Science, and EOC assessments, as well as in the FCAT 2.0 Writing assessment, no matter how long these students have been receiving services. For additional information about the statewide assessment program as it pertains to ELLs, please contact the Bureau of Student Achievement through Language Acquisition.

Administration Procedures

The test administration manuals for the Florida statewide assessments provide a thorough description of each step (including the test administrator scripts) of the test administration. These are available to administrators on the password-protected PearsonAccess website. A high-level overview of these administration procedures is provided below.

Administration Schedule

FCAT 2.0
FCAT 2.0 is administered each year within a two-week window in the spring. The administration window is published on the FDOE website one year prior to the beginning of the academic school year to allow for the proper planning by Florida school districts. Though there are 10 days in the administration window itself, the FDOE designates Day 1 through Day 4 as the time for administering the first and second sessions of the reading and mathematics assessments. The particular order of the reading and mathematics administrations during that four-day period may vary by grade level. The FDOE designates Days 7 and 8 for administering the first and second sessions of the grade 5 science assessment. Districts may administer grade 8 science on Day 4, Day 5, Day 6, Day 7, or Day 8 of the testing window (both sessions must be administered in one day).

Make-up tests are administered, as needed, from Day 2 through Day 10 of the administration window. However, if students started, but did not complete, a session of the assessment (either for illness or other reasons), they may not finish or make up that session on another day.

Florida End-of-Course Assessment
All students enrolled in an EOC-designated course take the associated EOC assessment. Because 30% of the course grade for first-time grade 9 students will be attributable to the student's score on the EOC assessment in the initial year of the operational assessment, each EOC assessment is administered as closely as possible to the end of the semester to allow students to demonstrate their mastery of the course material. For the Algebra 1 EOC Assessment, those courses include:

- Algebra 1 – 1200310
- Algebra 1 Honors – 1200320
- Algebra 1-B – 1200380
- Pre-AICE Mathematics 1 – 1209810
- IB Middle Years Program – Algebra 1 Honors – 1200390

For the Biology 1 EOC Assessment, those courses include:

- Biology 1 – 2000310
- Biology 1 Honors – 2000320
- Pre-AICE Biology – 2000322
- Biology Technology – 2000430
- Biology 1 Pre-IB – 2000800
- IB Middle Years Program – Biology Honors – 2000850
- Integrated Science 3 – 2002440
- Integrated Science 3 Honors – 2002450

For the Geometry EOC Assessment, those courses include:

- Geometry – 1206310
- Geometry Honors – 1206320
- IB Middle Years Program – Geometry Honors – 1206810
- Pre-AICE Mathematics 2 – 1209820

For the U.S. History EOC Assessment, those courses include:

- United States History – 2100310
- United States History Honors – 2100320

For the Civics EOC Assessment, those courses include:

- Civics – 2106010
- Civics – 2106015
- Civics & Career Planning – 2106016
- Advanced Civics – 2106020
- Advanced Civics – 2106025
- Civics, Advanced & Career Planning – 2106026
- Civics and Digital Technologies – 2106029

Each EOC assessment is a computer-based assessment. Before a student can participate in any computer-based administration, the student must participate in a practice test (ePAT) as directed by the test administrator using the ePAT script. This practice test will enable each student to learn how to use the computer-based system.

In the initial administration year for assessing a course, the assessment window is limited to a three-week period at the end of the spring semester. Districts are directed to pick one of the three weeks in which to administer an EOC assessment for spring 2014. This choice within the FDOE-mandated window allows each district to match the EOC assessment administration week with the close of its particular academic calendar.

Timed Assessment Sessions

Except for students with IEP or Section 504 plans with special provisions, the test administrators allow for the exact amount of time allotted for each session of FCAT 2.0 or EOC assessment, providing no special incentives to students to finish early, though students may do so if they wish. The testing time for each assessment is listed in Table 3-1.

Table 3-1. Testing Time per Session for Florida Statewide Assessments

Assessment	Testing Time per Session*	# of Sessions	Break Within Session
Reading	70 minutes	2	n/a
Mathematics	70 minutes	2	n/a
Science	80 minutes	2	n/a
Algebra 1 EOC	80 minutes	2	10 min. after 80 min. elapsed
Biology 1 EOC	80 minutes	2	10 min. after 80 min. elapsed
Geometry EOC	80 minutes	2	10 min. after 80 min. elapsed
U.S. History EOC	80 minutes	2	10 min. after 80 min. elapsed
Civics EOC	80 minutes	2	10 min. after 80 min. elapsed

*Students may have additional time, if needed, but must complete the test within the same school day.

Accommodations

Several accommodations are available for qualifying students taking the Florida statewide assessments. There are three general categories of students that qualify for accommodations: students with disabilities enrolled in public schools with current IEPs established according to the Individuals with Disabilities Education Act (IDEA); students with current plans developed according to Section 504 of the Rehabilitation Act and the Americans with Disabilities Act (ADA), which may not be individualized; and students identified as ELLs.

Students with Disabilities

Determination of appropriate accommodations in assessment situations for students with disabilities is based on the individual needs of each student. Decisions on accommodations are made by the IEP or Section 504 plan team and recorded on the IEP or Section 504 plan. Students with disabilities are oriented to any test situation through test-taking instruction designed to familiarize them with testing format and procedures.

This orientation takes place near the time of testing. Guidelines recommended for making accommodation decisions include:

1. Accommodations should facilitate an accurate demonstration of what the student knows or can do.
2. Accommodations should not provide the student with an unfair advantage or interfere with the validity of a test; accommodations must not change the underlying skills that are being measured by the test.
3. Accommodations must be the same or nearly the same as those needed and used by the student in completing classroom instruction and assessment activities.
4. Accommodations must be necessary for enabling the student to demonstrate knowledge, ability, skill, or mastery.

Accommodations for the Florida statewide assessments fall into five broad categories. An abridged list of the several accommodations that are available is provided parenthetically below. A full listing of the available accommodations for each category is provided in an appendix of the Test Administration Manual for each administration.

- **Presentation** (large print, braille, one-item-per-page, color transparencies)
- **Responding** (proctor transcription, speech-to-text technology, switches, special keyboards, and other communication devices to indicate answers)
- **Scheduling** (several brief sessions, a finite amount of extended time if allowed in the IEP or Section 504 plan, a specified time of day)
- **Setting** (a familiar space, a small group, special lighting or furniture)
- **Assistive Devices** (visual magnification and auditory amplification devices)

While calculators are allowed for all students taking grade 7 mathematics and above, calculators may not be used by students in grades 3–6 mathematics, even as an accommodation for students with disabilities.

Another accommodation not permitted is the use of manipulative materials, including, but not limited to, counters, base-10 blocks, clock faces, or geometric shapes. Occasionally, an item in a braille test will require the test administrator to modify the test materials to give the student a tactile reference (e.g., a shape may need to be cut out so the student can identify it in order to respond to the item). If such items are included in the braille test, district assessment coordinators are notified and supplied with necessary instructions to communicate to appropriate school personnel prior to the test administration. These modifications are approved by FDOE and apply to braille test materials only.

Special Accommodation for Computer-Based Testing

When the primary administration mode for a Florida statewide assessment is computer-based, a paper form of the assessment is made available as an accommodation for students who require it. In addition to the large print, braille, and one-item-per-page accommodation versions of the assessment, a “regular print” version of the CBT is produced and made available for students for whom it is deemed a necessary accommodation in their IEPs or Section 504 plans.

English Language Learners

Allowable accommodations for students identified as ELLs include:

- Flexible setting
- Flexible scheduling
- Assistance in heritage language
- Approved dictionary

A full elaboration of the definitions and conditions for the use of these accommodations for ELLs can be accessed via http://www.fldoe.org/ese/pdf/fcatteam.pdf.

Test Security

Florida State Board of Education Rule 6A-10.042, FAC, was developed to meet the requirements of the Test Security Statute, s. 1008.24, F.S., and applies to anyone involved in the administration of a statewide assessment. The State Board of Education Rule prohibits activities that may threaten the integrity of the test. The full text of both the Florida Test Security Statute and the State Board of Education Rule are reprinted in an appendix of each Test Administration Manual.

Examples of prohibited activities are listed below:

- Reading the passages or any test items
- Revealing the passages or test items
- Copying the passages or test items
- Explaining or reading passages or test items for students
- Changing or otherwise interfering with student responses to test items
- Copying or reading student responses
- Causing achievement of schools to be inaccurately measured or reported

If students with current IEPs, Section 504 plans, or ELL plans have allowable accommodations documented, test administrators may provide the accommodations as described in the appendix of the Test Administration Manual and may modify the scripts as necessary to reflect the allowable accommodations. This stipulation does not give test administrators permission to paraphrase items. Modified scripts for students using braille versions of the test are provided with the braille test materials. Modifications to the regular administration scripts for students using large print and one-item-per-page materials are located in the *Special Documents Instructions* section of the associated Test Administration Manual.

The security of all test materials must be maintained before, during, and after test administration. Under no circumstances are students permitted to assist in preparing secure materials before testing or in organizing and returning materials after testing.

For computer-based testing, any monitoring software that would allow test content on student workstations to be viewed on another computer during testing must be turned off.

After any administration, initial or make-up, student authorization tickets, session rosters, hardcopy reference sheets and periodic tables, and used work folders must be returned immediately to the school assessment coordinator and placed in locked storage. No more than three persons may have access to the locked storage room. Secure materials must not remain in classrooms or be taken off the school's campus overnight.

District assessment coordinators must require that all school administrators, school assessment coordinators, technology coordinators, test administrators, and proctors sign and return an FCAT/FCAT 2.0/EOC Test Administration and Security Agreement, provided in an appendix in the Test Administration Manual, stating that they have read and agree to abide by all test administration and test security policies and procedures. Additionally, any other person who assists the school assessment coordinator, technology coordinator, or test administrator must sign and return an agreement.

Each school is required to maintain an accurate security log for each testing room. Anyone who enters a testing room for the purpose of monitoring the test is required to sign the log. This rule applies to test administrators, proctors, and anyone who relieves a test administrator, even for a short break, regardless of how much time he or she spends monitoring a testing room.

Test administrators must NOT administer tests to their family members. Students related to their assigned test administrator should be reassigned to an alternate test administrator.

Inappropriate actions by school or district personnel may result in student or classroom invalidations and/or the loss of teaching certification.

Data Forensics Program

FDOE utilizes statistical analyses of test data for the FCAT, FCAT 2.0, and Florida EOC Assessments to help ensure fair and valid results statewide. Through Pearson, the FCAT/FCAT 2.0/EOC assessment contractor, FDOE has contracted with Caveon Test Security to provide its *Caveon Data Forensics™* for all statewide assessments. Caveon analyzes data to identify highly unusual test results for two primary groups: 1) students with extremely similar test responses; and 2) schools with improbable levels of similarity, gains, and/or erasures. Although FDOE has examined data to identify testing irregularities in the past, the spring 2014 FCAT, FCAT 2.0, and Florida EOC Assessment test administration was the fourth year that FDOE contracted with Caveon Test Security to implement the data forensics program.

Students with Extremely Similar Test Responses

FDOE flags and invalidates the test scores of students whose test responses are so similar (and statistically aberrant) that the validity of the test responses must be questioned. For this spring, a very conservative threshold was used. FDOE flags and invalidates only those instances where the chance of two or more tests being so similar while taken independently is one chance in a trillion when tests are taken normally. Despite such a conservative threshold (the chance of being struck by lightning is one in a million), FDOE has instituted an appeal process where school districts have the option to present evidence that explains and justifies why two or more students' test responses are so similar. Based upon an individual appeal, FDOE may reverse the score invalidation.

Those schools that choose to submit a written request for student appeals are asked to detail the circumstances surrounding the student's testing experience and include supporting or related evidence, such as:

- Testing conditions and test security protocol at the testing site
- Testing location of students
- FCAT testing history of students
- Student statements
- Test administrator/proctor statements

Schools with Improbable Similarity, Gains, and Erasures

FDOE flags schools when test results are marked by extremely unusual levels of similarity, score gains, and erasures. The erasure analysis is based on an examination of answer-changing rates from wrong-to-right, taking into account other types of erasures. This analysis identifies extreme statistical outliers that could involve tampering with test answer documents. The erasure rate index used is at least a 12. An erasure rate index of 12 represents a level of erasures that would be expected to occur once in a trillion times when tests are taken under standardized conditions.

In 2014, FDOE flagged schools with improbable similarities, gains, and erasures. Schools are flagged in those extreme cases where their rate far exceeds the state mean for similarities, gains, and erasures. This flagging may prompt a formal investigation led by FDOE, its Office of Professional Practices, and the Inspector General's Office, possibly leading to sanctions. School districts with flagged schools are required to conduct an internal investigation and send their results to FDOE. Based on their findings, districts are able to submit appeals. For the spring test administration, each school identified for investigation is initially assigned an "I" for its accountability outcome(s), including each of the following: school grade, AYP outcome, and school improvement rating for alternative schools. The "I" rating indicates that FDOE is not able to accurately evaluate the accountability rating(s) for the identified schools based on available data. This rating may be changed to another accountability rating (a regular school grade, AYP outcome, school improvement rating) after the erasure issues are resolved, or when the Commissioner of Education determines that sufficient data are available to accurately assign a school grade, AYP outcome, or school improvement rating. Superintendents may request during the appeals process that the schools in question have the "I" rating replaced by a regular school grade, AYP rating, or school improvement rating if or when sufficient data become available to make this determination.

School district investigations should produce documentation that should include but is not limited to the following:

- Testing conditions and test security protocol at the testing site
- Signed statements that address possible testing irregularities
- Staff assignments as they relate to testing
- Stakeholder interviews
- Research using erasure data for student answer documents and student results files

Release/Uphold Invalidation

K-12 Management makes the final determination on whether student scores are released or held up. Districts are informed of the determination by a faxed memo. The contractor Pearson is informed of the updated scores. Scores that have been determined in this way are released during later rounds of reporting. K-12 Management along with Professional Practices and the Inspector General's office make the final determination on whether schools flagged for similarity, gains, and erasures are to be released. Those schools determined to be released will receive their school grade. In those extreme cases further investigation by FDOE may be needed.

Chapter 4. Reports

Appropriate Uses for Scores and Reports

As with any large-scale assessment, Florida statewide assessments provide a point-in-time snapshot of information regarding student achievement. For that reason, scores must be used carefully and appropriately if they are to permit valid inferences to be made about student achievement. Because all tests measure a finite set of skills with a limited set of item types, decisions about student placement or instructional interventions should be based on multiple sources of information—including, but not limited to, test scores.

Information about student performance is provided on individual student reports and summary reports for schools, districts, and the state. This information may be used in a variety of ways. Interpretation guidelines were developed and published as a component of the release of public data; these documents, *Understanding FCAT 2.0 Reports* and *Understanding Florida EOC Assessment Scores,* are located on the FDOE website at http://fcat.fldoe.org/fcat2/ and http://fcat.fldoe.org/eoc/ respectively. These documents are comprehensive and updated annually as changes to reports or reporting are enacted.

Reports Provided

Florida publishes several reports for each year. Table 4-1 provides a list of reports and score types appearing on these reports. These reports are published for both FCAT 2.0 and EOC assessments. The following section contains a description of the scores provided. In 2014, only Civics assessment scale scores were reported using a temporary scaling system pending standard-setting outcomes in the fall of 2014. Civics EOC Assessment was reported using the T-scale (standard scores with a theoretical mean of 50 and standard deviation of 10). No achievement level information was provided for the 2014 Civics EOC Assessment. Meanwhile, the FCAT 2.0 Reading and Mathematics assessments used new developmental scale scores with approved cut scores in 2011 (discussed in “Chapter 5. Performance Standards”). The FCAT 2.0 Science Assessment also used its own unique scale with the mean of 200 and standard deviation of 20 and with new cut scores established in 2012. The Algebra 1, Biology, and Geometry EOC Assessments also used their own unique scale with the mean of 400 and standard deviation of 25. Algebra EOC achievement level cut scores were established in 2011, meanwhile the others were established in 2012 the same as FCAT 2.0 Science (discussed in “Chapter 5. Performance Standards”).Furthermore U.S. History cut scores were established in the fall of 2013.

Table 4-1. Types of Reports Provided for FCAT 2.0 and EOC Assessments

	Types of Scores			
Report	**Raw Scores**	**Scale Score**	**Achievement Level**	**Longitudinal Scores***
Individual Student Report	X	X	X	X
Parent Report	X	X	X	X
School Report of Students	X	X	X	
District Report of Schools	X	X	X	
District Summary Report	X	X	X	
State Report of Districts	X	X	X	
State Summary Report	X	X	X	

Note: Longitudinal scores refer to the same student' historical scores (e.g., scale scores from previous tests).

For FCAT 2.0 Writing at grades 4, 8, and 10, only raw scores (1–6) are reported at all levels of reports. Beginning in 2010, each essay was scored by one rater. In previous years, two raters were used and the scores were averaged. Therefore, a student could have received a half-point score, such as 4.5, before 2010, whereas in 2010 and 2011, no half-point scores were possible. In 2012, 2013, and 2014, in contrast to 2010 and 2011, each response to a given prompt was rated by two raters and half point scores were again reported.

Individual Student Reports

Florida's individual student reports provide information on a student's performance in each subject measured as well as a comparison of his or her performance relative to other students in the state. The reports contain numeric and graphical information about student overall performance. Students may use this information to understand their achievement on each subject as well as their relative performance across subjects. The report provides the number of points earned (raw score) by the student in each reporting category, which can be used to evaluate relative strengths and weaknesses within a subject. The maximum possible points and state mean are provided to assist the student in interpreting his or her performance. The reports provide verbal definitions of scores in English, Spanish, and Haitian Creole.

The reports for grades 3 and 10 reading indicate whether or not the student has met the required score. Florida statute requires grade 3 students to earn at least a Level 2 score in order to progress to grade 4 (in lieu of passing FCAT 2.0 the school can provide other forms of evidence). Florida statute requires a student to earn a passing score on the grade 10 reading test in order to be awarded a high school diploma.

These reports also contain the student's prior year scale scores and achievement levels. All available historical scores for the student are presented for each subject, including cases where the student took the same grade-level test multiple years in a row. In 2014, Civics EOC Assessment students received raw scores and an overall scale score. The use and interpretation of student performance on this test was determined locally in 2014 in the absence of achievement levels.

Parent and Student Report

The Parent and Student Report contains the information presented in the Individual Student Report, in the manner in which it was presented for that report. Additional information includes opening comments from the commissioner of education and further definition of the reported scores. The information presented in this report can be used by parents to help them understand their child's achievement across all subjects tested.

Again, no achievement level information was reported for the 2014 Civics Assessment.

School Report of Students

This report contains a listing of all students tested at the school and their performance on each test. Only numerical information is provided. Scores provided include scale scores, raw scores for reporting category, the maximum possible point total for reporting category, the achievement level, and last year's score (if available). This report is generated separately for each subject and grade.

District Report of Schools

The District Report of Schools contains a summary record for each school in a district. Scores reported include the average scale score, the percentage of students in each achievement level, and the average raw score for each reporting category. The maximum possible point total is provided for each reporting category. The total number of tested students is also provided. This report is generated separately for each subject and grade.

District Summary Report

The District Summary Report contains a summary record for each grade. Scores include the average scale score, the percentage of students in each achievement level, and the average raw score for each reporting category. The maximum possible point total is provided for each reporting category. The total number of tested students is also provided. The statewide value for each score is provided for comparison purposes. The report is generated separately for each subject.

State Report of Districts

The State Report of Districts contains a summary record for each district in the state. Scores reported include the average scale score, the percentage of students in each achievement level, and the average raw score for each reporting category. The maximum possible point total is provided for each reporting category. The total number of tested students is also provided. The last record in the report contains the statewide values for the scores. This report is generated separately for each subject and grade.

State Summary Report

The State Summary Report contains a summary record for each grade. Scores include the average scale score, the percentage of students in each achievement level, and average raw score for each reporting category. The maximum possible point total is provided for each reporting category. The total number of tested students is also provided. The report is generated separately for each subject.

Description of Scores

Scores are the end product of the testing process. They provide information about how each student performed on the tests. Four different types of scores are used on the Florida statewide assessments reports: scale scores, subscale raw scores, achievement levels, and historical scores. These four scores are related to one another. This section briefly describes each type of score.

Subscale Raw Score

The subscale raw score is the sum of points earned across items from a particular reporting category. By themselves, these raw scores have limited utility. They can be interpreted only in reference to the total number of items in a reporting category or using normative information. They cannot be compared directly across test forms or years of administrations. Several values derived from raw scores are included to assist in interpreting the raw scores: maximum points possible and aggregate averages (for school-, district-, and state-level reports).

Scale Score

Scale scores are mathematically determined based on the patterns of student responses. The use of scale scores allows for the maintenance of a consistent metric across test forms and the comparison of scores across all test administrations within a particular grade and subject. They can be used to determine whether a student met the standard or achievement level in a manner that is fair across forms and administrations because scale scores adjust for different form difficulties. Schools can also use scale scores to compare the knowledge and skills of groups of students within a grade and subject across years. These comparisons can be used in assessing the impact of changes or differences in instruction or curriculum.

Student results for the spring 2014 FCAT 2.0 Reading and Mathematics assessments were reported on the new developmental scale scores (DSS). On the other hand, the FCAT 2.0 Science assessment was reported on the new FCAT 2.0 Science scale scores and achievement levels. Details about the new FCAT 2.0 score scale are described in "Chapter 6. Scaling."

Also in 2014, new scale scores for the U.S. History EOC Assessment were reported on an EOC assessment scale customized with the slope of 25 and the intercept of 400, which has a mean of 400 and a standard deviation of 25. The scale scores for the Civics EOC Assessments were reported on a T-scale, which had a theoretical mean of 50 and a standard deviation of 10. The T-scores ranged from 20 to 80. Hence, during this initial operational year for Civics, the school districts were directed to the T-scale, rather than to a customized EOC assessment scale, to determine 30% of the student grades.

Details about how scale scores are computed are given in “Chapter 6. Scaling.”

Achievement Levels

To help parents and schools interpret scale scores, achievement levels are reported. Each achievement level is determined by the student’s scale score. The range for an achievement level is set during the standard-setting process. Each time a new test is implemented, panels of Florida educators set the achievement levels. For 2014, achievement levels for FCAT 2.0 Reading, Mathematics, Science, the Algebra 1, Biology 1, Geometry, and U.S. History EOC Assessments were reported to the stakeholders. No achievement levels were reported for the Civics EOC Assessment.

Historical Scores

Students who have taken FCAT 2.0 in previous years have their historical test performances reported, such as in 2011 and 2012, provided the correct student identifications were submitted on the student answer documents. Historical scores include FCAT 2.0 scale scores and FCAT 2.0 achievement levels.

Appropriate Score Uses

The tests in the Florida statewide assessment system are designed primarily to measure student achievement and to determine school and district accountability related to the implementation of the Next Generation Sunshine State Standards (NGSSS). Students must earn specific scores on the grade 3 reading and grade 10 reading tests for grade promotion and graduation purposes. They are summative measures of a student’s performance in a subject at one point in time. They provide a snapshot of the student’s overall achievement, not a detailed accounting of the student’s understanding of specific content areas defined by the standards. Test scores from Florida statewide assessments, when used appropriately, can provide a basis for making valid inferences about student performance. The following list outlines some of the ways the student scores can be used.

- *Reporting results to parents of individual students*
  The information can help parents begin to understand their child’s academic performance as related to the NGSSS.
- *Evaluating student scores for placement decisions*
  The information can be used to suggest areas needing further evaluation of student performance. Results can also be used to focus resources and staff on a particular group of students who appear to be struggling with the NGSSS. Students may also exhibit strengths or deficits in reporting categories measured on these tests. Because the reporting categories are based on small numbers of items, the scores must be used in conjunction with other performance indicators to assist schools in making placement decisions, such as whether a student should take an improvement course or be placed in a gifted or talented program.
- *Evaluating programs, resources and staffing patterns*

> Test scores can be a valuable tool for evaluating programs. For example, a school may use its scores as one piece of evidence in evaluating the strengths and weaknesses of a particular academic program or curriculum in the school or district as it relates to the NGSSS.

Individual Students

Scale scores determine whether a student's performance has met or fallen short of the proficiency criterion level. Test results can also be used to compare the performance of an individual student with the performance of a similar demographic group or an entire school, district, or state group. For example, the score of a Hispanic student in a gifted program could be compared with the average scores of Hispanic students, gifted students, all the students on campus, or any combination of these aggregations.

Reporting category scores provide information about student performance in more narrowly defined academic content areas. For example, individual scores on reporting categories can provide information to help identify areas in which a student may be having difficulty, as indicated by a particular test. Once an area of possible weakness has been identified, supplementary data should be collected to further define the student's instructional needs.

Finally, individual student test scores must be used in conjunction with other performance indicators to assist in making placement decisions. All decisions regarding placement and educational planning for a student should incorporate as much student data as possible.

Groups of Students

Test results may be used to evaluate the performance of student groups. The data should be viewed from different perspectives and compared with district and state data to gain a more comprehensive understanding of group performance. For example, the average scale score of a group of students may show they are above the district and/or state average, yet the percentage of students who are proficient in the same group of students may be less than the district or state percentage. One perspective is never sufficient.

Test results may also be used to evaluate the performance of student groups over time. Average scale scores can be compared across test administrations within the same grade and subject area to provide insight into whether student performance is improving across years. The percentages of students in each achievement level can also be compared across administrations within the same grade and subject area to provide insight into whether student performance is improving across years.

Test scores can also be used to compare the performance of different demographic or program groups (within the same subject and grade) on a single administration to determine which demographic or program group, for example, had the highest or lowest average performance, or the highest percentage of students considered proficient on the NGSSS. Other test scores can be used to help evaluate academic areas of relative strength or weakness. Average performance on a reporting category can help identify areas where further diagnosis may be warranted for a group of students.

Test results for groups of students may also be used when evaluating instructional programs; year-to-year comparisons of average scale scores or the percentage of students considered proficient in the program will provide useful information. Considering test results by subject area and by reporting category may be helpful when evaluating curriculum, instruction, and their alignment to standards because all the Florida statewide assessments are designed to measure content areas within the required state standards.

Generalizations from test results may be made to the specific content domain represented by the reporting categories being measured on the test. However, because the tests are measuring a finite set of skills with a limited set of items which vary from year to year, any generalizations about student achievement derived solely from a particular test should be made cautiously and with full reference to the fact that the conclusions were based on only one test. All instruction and program evaluations should include as much information as possible to provide a more complete picture of performance.

Cautions for Score Use

Test results can be interpreted in many different ways and used to answer many different questions about a student, educational program, school, or district. As these interpretations are made, there are always cautions to consider.

Understanding Measurement Error

When interpreting test scores, it is important to remember that test scores always contain some amount of measurement error. That is to say that test scores are not infallible measures of student characteristics. Rather, some score variation would be expected if the same student is tested across occasions using equivalent forms of the test. This effect is due partly to day-to-day fluctuations in a person’s mood or energy level that can affect performance and partly to a consequence of the specific items contained on a particular test form the student takes. At an individual level, one form may result in a higher score for a particular student than would another form. This difference in score may occur even though all testing programs in Florida conduct a careful equating process (described in “Chapter 7. Equating”) to ensure that test scores from different forms can be compared. Because measurement error tends to behave in a fairly random fashion, when aggregating over students these errors in the measurement of students tend to cancel each other out. “Chapter 8. Reliability” describes measures that provide evidence indicating measurement error on Florida statewide assessments is within a tolerable range. Nevertheless, measurement error must always be considered when making score interpretations.

Using Scores at Extreme Ends of the Distribution

As with any fixed-length test, student scores at the extremes of the score range must be viewed cautiously. For instance, if a student achieves the maximum scale score for the Grade 5 FCAT 2.0 Mathematics assessment, it cannot be determined whether the student would have achieved a higher score if a higher score were possible. Caution should be taken when comparing students who score at the extreme ends of the distribution.

Analyses of student scores at extreme ends of the distribution should also be undertaken cautiously because of a phenomenon known as regression toward the mean. Students who scored high on the test may achieve a lower score the next time they test because of regression toward the mean. (The magnitude of this regression effect is proportional to the distance of the student's score from the mean and bears an inverse relationship to reliability.) For example, if a student who obtained a high score of 38 out of 40 took the same test again, there would be many more opportunities—compared to a student with a score close to the mean—to incorrectly answer an item that he or she originally answered correctly (38 opportunities, in fact), while there would only be two opportunities to correctly answer items missed the first time. If an item is answered differently, it is more likely to decrease the student's score than to increase it. The converse of this is also true for a student with a very low score; the next time the student tests, he or she is more likely to achieve a higher score, and this higher score may be a result of regression toward the mean rather than an actual gain in achievement. It is more difficult for students with very high or very low scores to maintain their scores than it is for students in the middle of the distribution. The regression toward the mean phenomenon applies to any test and is another reason to be cautious when interpreting any scores at extreme ends of the distribution.

Interpreting Score Means

The scale score mean (or average) is computed by summing each student's scale score and dividing by the total number of students. Although the mean provides a convenient and compact representation of where the center of a set of scores lies, it is not a complete representation of the observed score distribution. Very different scale score distributions in two groups could yield the same mean scale score. When a group's scale score mean falls above the scale score designated as the passing or proficient cut score, it does not necessarily follow that most students received scale scores higher than the cut score. It can be the case that a majority of students received scores lower than the cut score while a small number of students got very high scores. Only when more than half of the students score at or above the particular scale score can one conclude that most students pass or are proficient on the test. Therefore, both the scale score mean and percentage at or above a particular scale cut score should be examined when comparing results from one administration to another.

Using Reporting Category Information

Reporting category information can be useful as a preliminary survey to help identify skill areas in which further diagnosis is warranted. The standard error of measurement associated with these generally brief scales makes drawing inferences from them at the individual level very suspect; more confidence in inferences is gained when analyzing group averages. When considering data at the reporting category level, the error of measurement increases because the number of possible items is small. In order to provide comprehensive diagnostic data for each reporting category, the tests would have to be prohibitively lengthened. Once an area of possible weakness has been identified, supplementary data should be gathered to understand strengths and deficits.

Also, because the tests are equated only at the total subject-area test scale score level, year-to-year comparisons of reporting-category-level performance should be made cautiously. Significant effort is made to approximate the overall difficulty of the test from year to year during the test construction process, but some fluctuations in difficulty do occur at the reporting category level across administrations. Observing trends in reporting category performance over time, identifying patterns of performance in clusters of benchmarks testing similar skills, and comparing school or district performance to district or state performance are appropriate uses of group reporting category information.

Furthermore, for tests under development with new content standards, as with FCAT 2.0, changes to the test content and the percentage of score points allotted to each reporting category may change. Some of these changes may be significant. When changes in test content occur, comparing student performance across years is particularly difficult, and under these circumstances the advice from measurement professionals is likely to discourage making such comparisons.

Program Evaluation Implications

Test scores can be a valuable tool for evaluating programs, but any achievement test can give only one part of the picture. Standard 15.4 in the *Standards for Educational and Psychological Testing* states, “In program evaluation or policy studies, investigators should complement test results with information from other sources to generate defensible conclusions based on the interpretation of the test result.” The FCAT 2.0 and EOC assessments are not all-encompassing assessments measuring every factor that contributes to the success or failure of a program. Although more accurate evaluation decisions can be made by considering all the data the test provides, users should consider test scores to be only one component of a comprehensive evaluation system.

Chapter 5. Performance Standards

Introduction

Standard setting is the process of establishing cut scores on a test. For FCAT 2.0 and EOC assessments in Florida, two panels of Florida stakeholders meet separately to make recommendations regarding cut scores. Committees of Florida educators—referred to in this document as the Educator Panel—make content-based recommendations for cut scores. A separate panel of superintendents, business leaders, and other community stakeholders—referred to in this document as the Reactor Panel—then review the outcomes from the Educator Panel meeting and form their own set of recommended cut scores.

In 2011, standard setting was conducted for FCAT 2.0 Reading and Mathematics assessments and Algebra 1 EOC Assessment. In 2012, standard setting was conducted for FCAT 2.0 Science assessments and Biology 1 and Geometry EOC Assessments. As recommended by Dr. Mark Reckase (2010), the Modified Angoff[1] standard setting method (Angoff, 1971; Cizek & Bunch, 2007) is used by Educator Panel participants to set their recommended cut scores. The outcomes of the Educator Panel and Reactor Panel meetings are described in this chapter, as are the final decisions made by the State Board of Education.

Interim Performance Standards for the 2011 Administration of Reading and Mathematics

The first administration of FCAT 2.0 Reading and Mathematics assessments occurred in spring 2011, followed by the establishment of interim performance levels on the FCAT 2.0 through the use of concordant FCAT scores and FCAT performance levels. These interim performance level cut scores were applied to the 2011 score reports.

Interim performance standards on FCAT 2.0 Reading and Mathematics tests were established through a statistical linkage between state results in 2010 and 2011. (See Table 5-1.) The procedures used to establish this linkage were reviewed and approved by the FDOE leadership and advisors—including the National Technical Advisory Committee, which includes nationally known psychometricians and policy makers.

The high-level linking process included the following steps:

1. Establish an interim FCAT 2.0 scale using the procedures defined in “Chapter 6. Scaling.” Obtain the distribution of reportable interim scores for 2011.
2. Obtain the distribution of reported scores from the 2010 administration of FCAT.
3. Establish a concordance between FCAT and FCAT 2.0 reported scores using the distributions from Steps 1 and 2. See “Chapter 7. Equating.”

[1] A judgmental method commonly used to set cut scores, which requires subject matter experts to estimate the percentage of students at a particular achievement level who should correctly answer a given item.

4. For each student, use the interim FCAT 2.0 score to obtain the FCAT concordant score derived in Step 3. That concordant FCAT score is the reported score for 2011.
5. Assign the published FCAT performance levels to the concordant FCAT score.

Table 5-1. FCAT Scale Score Ranges for Performance Levels on 2011 Tests

Reading					Grade	Mathematics				
Level 1	Level 2	Level 3	Level 4	Level 5		Level 1	Level 2	Level 3	Level 4	Level 5
100-258	259–283	284–331	332–393	394–500	3	100–252	253–293	294–345	346–397	398–500
100–274	275–298	299–338	339–385	386–500	4	100–259	260–297	298–346	347–393	394–500
100–255	256–285	286–330	331–383	384–500	5	100–287	288–325	326–354	355–394	395–500
100–264	265–295	296–338	339–386	387–500	6	100–282	283–314	315–353	354–390	391–500
100–266	267–299	300–343	344–388	389–500	7	100–274	275–305	306–343	344–378	379–500
100–270	271–309	310–349	350–393	394–500	8	100–279	280–309	310–346	347–370	371–500
100–284	285–321	322–353	354–381	382–500	9	No Longer Tested				
100–286	287–326	327–354	355–371	372–500	10					

The achievement level reports in the 2011 yearbook provided the percentage of students by achievement level. Summaries of the scale score distribution can be found in the scale distribution and statistics reports of that yearbook. Because equipercentile linking was used in 2011, the achievement level percentages and the scale score distributions were very similar to those obtained in 2010.

Setting Performance Standards—2011

The FCAT 2.0 Reading and Mathematics and the Algebra 1 EOC Assessment are statewide tests. FCAT 2.0 Reading is currently administered in grades 3–10, FCAT 2.0 Mathematics is currently administered in grades 3–8, and the Algebra 1 EOC Assessment is administered at the end of the Algebra 1 (or equivalent) course. These testing programs assess student achievement on the Next Generation Sunshine State Standards (NGSSS). FCAT 2.0 Reading is composed entirely of multiple-choice items, as is FCAT 2.0 Mathematics in grade 3. Each FCAT 2.0 Mathematics Assessment in grades 4–8 includes a few gridded-response items in addition to the multiple-choice items. The Algebra 1 EOC Assessment is composed of multiple-choice and fill-in response items.

The goal of Florida standard setting is to establish cut scores on a test marking different levels of achievement. Achievement levels describe the success a student has achieved with the content assessed. For Florida assessments, achievement levels range from 1 to 5, with Level 1 being the lowest and Level 5 being the highest. To be considered on grade level, students must achieve Level 3 or higher. Level 3 indicates satisfactory performance. The following Achievement Level Policy Definitions apply to all FCAT 2.0 and Florida EOC assessments:

- **Level 5:** Students at this level demonstrate mastery of the most challenging content of the *Next Generation Sunshine State Standards*.
- **Level 4:** Students at this level demonstrate an above satisfactory level of success with the challenging content of the *Next Generation Sunshine State Standards*.
- **Level 3:** Students at this level demonstrate a satisfactory level of success with the challenging content of the *Next Generation Sunshine State Standards*.

- **Level 2:** Students at this level demonstrate a below satisfactory level of success with the challenging content of the *Next Generation Sunshine State Standards*.
- **Level 1:** Students at this level demonstrate an inadequate level of success with the challenging content of the *Next Generation Sunshine State Standards*.

Standard Setting Process Overview

After the interim standards were applied to 2011 results, formal standard setting meetings were undertaken to set cuts that would be applied to the results of the 2012 administration and subsequent administrations. From September 20 through September 23, 2011, several panels of Florida educators—collectively known as the Educator Panel—were convened in Orlando, Florida, to make content-based recommendations for cut scores for the FCAT 2.0 Grades 3–8 Mathematics and Grades 3–10 Reading, as well as the Algebra 1 End-of-Course (EOC) Assessment. A separate panel of superintendents, business leaders, and other community stakeholders—Reactor Panel #1—was convened September 29 through September 30, 2011, in Tallahassee, Florida, to review the outcomes from the Educator Panel meeting and to form their own set of recommended cut scores for the FCAT 2.0 and Algebra 1 EOC Assessments. At the request of the SBE, a second Reactor Panel—Reactor Panel #2—was convened on November 30, 2011, in Tallahassee, Florida, to provide additional input.

A total of 267 Florida educators participated in the Educator Panel meeting, a total of 21 community stakeholders participated in the first Reactor Panel meeting, and eight community stakeholders participated in the second Reactor Panel meeting.

The FCAT 2.0 and Algebra 1 EOC Assessment standard setting process can be summarized as taking place in six phases:

1. Updates and revisions to the Achievement Level Descriptions,
2. Educator Panel meeting,
3. Reactor Panel meeting,
4. Commissioner’s recommendations,
5. Public input, and
6. State Board of Education final decision.

The main purposes of this chapter are to document the methodologies and outcomes from the Educator Panel and the Reactor Panels and to document the final cut scores set by the SBE. Prior to the Educator Panel meeting, the Achievement Level Descriptions were revised by FDOE staff based on feedback from the public review. Activities associated with those revisions will not be dealt with in detail in this report.

Panelists

The success of standard setting requires the involvement of a variety of participants from facilitators to primary stakeholders. FDOE was responsible for recruiting and selecting individuals to serve in one of two groups: the Educator Panel—those who engaged in the standard setting process (detailed description is provided below)—and the Reactor Panels—individuals who reviewed the activities and results of the standard setting and provided feedback to FDOE, proposing modifications to the Educator Panel’s recommendations as deemed necessary. FDOE also reviewed the list of participants for the Educator Panel and nominated individuals to serve as table leaders throughout the standard setting process. Panelists in each committee room were grouped into three tables, each with a table leader. Table leaders assisted the meeting facilitators with the maintenance of meeting materials, dialogue among table participants during discussions, and determining the readiness of the participants to engage in the various rounds of judgment.

As with prior recruiting practices for FCAT 2.0 development activities, FDOE recruited members of the Educator Panel considering the following criteria:

- Regional representation
- Teaching and subject matter expertise
- Understanding of the Florida instructional environment
- Experience with diverse groups of students
- Special education expertise
- English Language Learner expertise
- Individual demographics

As shown in the following tables, there were 15 separate committees (Grades 3–10 FCAT 2.0 Reading, Grades 3–8 FCAT 2.0 Mathematics, and Algebra 1 EOC Assessment), and a total of 267 Educator Panel participants (15 to 21 panelists per committee).

All panelists were asked to provide voluntary demographic information. The Educator Panel participants' professional backgrounds are summarized in Table 5-2.

Table 5-2. Educator Panel: Percentages of Panelists by Professional Background

Subject	Grade	TCH	COA	SPC	ADM	OTH	NR
Reading	3	47	21	5	21	5	11
	4	41	18	24	6	0	12
	5	44	22	11	11	0	11
	6	50	14	0	14	7	21
	7	33	28	11	22	11	6
	8	35	12	24	12	6	18
	9	40	20	7	40	0	7
	10	26	11	21	32	0	11
Mathematics	3	43	14	10	14	10	14
	4	47	18	24	6	12	0
	5	40	40	0	15	0	5
	6	37	5	37	11	0	11
	7	24	33	19	14	0	14
	8	35	25	10	25	0	10
	Algebra 1	26	11	16	58	0	5

Note. TCH=Teacher, COA=Coach, SPC=Specialist, ADM=Administrator, OTH=Other, NR=No Response. Some participants indicated multiple professional backgrounds in their responses, so percentages may not sum to 100%.

The Educator Panel participants' experience is summarized in Table 5-3. This table lists the minimum, mean, and maximum value for both overall teaching experience and experience within the panelists' assigned grade and subject for each committee.

Table 5-3. Educator Panel: Teaching Experience

		Total Yrs. Experience			Yrs. In This Grade/Subj.		
Subject	**Grade**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**
Reading	3	7	18.8	37	0	7.1	15
	4	5	16.1	37	0	5.9	19
	5	6	18.0	37	2	13.4	30
	6	6	16.9	30	2	5.0	9
	7	6	16.2	32	0	8.9	32
	8	8	20.3	40	0	8.1	28
	9	4	16.5	38	0	6.6	14
	10	5	14.3	30	2	10.8	24
Mathematics	3	5	17.4	27	0	7.6	26
	4	5	14.9	34	3	8.7	26
	5	5	14.9	34	1	9.8	20
	6	5	21.2	36	0	6.5	25
	7	5	16.1	37	1	12.2	33
	8	4	18.3	38	2	11.1	35
	Algebra 1	5	19.6	38	5	15.6	35

Pre-Workshop: Table Leader Training

Prior to the standard setting workshop, Pearson and FDOE facilitated a table leader training session. This training involved only those panelists chosen to serve as table leaders as well as key staff from FDOE and Pearson. The intent of this training was to provide an overview of the standard setting process and to explain the table leaders' roles during the process. Table leaders served a key role in distributing and collecting materials, monitoring the participation and understanding of tasks from all participants, guiding table-level discussions, and monitoring the use and security of documents throughout the workshop.

Training for table leaders involved a PowerPoint presentation that was an abbreviated version of the presentation for the entire Educator Panel. This presentation for table leaders focused on the standard setting process, the Modified Angoff procedure, and the kinds of feedback provided after rounds of judgments. During training, a table leader folder containing examples of the rating forms, the readiness survey, and the evaluation form was provided to each table leader.

The lead facilitator discussed the agenda and instructed the table leaders to help keep their table members on track to complete the required tasks. The lead facilitator stressed to the table leaders the importance of discussions remaining constructive and relevant and encouraged table leaders to help prevent their tables from straying off track.

Panelist Training

After introductions and other general logistical comments within the committees, the committees reviewed the Achievement Level Descriptions (ALDs) for their particular subject and grade. This task was aimed at determining distinguishing features that separate one achievement level from another. From these distinctions, the committees developed a list of general behaviors and themes for those students who could be described as “just-barely” at a particular achievement level. Each committee performed this task at the table level followed by a discussion with all panelists in order to arrive at a shared and concrete understanding of what the “just-barely” level students at each achievement level should be capable of. The panelists were asked to record three behavioral descriptors at each of their tables. These descriptions were shared across tables and typed and printed for the committee to work with for the remainder of the standard setting workshop. The table leaders helped the facilitator by recording these descriptions at their tables and by actively working to engage all of the panelists in this discussion.

Following this activity, the panelists completed the test for their subject and grade to get an appreciation of the student test-taking experience and to begin making some connections between the test and the descriptions of the “just-barely” level students generated earlier. Once the panelists completed the test, they received answer keys and were given the opportunity to briefly discuss the overall test and to revisit the descriptions of “just-barely” level students at each achievement level as necessary.

The next stage of the meeting was the beginning of the standard setting process itself. The panelists were trained on the Modified Angoff procedure, with opportunities for questions and further clarification as necessary. For the process to succeed, it was critical that the panelists understood the standard setting process and what was required of them. In making their judgments, panelists were asked to respond to the following question:

> Given the knowledge, skills, and abilities that are required in this question, what percentage of students just barely at that achievement level should get this item correct?

Panelists were told that the goal for each achievement level cut was to provide an estimate of the percentage of students who should get this item correct given the description of the minimum level of knowledge and skill required to be considered in that achievement level. To help with this task, facilitators told panelists to visualize a hypothetical group of 100 students at each achievement level. This hypothetical group should comprise students who possess enough knowledge and skills to just make it into a particular achievement level. They are neither highly skilled nor average, but are just barely qualified for that achievement level. And, referencing those 100 hypothetical students, facilitators told panelists to approximate the percentage of those students who should get the item correct. Panelists were instructed to render their judgments (percentages) in increments of 5.

Panelists were instructed to use a recording form and a data entry remote to indicate the percentage of “just-barely” level students who should get the item correct for each of the four achievement level cut points—2, 3, 4, and 5 (students not meeting the requirements of achievement level 2 are classified as level 1). Panelists were allowed to select percentages that reflected their judgments, but the percentages were required to increase from one achievement level to the next as per instructions that were provided by the facilitator. As previously mentioned, panelists were reminded that their judgments could not fall below the chance level of a correct response, which is 25% for multiple-choice items with four response options or 0% for gridded-response or fill-in response items. Each panelist was given an item map, which provided the content standard to which each individual item was linked. Panelists were encouraged to use the item map as a resource during the standard setting process, considering not only the difficulty of each unique item, but also the content the item tested, particularly as it related to expectations of the “just-barely” level students.

Following training on the Modified Angoff method, the panelists engaged in a practice round of making judgments using the recording forms and the data entry remotes. This exercise consisted of making percent-correct judgments based on Achievement Level 3 for 15 released items. Following this activity, the facilitator led the panelists in a discussion of the practice round. Table leaders assisted the facilitator by verifying that each of the panelists at their table accurately completed the practice task. The facilitator reiterated that judgments are made for the “just-barely” level students only. Following the practice round and feedback, the panelists proceeded with the actual rounds of judgments for standard setting. The panelists used test booklets, item information sheets, and data entry remotes throughout the remainder of the standard setting meeting. The test booklets used in the standard setting include only operational items (which count towards student total scores).

Round One

Prior to the first round of judgments, the panelists completed a readiness survey to indicate their level of confidence for making judgments using the procedures identified. The table leaders alerted the facilitator if any panelist was not comfortable with the task so the facilitator could address any outstanding questions or concerns. Once all panelists indicated they were comfortable with the task, they began the first round of standard setting. Panelists were instructed to work independently, making a judgment for each item for four achievement level cuts: Level 2, Level 3, Level 4, and Level 5. In making their judgments, panelists were asked to respond to the following question:

> Given the knowledge, skills, and abilities that are required in this question, what percentage of students just barely at that achievement level should get this item correct?

Panelists were reminded

1. That the percent-correct judgments for an item must increase from Level 2 through Level 5,

2. To completely erase judgments when changing them within rounds on their recording forms,
3. That the minimum allowable percent-correct judgment for a multiple-choice item is 25%, and
4. That the minimum allowable percent-correct judgment for a gridded-response or fill-in response item is 0%.

Once all panelists within a committee completed the first round of judgments, the facilitator worked with the table leaders to check the recording of judgments onto the recording forms and the correct input of judgments using the data entry remotes. The facilitator also checked data entry by monitoring panelists as they entered their information into the remotes. Once all of the panelists completed entering their judgments into the remotes, the facilitator checked the entry of data one more time before closing the session in the data entry software and delivering the completed judgments to the data analysts for verification and analysis.

Round Two

Following the completion of Round One and data processing, the facilitator provided the panelists with the results from the Round One judgments. A brief facilitator-led discussion occurred to describe how the cut scores for the achievement levels were determined at the table level. Each table was provided descriptive statistics (mean, median, mode, min, and max) for panelists' Angoff ratings across achievement levels from Round One.

The table leaders, with guidance from the facilitator, led a discussion of individual judgments. Instead of reviewing all items, table leaders were provided a handout as a guide for which items to discuss. For each achievement level cut, tables discussed items that had the greatest range of disagreement among the panelists at each table. The facilitator also informed the committee that the panelists at the table could discuss any of the items if discussion was requested by a panelist. The table leaders were instructed to start with the Level 2 cut, go through all items flagged for review, and then move to the Level 3, Level 4, and Level 5 cuts until all flagged items were discussed. As items were discussed, panelists considered the following questions:

- How similar were their cut scores and judgments to that of the group (i.e., were there panelists that were more lenient or stringent than the other panelists? If so, why)?
- Did panelists have different conceptualizations of the "just-barely" level students?

Each panelist, in turn, shared his/her reasons for rating the item as he/she did. The table leader led a similar discussion for the item ratings at each achievement level. While panelists were encouraged to reassess their ratings based on these discussions, the main purpose of this activity was to allow panelists to think through and discuss the rating process, not to reach consensus. Panelists were reminded that consensus is not required, but they should discuss differences to get an understanding for why differences exist.

Empirical item difficulty data were provided as process feedback, as a check for widely discrepant ratings by the judges, and to verify that they understood the judgment activity. The

25% most difficult items (in terms of *p*-value) were classified into the high difficulty category, the 25% least difficult items were classified into the low difficulty category, and the remaining 50% of items were classified as medium difficulty. Facilitators guided panelists' understanding of how item difficulty estimates might inform their judgments. For those items that performed significantly differently than expected, panelists were instructed to do the following:

- Remember that the item difficulty estimates provided are based on the total population of students tested, so they are not appropriate for describing the difficulty of a given item for a specific group of students.
- Also consider that empirical item difficulty reflects how students actually performed, but not how they *should* perform, which is the goal of standard setting.
- Compare the percentage assigned for a given item at a given level against the provided item difficulty categorization and think about whether they appear to align given these two populations.
- If the judgment value does not make sense, determine if it should be revised given this new information, and if so, in what way?

After discussing the results from Round One, the panelists were instructed to independently revise their judgments as they felt necessary. Within the instructions, the facilitators reminded the panelists that the judgments were to be made based on "just-barely" level students for each achievement level only, not all students, and that consensus was not required. The panelists indicated their confidence in performing this task by completing the readiness survey for Round Two. Once all panelists indicated on the readiness survey that they were comfortable with beginning Round Two, they provided revised judgments on their recording forms in a separate column and entered their judgments into the data entry remotes. Once all of the panelists completed entering their judgments into the remotes, the facilitator checked the entry of data one more time before closing down the session in the software and delivering judgments to the data analysts for verification and analysis.

Round Three

Once the Round Two judgments were compiled, the panelists received the same type of individual and table-level feedback provided in the previous round, and they also received committee-level feedback as well. The committee-level feedback reports were identical to the table-level reports, but the data contained within these reports were based on all committee members instead of being limited to specific tables.

After sufficient conversation had taken place regarding the results from Round Two, the panelists were instructed to independently revise their judgments during the next round as they felt necessary. Once all panelists indicated on the readiness survey that they were comfortable with beginning Round Three, they provided revised judgments on the recording forms in a separate column and entered their data into the data entry remotes, which followed the same process as the first two rounds.

Round Four

Once the Round Three judgments were compiled, the panelists received the same type of individual, table-level, and committee-level feedback provided in the previous round. For this round, as with the previous rounds, the review of items only occurred for items with the widest ranges of percent-correct judgments. Once these discussions were complete, the panelists were shown impact data—percentages of the spring 2011 testing population classified into each achievement level based on the committee's recommended cut scores—and the impact data from the different grades (i.e., the vertical articulation results). The impact data showed percentages in achievement levels for all students. The vertical articulation results showed the scale cut scores as well as impact data within a subject across grades. The facilitator led the committee in a discussion about the impact data and vertical articulation results. Using these data, the panelists were asked to consider the following questions:

- Given the description of what students should know and be able to do at each achievement level, is that the percentage of students you expect to see in each achievement level?
- Given the results across grades, do your recommended cut scores appropriately align with the expectations set by the other grade levels?

In much the same way that the item difficulty information was intended to validate panelists' item difficulty estimates, the impact data were intended to assist panelists in refining their perception of the examinee population (relative to this assessment) and the achievement level descriptions. The facilitator led the panelists in a discussion concerning the appropriateness of the current cut score recommendations, given the percent of students that would be classified in each level. The facilitator also reminded panelists that, much like the previously discussed empirical item difficulty data, impact data were a reflection of how students actually performed, not how they should perform.

Similar to the impact data provided for their subject and grade, the vertical articulation results provided another source for validating judgments. Panelists were able to see how students were distributed across different grade levels and how this information could be used to inform their judgments.

Once the panelists had discussed the impact data and vertical articulation results, they completed the readiness survey for Round Four. Then they provided their revised judgments in a new column on their recording forms and entered data into the data entry remotes, which followed the same process as the previous rounds.

Round Five

The beginning of this round consisted of a large-group presentation of the individual committee results and impact data. This presentation was in the form of the vertical trends across grades, modified from Round 4. FDOE leadership offered panelists feedback and reactions, reiterating the policies and goals of the FCAT 2.0 and Algebra 1 EOC Assessment programs. During the presentation, FDOE discussed historical impact data for the assessments, external data from the National Assessment of Educational Progress (NAEP), and measures of college readiness.

After the large-group presentation, the panelists reconvened in their respective committee rooms. Each committee facilitator led a discussion on the vertical articulation results, reiterating the feedback and guidance provided by FDOE. After this discussion, the panelists responded to the readiness survey for the final round of judgments, indicating their understanding of the judgment data, impact data, the vertical articulation results, and their confidence in providing the final set of judgments. Once the panelists made their judgments, the final set of percent-correct judgments were submitted for data processing.

The final results, including descriptive statistics for the panelists' ratings, impact data based on the total student population and by subgroup (i.e., gender and ethnicity), vertical articulation across grades, and mean scale scores by achievement level (overall and by gender and ethnicity subgroups), were shown to the individual committees as a concluding step in the standard setting workshop.

Algebra 1 EOC Assessment College Readiness

For the Algebra 1 EOC Assessment committee only, the committee facilitator instructed the panelists to select one of the achievement level cuts that they believe best indicates a student is high achieving and has the potential to meet college-readiness standards by the time the student graduates from high school. Of the 19 Algebra 1 EOC Assessment committee members who participated in this phase of the process, 18 indicated that the Level 3 cut best represents college readiness and one member indicated college readiness is best represented at Level 4. Based on these responses from panelists, Level 3 was determined to be the Educator Panel's recommendation as the best representation of college readiness among the cut scores.

Workshop Wrap-up

After the final recommendations were captured, the panelists in each committee completed evaluation forms on their experience during the workshop, including their confidence in the process and the final cut score recommendations. Feedback from the panelists was documented for the consideration of FDOE and the Reactor Panel.

Final Cut Score Recommendations

Scale cut scores recommended by the Educator are plotted across grades for Reading in Figure 5-1 and for Mathematics in Figure 5-2. For each cut score in these figures, error bars extending ±1 standard error are included as well. Please note that Algebra 1 is not included in the Mathematics vertical scale or associated graphics; the final cuts for Algebra 1 are included at the end of the chapter (see Table 5-5. Final Approved Cut Scores).

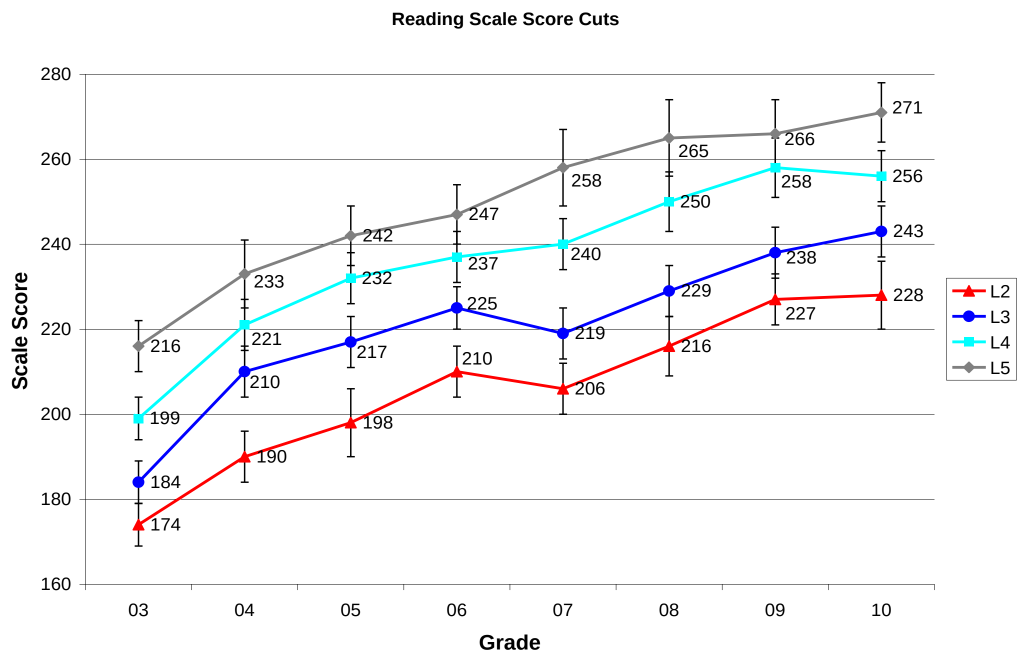

Figure 5-1. Educator Panel: Cut Scores for Reading

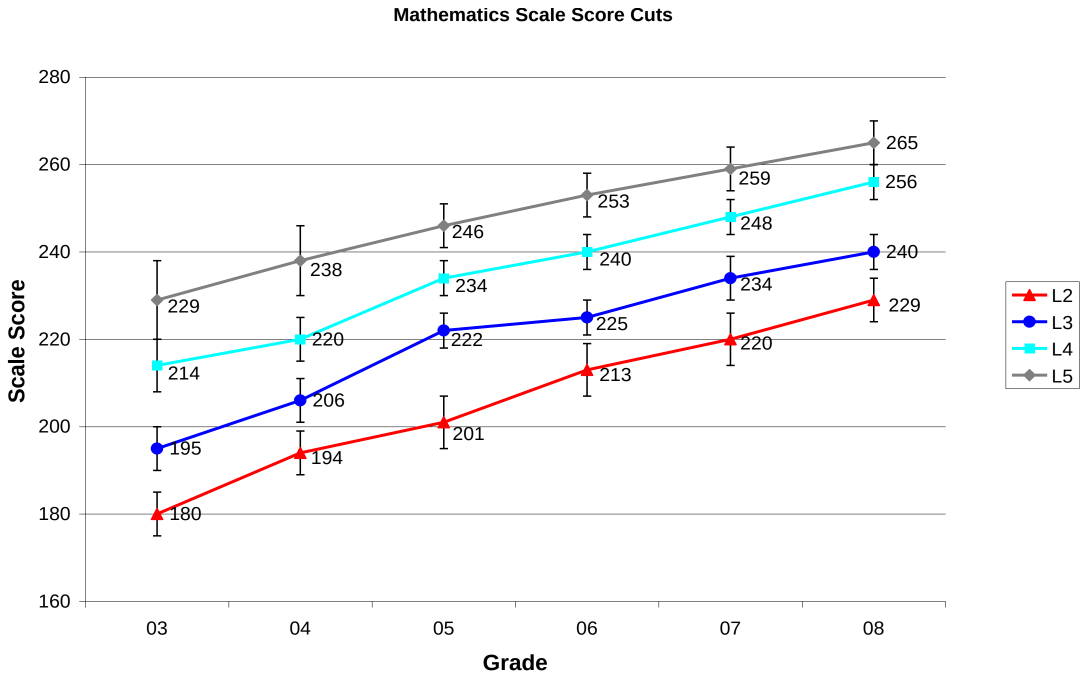

Figure 5-2. Educator Panel: Cut Scores for Mathematics

The percentages of spring 2011 students grouped into the five achievement levels based upon these final cut score recommendations are plotted across grades for Reading in Figure 5-3 and for Mathematics in Figure 5-4. In addition to impact data for the entire student population, impact data for gender and ethnic subgroups were presented to panelists.

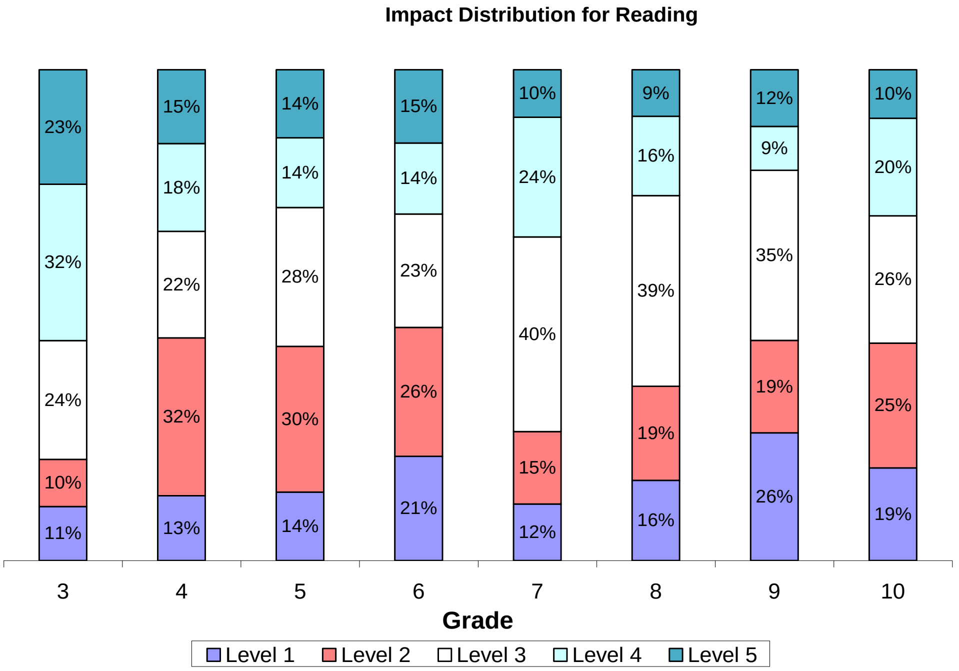

Figure 5-3. Educator Panel: Impact across Grades for Reading

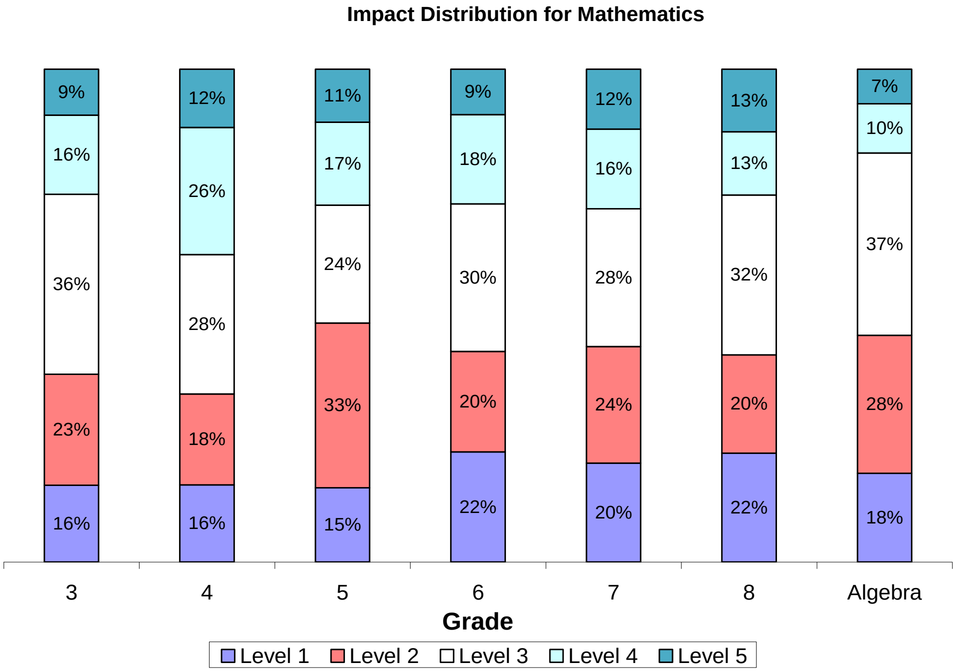

Figure 5-4. Educator Panel: Impact across Grades for Mathematics

Panelist Variability

In order to describe the variability in panelists' judgments, a Generalizability Theory (G-Theory; Brennan, 2001) study was performed. This information was used to determine how similar the cut scores might be if a different set of panelists (although equivalent) or different composition of small groups were used to set cut scores. For this investigation, the sources of variability of interest were panelists, small groups, and rounds. For each cut score, the variance associated with each of these sources was estimated using the maximum likelihood SAS VARCOMP procedure. For this study, the number of rounds was treated as a fixed factor (i.e., 5 rounds in total), meaning that if the standard setting meeting was held again, the same number of rounds would be used. In addition, because panelists discussed all activities in small groups, their judgments were considered dependent on group membership. Therefore, panelists were considered "nested" within tables. Variance components for tables (σ^2_{Tables}) and panelists within tables ($\sigma^2_{Judges:Tables}$) were computed. The computation of the standard errors was done using the following formula (Lee & Lewis, 2001):

$$SE_{cut} = \sqrt{\frac{\sigma^2_{Tables}}{N_{Tables}} + \frac{\sigma^2_{JudgeTable}}{N_{Judges} \cdot N_{Tables}} + \frac{\sigma^2_{Error}}{5N_{Tables} \cdot N_{Judges}}} \qquad (5.1)$$

Because round was treated as a fixed facet, its variance component was not included in the error term. The σ^2_{error} is a confounding term and includes the variance from the interaction between tables and panelists within tables as well as variances unexplained by the defined facets. The sample size in the equation refers to the sample size likely to occur in the Decision Study[2] (D study). Without loss of generality, the sample sizes for the D study were assumed to be the same as the sample size in the Generalizability study (G study). Standard errors were computed for each of the recommended cut scores for each FCAT 2.0 and the Algebra 1 EOC Assessment. For the purposes of this analysis, the recommended cut scores were the scale scores associated with the judgments across panelists derived during standard setting. Different patterns of variance component estimates and hence standard errors for cut scores were anticipated for different cut scores (Lee & Lewis, 2001).

The conditional standard error of measurement (CSEM) for each recommended scale score cut for each FCAT 2.0 and the Algebra 1 EOC Assessment were calculated using the following formula:

$$CSEM = {1}\Big/{\sqrt{I(SS)}} \, . \qquad (5.2)$$

In this formula, *I(SS)* is the amount of psychometric information at a given scale score point. In this case, this is the amount of information at each of the recommended scale score cuts.

[2] A decision study uses information from a G study to design a measurement procedure that minimizes error for a particular purpose.

The standard error of the cut score (SE_{cut}) and the CSEM were used to compute a composite standard error ($SEM_{combined}$), which was calculated using the following formula (Lee & Lewis, 2001):

$$SEM_{combined} = \sqrt{(SE_{cut})^2 + (CSEM)^2}\ . \qquad (5.3)$$

Each of the composite standard error estimates was applied to the corresponding panel-recommended cut score to produce 1, 2, and 3 standard error bands around the cut score. Table 5-4 reports $SEM_{combined}$ at the cut point for each achievement level by subject and grade.

Table 5-4. $SEM_{combined}$ Summary

Subject	Grade	Level 2	Level 3	Level 4	Level 5
Reading	3	5.469	5.144	4.961	6.291
	4	5.952	6.361	6.367	7.688
	5	8.061	6.307	6.324	7.099
	6	5.891	5.237	5.848	7.234
	7	5.715	5.539	6.366	8.725
	8	6.552	5.604	6.592	9.416
	9	6.215	5.989	6.586	7.784
	10	7.658	6.224	5.635	7.174
Mathematics	3	5.171	4.683	5.870	9.228
	4	5.061	4.703	5.482	8.385
	5	5.753	4.250	4.353	5.391
	6	5.903	4.393	3.962	5.202
	7	5.789	4.578	4.384	5.226
	8	5.260	4.467	4.088	4.662
	Algebra 1	11.732	6.832	5.446	5.651

Reactor Panel Meeting #1—2011

Meeting Overview

A panel of community and business leaders was assembled for one and one half days from September 29–30, 2011, in Tallahassee, Florida, to review the Educator Panel's standard setting process, the recommended cut scores, impact data, and external test data, and to provide feedback and recommendations regarding the Educator Panel's recommended cut scores.

Pearson and FDOE jointly facilitated this meeting with Pearson providing information on the standard setting process and results, external tests and test data, and guiding the panel on the process for modifying cut scores.

The Reactor Panel opened with a round of introductions and a discussion of the purpose of the meeting. Next, the Reactor Panel reviewed the methods used during the Educator Panel standard setting workshop and the outcomes of that study. Members of the Reactor Panel completed several different tests similar to the Educator Panel's test-taking experience activity and were provided an opportunity to make general observations about the tests they completed. Following the test-taking activity, the Reactor Panel was provided training on the Modified Angoff procedure and performed a practice Modified Angoff activity.

Members of the Reactor Panel were provided with comparisons between the Florida assessments and several external tests, including NAEP, Stanford 10, PSAT, SAT, PLAN, and ACT. Among the external data that were considered, comparisons to NAEP were highly emphasized, particularly in terms of aligning expectation of the Grades 4 and 8 FCAT 2.0 Reading and Mathematics with the corresponding grades on the NAEP Reading and Mathematics assessments and then making comparisons across grades to assess reasonableness.

Using the Educator Panel's cut scores and the computed SEM_{combined} values, the facilitator presented the Reactor Panel with ranges of scale scores to consider when making its recommendations. The Reactor Panel was encouraged to recommend cut scores that were within ±2 standard errors of the cuts recommended by the Educator Panel, though it could elect to either modify or adopt the recommended cut scores for any subject and grade achievement level.

As the Reactor Panel considered options about retaining or modifying cut scores, live updates were made to graphical displays of the cut scores and impact data to show the Reactor Panel the effect of their changes. While consensus was not required, the majority of the Reactor Panel members supported the final panel recommendations, which are documented in a subsequent section of this report.

Results

The Reactor Panel's recommended cut scores for Achievement Levels 3 and 4 for Grade 3 FCAT 2.0 Reading were slightly outside the ±2 standard error bounds around the Educator Panel's recommended cuts. After the Reactor Panel considered the Educator Panel's cut score recommendations across grades, as shown in Figure 5-1, and the impact of those cuts, as shown in Figure 5-3, the Reactor Panel determined that modifications to the Educator Panel's cuts beyond two standard errors for Achievement Levels 3 and 4 for Grade 3 FCAT 2.0 Reading were warranted to ensure consistency in expectations and outcomes across all grades. All other cut scores recommended by the Reactor Panel were within 2 standard errors of the Educator Panel's recommended cuts. Cut scores recommended by the Reactor Panel are plotted across grades for Reading in Figure 5-5 and for Mathematics in Figure 5-6.

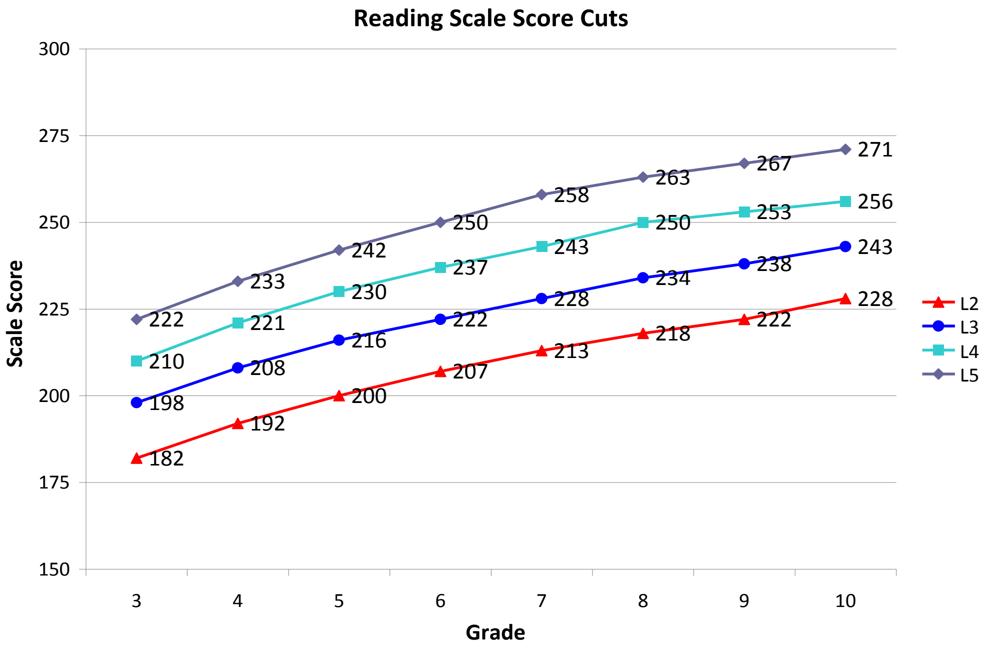

Figure 5-5. Reactor Panel #1: Cut Scores for Reading

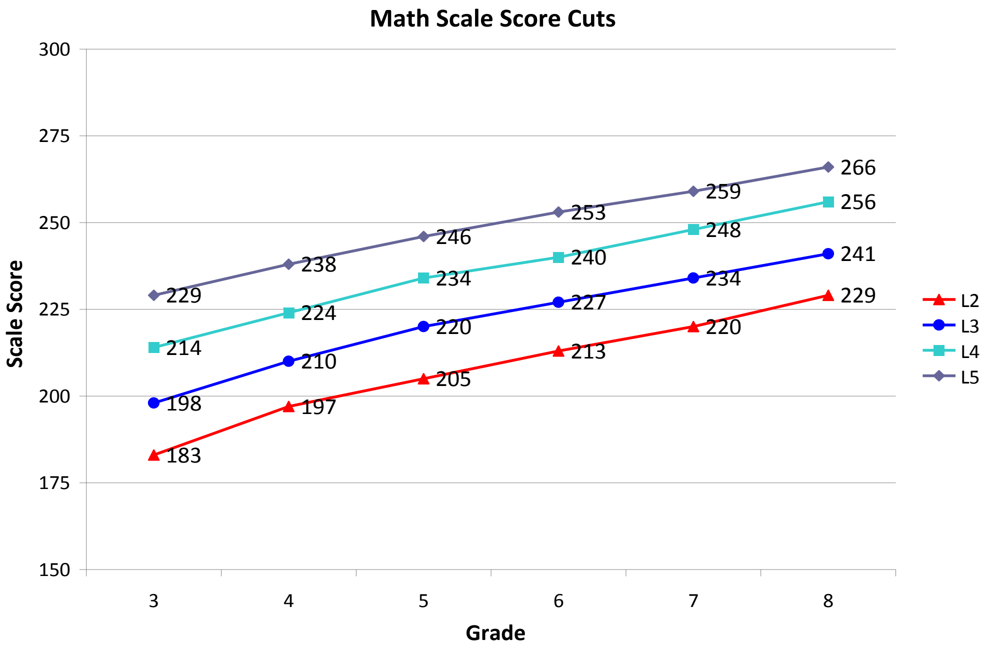

Figure 5-6. Reactor Panel #1: Cut Scores for Mathematics

The percentages of spring 2011 students grouped into the five achievement levels based upon these final cut score recommendations are plotted across grades for Reading in Figure 5-7 and for Mathematics in Figure 5-8. Impact data were also generated for gender and ethnic subgroups.

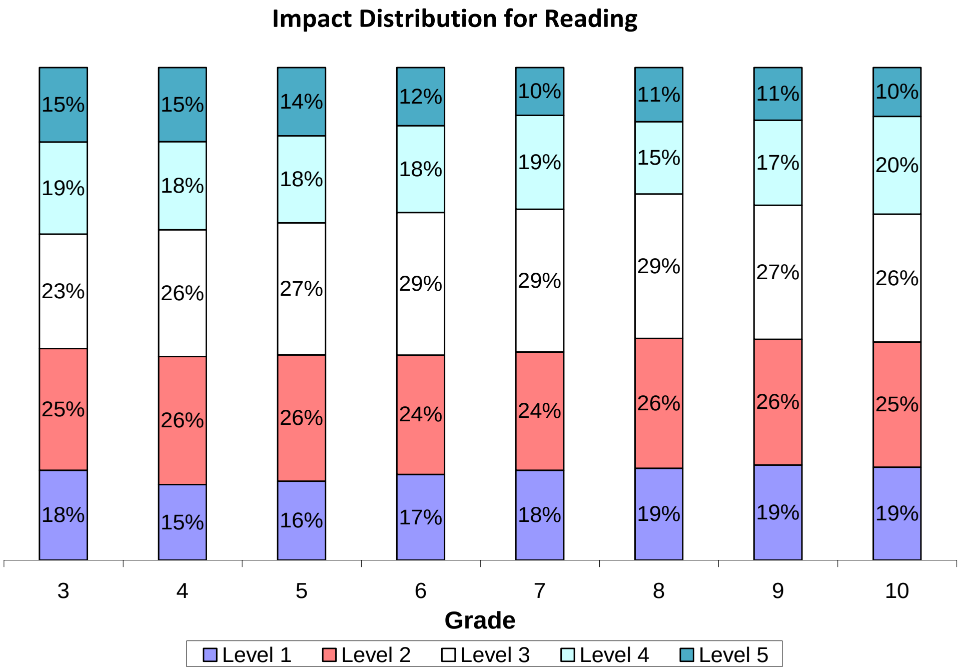

Figure 5-7. Reactor Panel #1: Impact across Grades for Reading

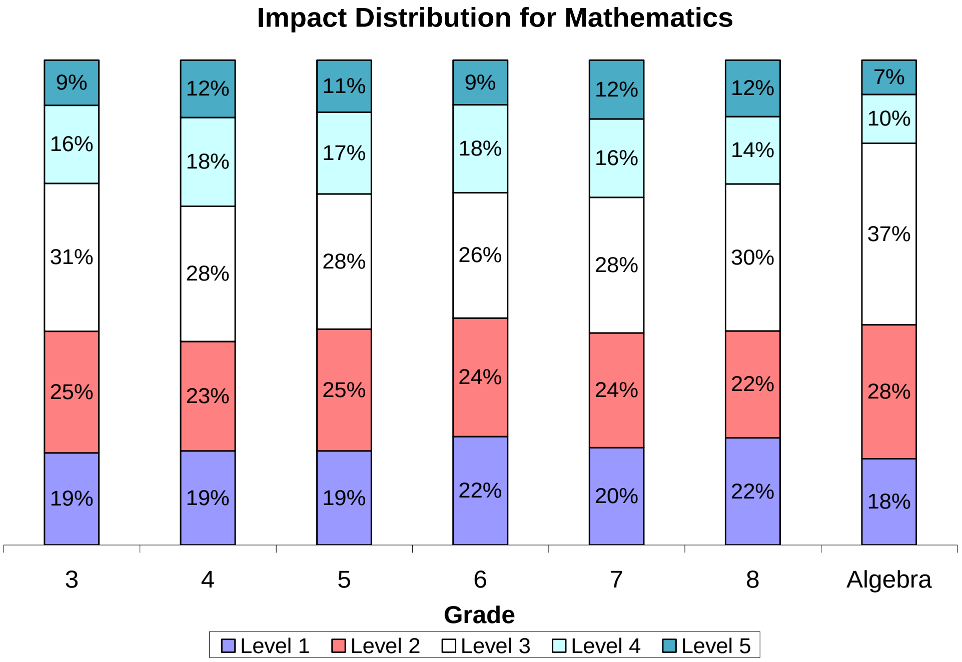

Figure 5-8. Reactor Panel #1: Impact across Grades for Mathematics

Algebra 1 EOC Assessment College Readiness

The Reactor Panel was briefed on the discussions that took place in the Algebra 1 EOC Assessment Educator Panel committee room regarding the college-readiness decision. Of the

Reactor Panel participants, 19 out of 20 agreed with the Educator Panel's selection of Achievement Level 3 as the college-readiness cut, and one individual selected Level 4 as the appropriate college-readiness cut. Overall, the Reactor Panel agreed with the Educator Panel that Achievement Level 3 best indicates that a student is high achieving and has the potential to meet college-readiness standards by the time the student graduates from high school, although the Reactor Panel also recommended that a future research study be conducted to evaluate the efficacy of this decision.

Reactor Panel Meeting #2—2011

Meeting Overview

At the request of the Florida State Board of Education, a second panel of community and business leaders was assembled in Tallahassee, Florida, for a one-day meeting on November 30, 2011, to review the outcomes from the previous phases of the standard setting and to provide additional feedback.

The meeting began with a review of the history of the FCAT Reading and Mathematics assessments' cut scores and student performance on the assessments. Following this review of historical data, the presentation turned to focus on the standard setting activities, including a review of key concepts associated with the chosen methodology and outcomes from the Educator Panel and the first Reactor Panel. Next, the members of this Reactor Panel considered the draft proposed rule to modify cut scores recommended by Florida's Commissioner of Education:

- Set Achievement Level 5 cut scores to the level at which no more than 10% of 2011 students would have achieved Level 5 in all grades and subjects, and
- Make a small adjustment to the Achievement Level 4 cut score in Grade 8 FCAT 2.0 Reading to ensure consistency in expectations and outcomes across all grades in Reading.

To evaluate the appropriateness of these proposed changes, the panel reviewed cut scores recommended by the first Reactor Panel, the new cut scores proposed by the Deputy Commissioner, and impact data associated with each set of cut scores.

Following review of the draft proposed rule, discussion turned to college readiness and its relationship to student performance on the Algebra 1 EOC Assessment. During this portion of the presentation, Florida's definition of college readiness—that a student possess sufficient skills for enrollment into the lowest college-credit-bearing class (Intermediate Algebra or Freshman Composition I)—was reviewed and discussed, as were relevant external data from sources such as the PERT, ACT, SAT, and FCAT assessments. After this presentation, panelists reviewed the Algebra 1 EOC Assessment and held a follow-up discussion. Following these activities, members of the second Reactor Panel rendered their judgments either to endorse the draft proposed rule cut scores or to provide their own recommended adjusted cut scores. They also recommended a score on the Algebra 1 EOC Assessment that indicates a student is high achieving and has the potential to meet college-readiness standards by the time the student graduates from high school. The final recommendations by the second Reactor Panel were informed by the Achievement Level Descriptions, postsecondary readiness competencies, the overall rigor of the test, the difficulty of test questions, and a review of external data. Once the decisions were made, the members of the second Reactor Panel submitted rationales to support their recommendations regarding the draft proposed rule cut scores as well as Algebra 1 EOC Assessment college readiness.

Results

Recommendations regarding cut scores and college readiness were obtained from eight out of nine panelists; one panelist was not able to participate for the full duration of the second Reactor Panel meeting and did not participate in assisting the panel in forming its final recommendations.

Grade 10 FCAT 2.0 Reading

Of the eight remaining individuals from this Reactor Panel who provided recommendations, five recommended a scale score of 243 as the appropriate cut point for Achievement Level 3 on the Grade 10 FCAT 2.0 Reading Assessment, and two individuals recommended a scale score of 245 for this cut point. One of the eight individuals abstained from providing a recommended cut score, citing a lack of external benchmark evidence associating this proposed cut score with the Partnership for Assessment of Readiness for College and Careers (PARCC) as the rationale for this decision.

Algebra 1 EOC

Of the eight remaining panelists, seven selected a scale score of 399 (i.e., the Achievement Level 3 cut score) on the Algebra 1 EOC Assessment as indicating that a student is high achieving and has the potential to meet college-readiness standards by the time of graduation. As with the Grade 10 FCAT 2.0 Reading cut score decision, one member of the Reactor Panel abstained from providing a recommendation regarding college readiness, citing insufficient information as the reason for not providing a recommendation.

State Board of Education

Following conclusion of the Educator and Reactor Panel activities, the Commissioner reviewed the two panels’ recommendations and developed the Department’s recommended cut scores, which reflected both the Educator and Reactor Panels’ recommendations. The State Board of Education convened on December 19, 2011, and was briefed on the standard setting methodology. The State Board of Education approved the Commissioner’s recommended cut scores for FCAT 2.0 Reading and Mathematics and the Algebra 1 EOC Assessment, which were informed by the Educator Panel’s and the two Reactor Panels’ recommendations. Regarding college readiness, the Commissioner recommended that Achievement Level 4 be set as the college readiness cut—an increase from the Educator and Reactor Panels’ recommendation of Achievement Level 3—but noted that scores at Achievement Level 3 indicate that students are on the pathway to college and career readiness. The State Board of Education approved the cut scores recommended by the Commissioner are presented in Table 5-5. These final approved cut scores are graphically presented in Figure 5-9 and Figure 5-10.

Table 5-5. Final Approved Cut Scores

Subject	Grade	Level 2	Level 3	Level 4	Level 5
Reading	3	182	198	210	227
	4	192	208	221	238
	5	200	216	230	246
	6	207	222	237	252
	7	213	228	243	258
	8	218	235	249	264
	9	222	240	253	268
	10	228	245	256	271
Mathematics	3	183	198	214	229
	4	197	210	224	240
	5	205	220	234	247
	6	213	227	240	253
	7	220	234	248	261
	8	229	241	256	268
	Algebra 1	375	399	425	437

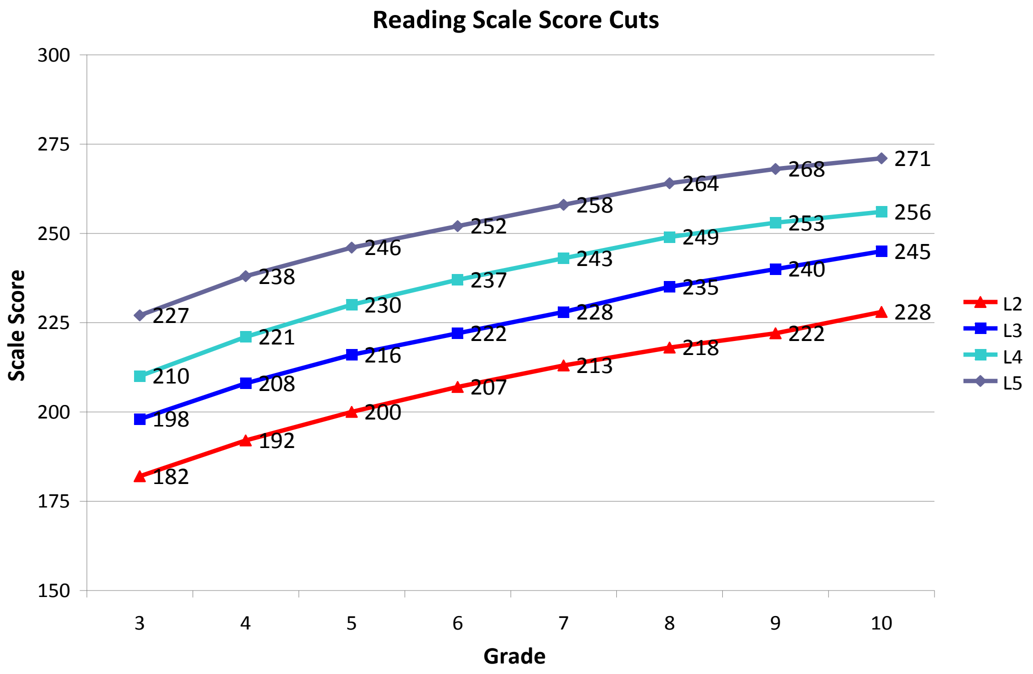

Figure 5-9. Final Approved Cut Scores for Reading

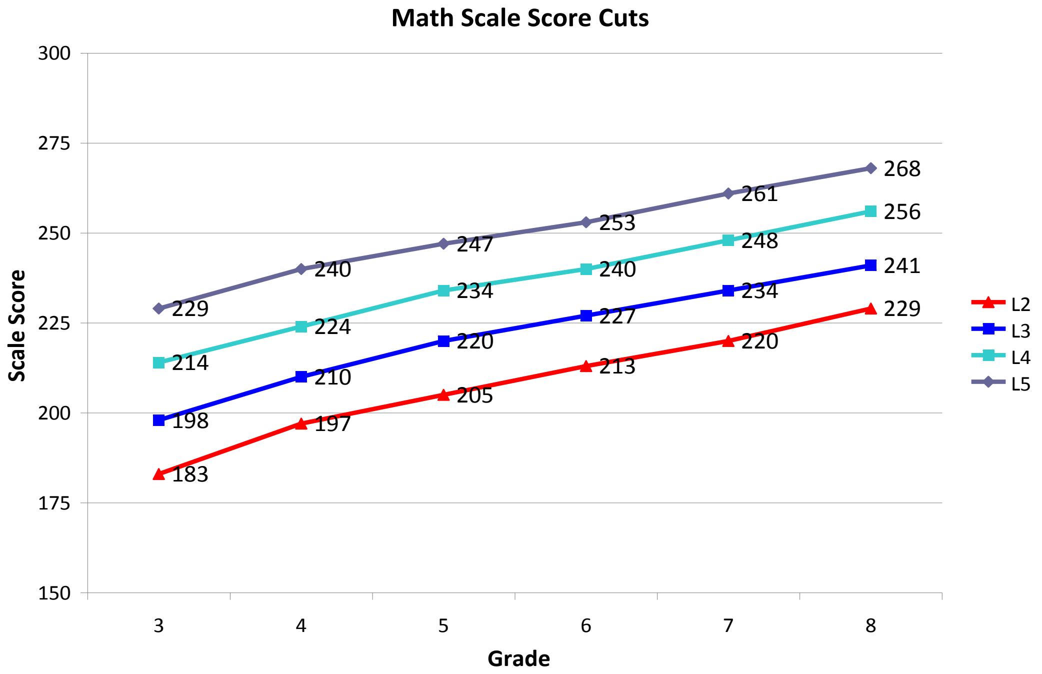

Figure 5-10. Final Approved Cut Scores for Mathematics

The percentages of spring 2011 students grouped into the five achievement levels based upon these final approved cut scores are presented in Table 5-6. Impact data are plotted across grades for Reading in Figure 5-11 and for Mathematics in Figure 5-12.

Table 5-6. Percentage of Spring 2011 Students in Each Achievement Level Based on Final Approved Cut Scores

Subject	Grade	Level 1	Level 2	Level 3	Level 4	Level 5
Reading	3	18%	25%	23%	24%	10%
	4	15%	26%	26%	23%	10%
	5	16%	26%	27%	22%	10%
	6	17%	24%	29%	19%	10%
	7	18%	24%	29%	19%	10%
	8	19%	28%	26%	17%	10%
	9	19%	29%	23%	18%	10%
	10	19%	30%	22%	20%	10%
Mathematics	3	19%	25%	31%	16%	9%
	4	19%	23%	28%	20%	10%
	5	19%	25%	28%	18%	10%
	6	22%	24%	26%	18%	9%
	7	20%	24%	28%	18%	10%
	8	22%	22%	30%	16%	10%
	Algebra 1	18%	28%	37%	10%	7%

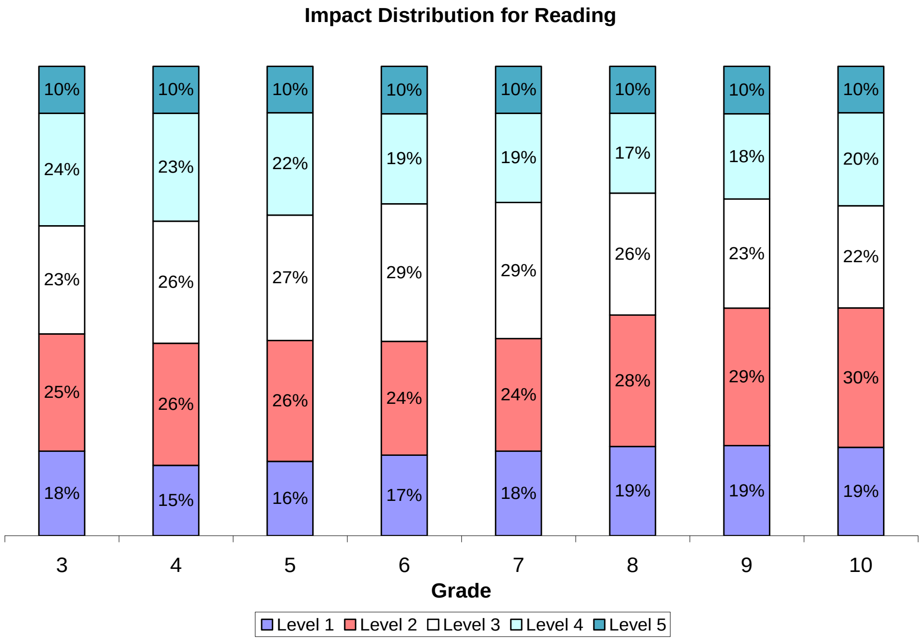

Figure 5-11. Impact across Grades for Reading Based on Final Cut Scores

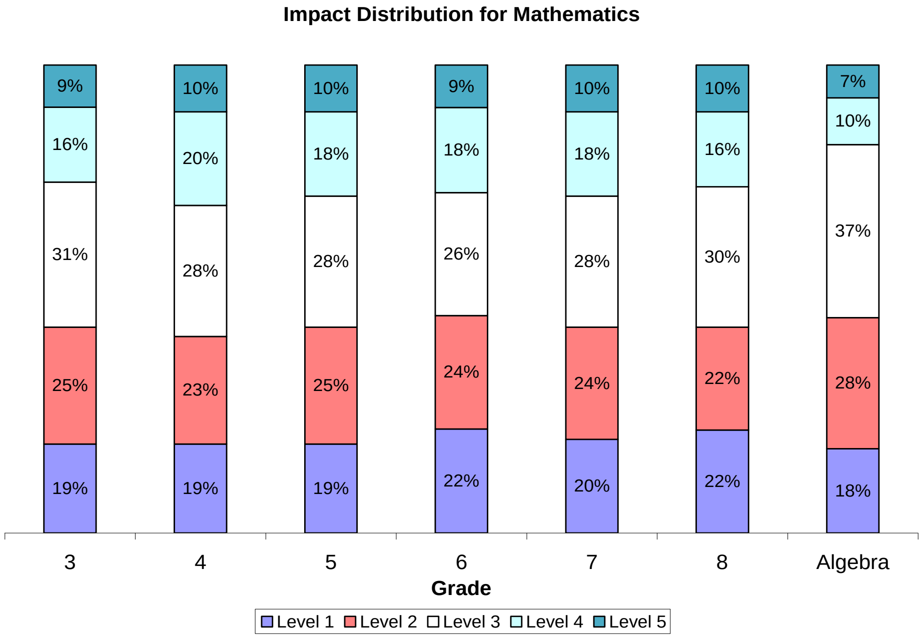

Figure 5-12. Impact across Grades for Mathematics Based on Final Cut Scores

Interim Performance Standards for the 2012 Administration of Science

The operational administration of the FCAT 2.0 Grades 5 and 8 Science Assessments occurred for the first time in spring 2012. Interim performance standards on these assessments were established through a statistical linkage between state results in 2011 and 2012 (see Table 5-7). The procedures used to establish this linkage were reviewed and approved by the FDOE leadership and advisors—including the National Technical Advisory Committee, which includes nationally known psychometricians and policy makers.

The high-level linking process included the following steps:

1. Establish an interim FCAT 2.0 scale using the procedures defined in "Chapter 6. Scaling." Obtain the distribution of reportable interim scores for 2012.
2. Obtain the distribution of reported scores from the 2011 administration of FCAT.
3. Establish a concordance between FCAT and FCAT 2.0 reported scores using the distributions from Steps 1 and 2. See "Chapter 7. Equating."
4. For each student, use the interim FCAT 2.0 score to obtain the FCAT concordant score derived in Step 3. That concordant FCAT score is the reported score for 2012.
5. Assign the published FCAT performance levels to the concordant FCAT score.

Table 5-7. FCAT Scale Score Ranges for Performance Levels on 2012 Science Tests

	Science				
Grade	Level 1	Level 2	Level 3	Level 4	Level 5
5	100–272	273–322	323–376	377–416	417–500
8	100–269	270–324	325–386	387–431	432–500

The achievement level reports in the 2012 yearbook provided the percentage of students by achievement level. Summaries of the scale score distribution can be found in the scale distribution and statistics reports of that yearbook. Because equipercentile linking was used in 2012, the achievement level percentages and the scale score distributions were very similar to those obtained in 2011.

Setting Performance Standards—2012

After the interim standards for FCAT 2.0 Grades 5 and 8 Science Assessments were applied to 2012 results, formal standard setting meetings were undertaken to set cuts that would be applied to the results of the 2013 administration and subsequent administrations. From September 18 through September 21, 2012, several panels of Florida educators—collectively known as the Educator Panel—were convened in Tallahassee, Florida, to make content-based recommendations for cut scores for the FCAT 2.0 Grades 5 and 8 Science as well as the Biology 1 and Geometry EOC Assessments. A separate panel of superintendents, business leaders, and other community stakeholders—Reactor Panel —was convened September 27 through September 28, 2012, in Tallahassee, Florida, to review the outcomes from the Educator Panel meeting and to form their own set of recommended cut scores for the FCAT 2.0 Grades 5 and 8 Science and the Biology 1 and Geometry EOC Assessments. A total of 75 Florida educators participated in the Educator Panel meeting, and a total of 20 community stakeholders participated in the Reactor Panel meeting.

FCAT 2.0 Science and the Biology 1 and Geometry EOC Assessments are statewide tests. FCAT 2.0 Science is currently administered in grades 5 and 8, the Biology 1 EOC Assessment is administered at the end of the Biology 1 (or equivalent) course, and the Geometry EOC Assessment is administered at the end of the Geometry (or equivalent) course. These testing programs assess student achievement on the NGSSS. FCAT 2.0 Science is comprised entirely of multiple-choice items, the Biology 1 EOC Assessment is also comprised entirely of multiple-choice items, and the Geometry EOC Assessment is comprised of multiple-choice and fill-in response items.

The standard setting for FCAT 2.0 Science and the Biology 1 and Geometry EOC Assessments generally followed the same process as standard setting of the FCAT 2.0 and Algebra 1 EOC Assessment in 2011 (see the prior section of this chapter for details). Features specific to 2012 standard setting are outlined below:

- To improve the precision of cut score estimation, panelists were instructed to render their judgments (percentages) in increments of 1 instead of 5.
- The practice round included 6-12 released items.
- FCAT 2.0 Science and Florida EOC assessments were not on vertical scales, so vertical articulation did not apply to 2012 standard setting. Four rounds of judgments were conducted in 2012 instead of five rounds.
    - Once the Round Three judgments were compiled, the panelists received the same type of individual, table-level, and committee-level feedback provided in the previous round. Following table-level and committee-level discussions, the panelists were shown impact data—percentages of the spring 2012 testing population classified into each achievement level based on the committee's recommended cut scores.
    - Following Round Four judgments, the final results, including descriptive statistics for the panelists' scale score cuts, committee's recommended cut scores, impact data based on the total student population and by subgroup (i.e., gender and ethnicity), and mean scale scores by achievement level (overall and by gender and ethnicity subgroups), were shown to the individual committees as a concluding piece of the standard setting workshop.

Panelists

As shown in the following tables, there were 4 separate committees (Grade 5 FCAT 2.0 Science, Grade 8 FCAT 2.0 Science, Biology 1 EOC Assessment, and Geometry EOC Assessment) for a total of 75 Educator Panel Participants. All panelists were asked to provide voluntary demographic information. The Educator Panel participants' professional backgrounds are summarized in

Table 5-8.

Table 5-8. Educator Panel: Percentages of Panelists by Professional Background

Grade/Subject	Number of Panelists	TCH (%)	COA (%)	SPC (%)	ADM (%)	OTH (%)
Science Grade 5	19	42	42	5	11	11
Science Grade 8	18	50	17	33	6	0
Biology 1 EOC	18	67	0	22	11	0
Geometry EOC	20	35	5	25	30	10

Note. TCH=Teacher, COA=Coach, SPC=Specialist, ADM=Administrator, OTH=Other. Some participants indicated multiple professional backgrounds in their responses, so the percentages may exceed 100%.

The Educator Panel participants' teaching experience is summarized in Table 5-9. This table lists the minimum, mean, and maximum values for both overall teaching experience and experience within the panelists' assigned grade and subject for each committee.

Table 5-9. Educator Panel: Teaching Experience

	Total Yrs. Experience			Yrs. In This Grade/Subj.		
Grade/ Subject	Minimum	Mean	Maximum	Minimum	Mean	Maximum
Science Grade 5	1	13.8	32	1	8.9	32
Science Grade 8	1	14.8	39	1	11.7	39
Biology 1 EOC	1	14.9	29	1	12.0	26
Geometry EOC	1	16.5	37	1	12.6	37

Final Cut Score Recommendations

Scale cut scores are plotted for grades 5 and 8 Science in Figure 5-13. Scale cut scores for Biology 1 and Geometry EOC Assessments are plotted in Figure 5-13Figure 5-14. For each cut score in these figures, error bars extending +/- 1 standard error are included as well.

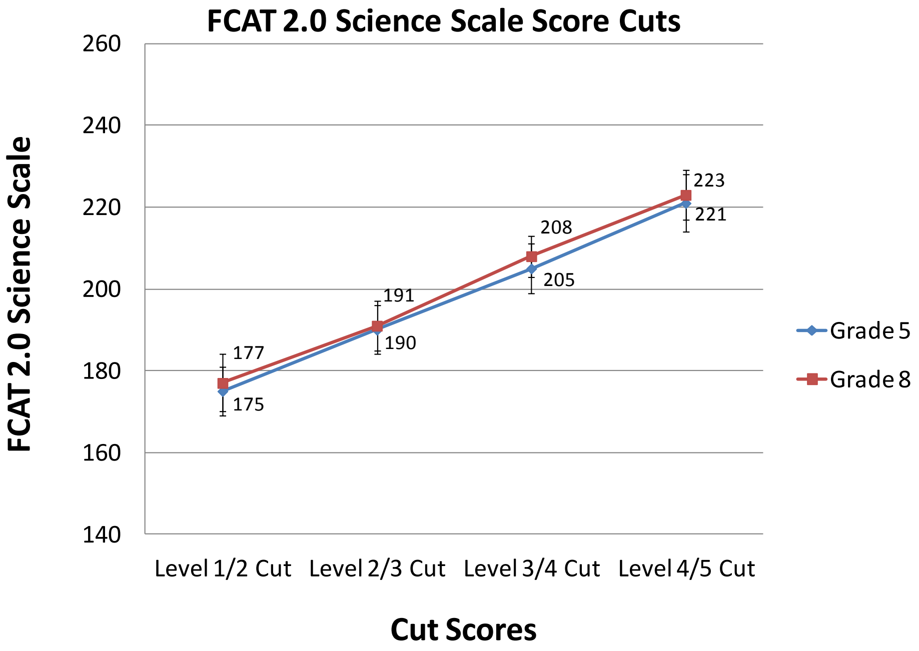

Figure 5-13. Educator Panel: Cut Scores for Science

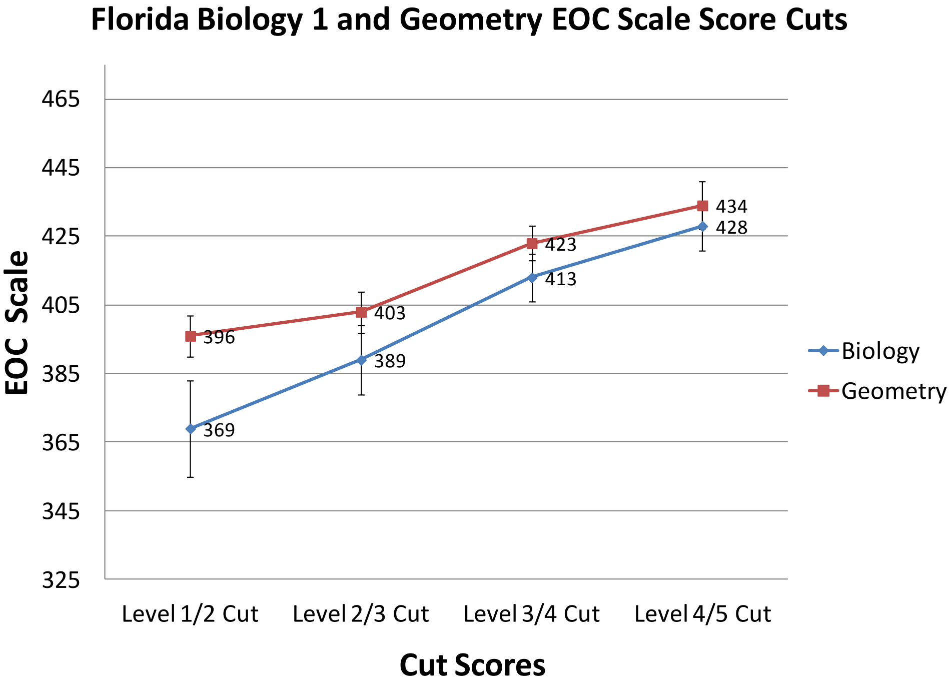

Figure 5-14. Educator Panel: Cut Scores for Biology 1 and Geometry EOC

The percentages of spring 2012 students grouped into the five achievement levels based upon these final cut score recommendations are plotted for FCAT 2.0 Science, Biology 1 EOC Assessment, and Geometry EOC Assessment in Figure 5-15.

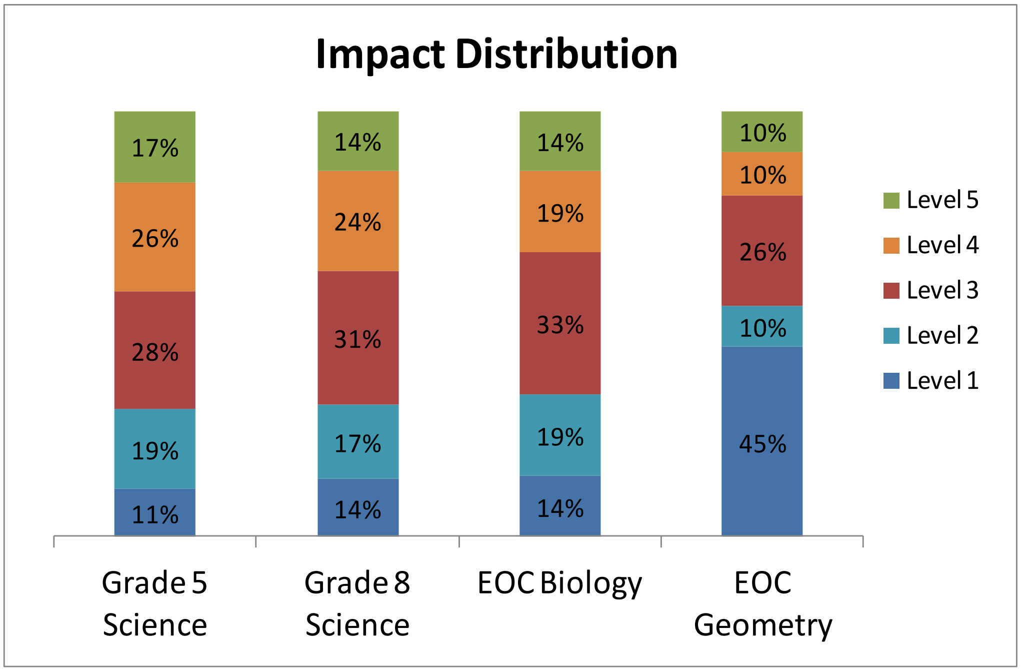

Figure 5-15. Educator Panel: Impact Distribution for FCAT 2.0 Science and the Biology 1 and Geometry EOC Assessments

Panelist Variability

The panelist variability was examined using the same procedure as in 2011 standard setting. The only change is the number of rounds, which is 4 for 2012. The formula below was changed accordingly to calculate the variability:

$$SE_{cut} = \sqrt{\frac{\sigma^2_{Tables}}{N_{Tables}} + \frac{\sigma^2_{JudgeTable}}{N_{Judges} \bullet N_{Tables}} + \frac{\sigma^2_{Error}}{4N_{Tables} \bullet N_{Judges}}}$$

Table 5-10 below reports *SEM*$_{\text{combined}}$ at each cut point for every grade and subject.

Table 5-10. *SEM*$_{\text{combined}}$ Summary

Assessment	Level 1/2	Level 2/3	Level 3/4	Level 4/5
Science Grade 5	6.497	5.602	5.649	7.055
Science Grade 8	7.078	5.767	5.320	6.477
Biology 1 EOC	13.576	9.796	6.974	7.284
Geometry EOC	6.384	5.785	5.134	7.060

Reactor Panel Meeting—2012

Meeting Overview

A panel of community and business leaders was assembled for one and one half days from September 27 through September 28, 2012, in Tallahassee, Florida, to review the Educator Panel's standard setting process, the recommended cut scores, impact data, and external test data, and to provide feedback and recommendations regarding the Educator Panel recommended cut scores. The Reactor Panel meeting followed the same process as was used in 2011.

Members of the Reactor Panel were provided with comparisons between the Florida assessments and several external tests (e.g., NAEP Science, PSAT, SAT, PLAN, and ACT), highlighting relevant interpretations and the unique information that each set of data provides. Historical Florida trend data for the FCAT Science tests were also reviewed by the Reactor Panel as part of evaluating the reasonableness of the cut scores.

Results

The Reactor Panel's recommended cut score for Achievement Level 2 for Geometry EOC was slightly outside the +/- 2 standard error bounds around the Educator Panel's recommended cuts. After the Reactor Panel considered the Educator Panel's cut score recommendations, as shown in
Figure 5-14, and the impact of those cuts, as shown in Figure 5-15, the Reactor Panel determined that modification to the Educator Panel's cut beyond two standard errors for Achievement Level 2 of Geometry EOC was warranted to ensure consistency in expectations and outcomes across EOC subjects. All other cut scores recommended by the Reactor Panel were within 2 standard errors of the Educator Panel's recommended cuts. Cut scores recommended by the Reactor Panel are plotted for FCAT 2.0 Grade 5 and 8 Science in Figure 5-16 and for Biology 1 and Geometry EOC in Figure 5-17.

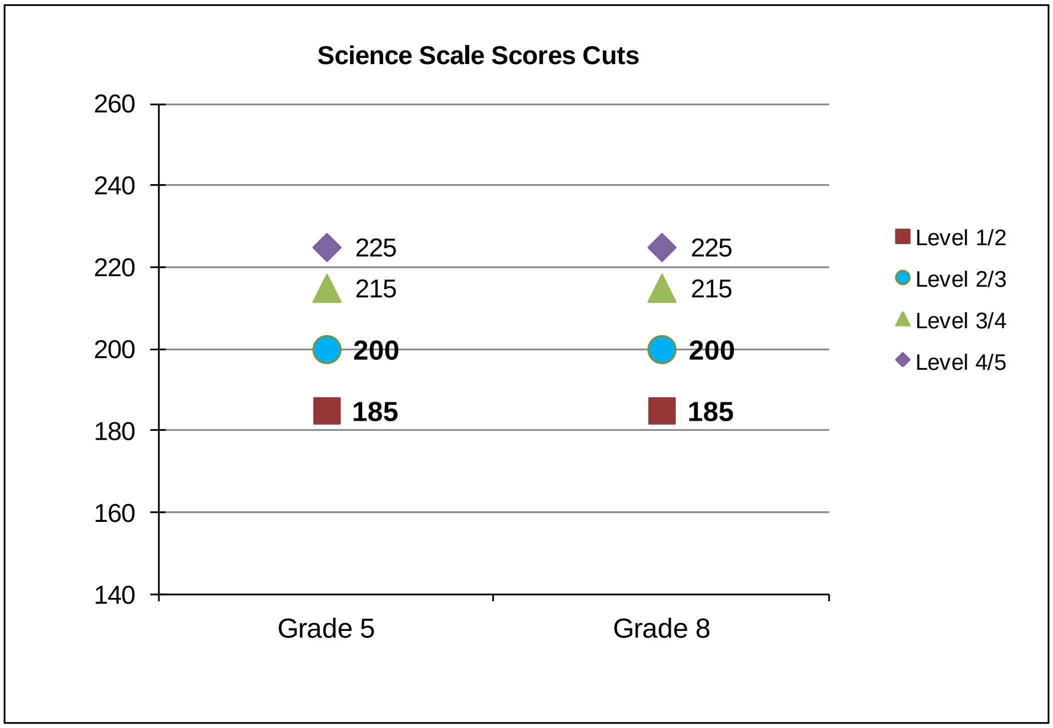

Figure 5-16. Reactor Panel: Cut Scores for Science

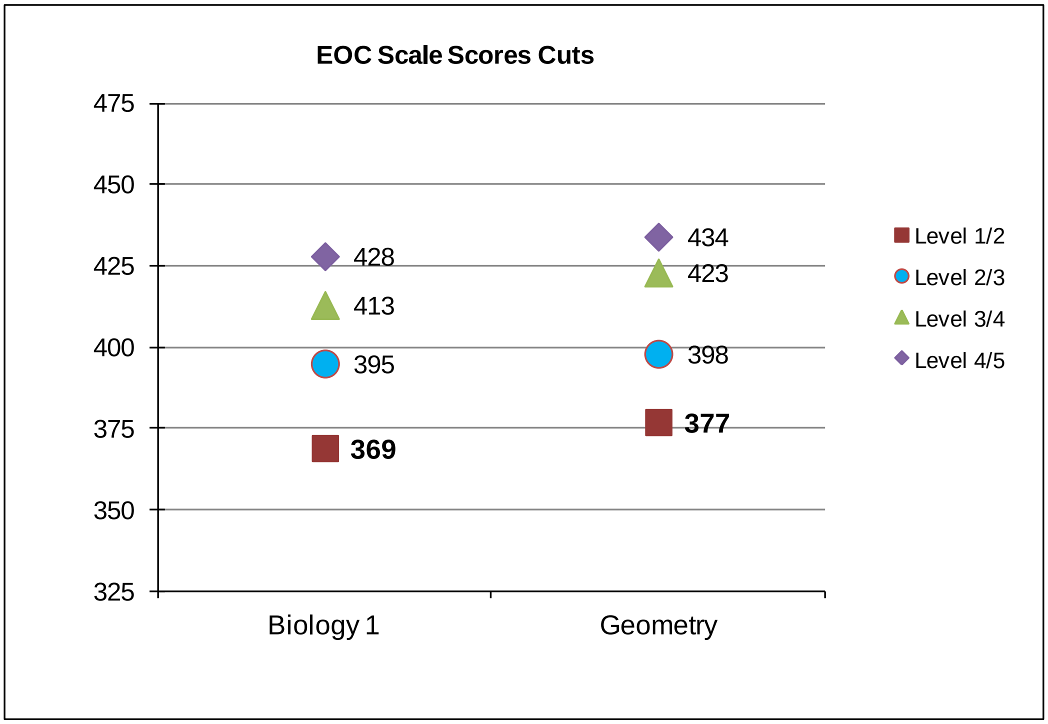

Figure 5-17. Reactor Panel: Cut Scores for Biology 1 and Geometry EOC

The percentages of spring 2012 students grouped into the five achievement levels based upon these final cut score recommendations are plotted in Figure 5-18.

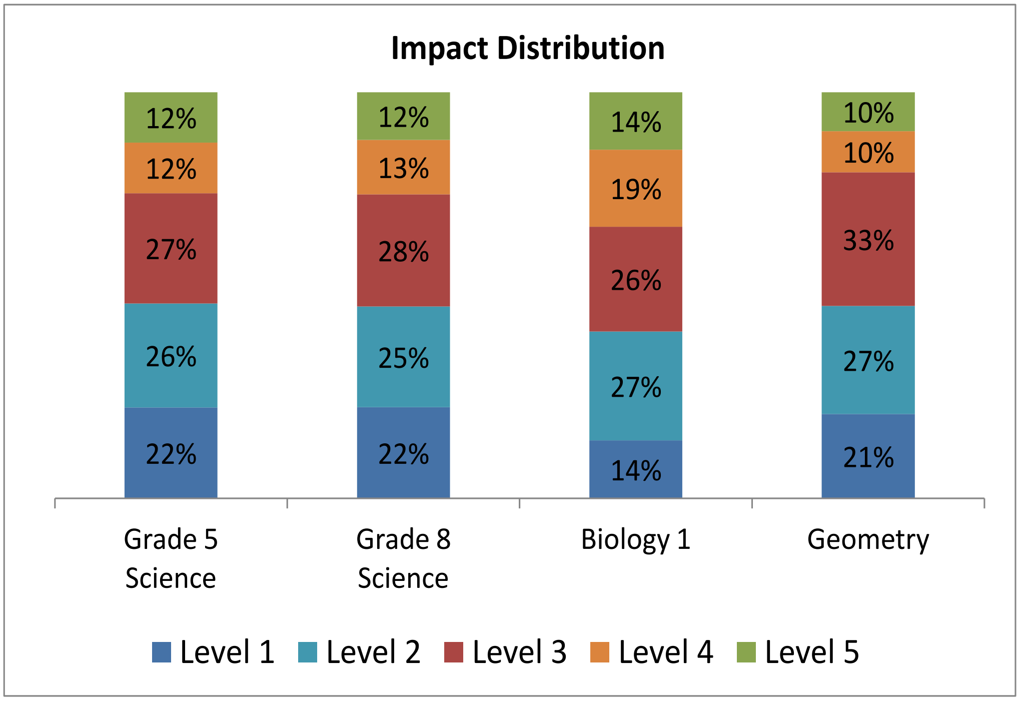

Figure 5-18. Reactor Panel: Impact Distribution

State Board of Education

Following the conclusion of the Educator and Reactor Panel activities, the Commissioner reviewed the two panels' recommendations and developed the Department's recommended cut scores, which reflected both the Educator and Reactor Panels' recommendations and public input. The State Board of Education convened on December 12, 2012, and was briefed on the standard setting methodology. The State Board of Education approved the Commissioner's recommended cut scores for FCAT 2.0 Science and the Biology 1 and Geometry EOC Assessments. Final approved cut scores are shown in Table 5-11 and Figure 5-20.

Table 5-11. Final Approved Cut Scores

Assessment	Level 1/2	Level 2/3	Level 3/4	Level 4/5
Science Grade 5	185	200	215	225
Science Grade 8	185	203	215	225
Biology 1 EOC	369	395	421	431
Geometry EOC	370	396	418	434

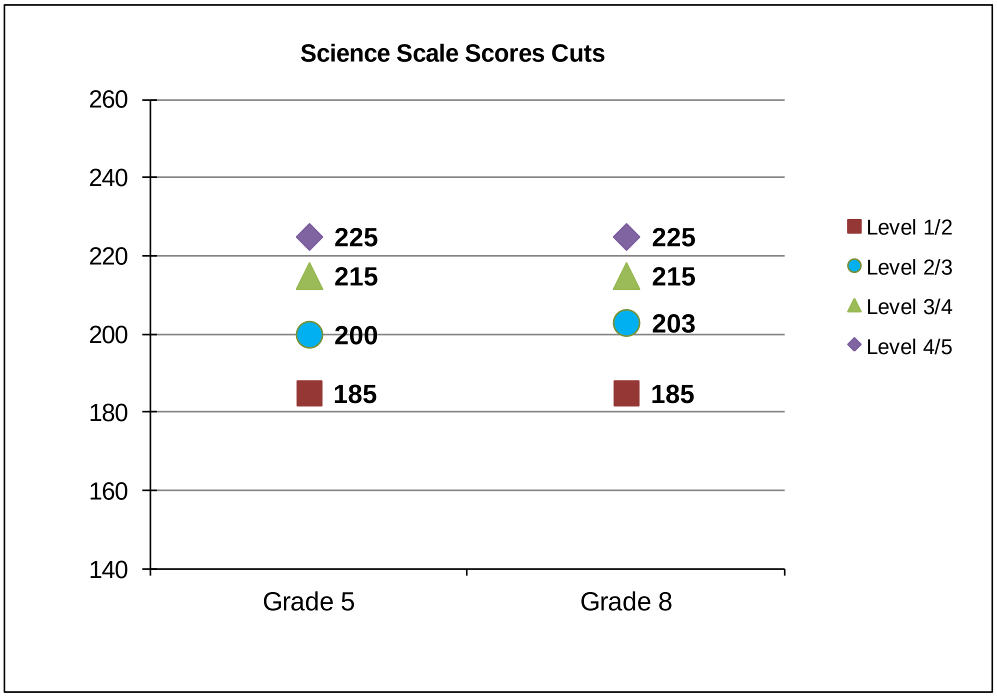

Figure 5-19. Final Approved Cut Scores for Science

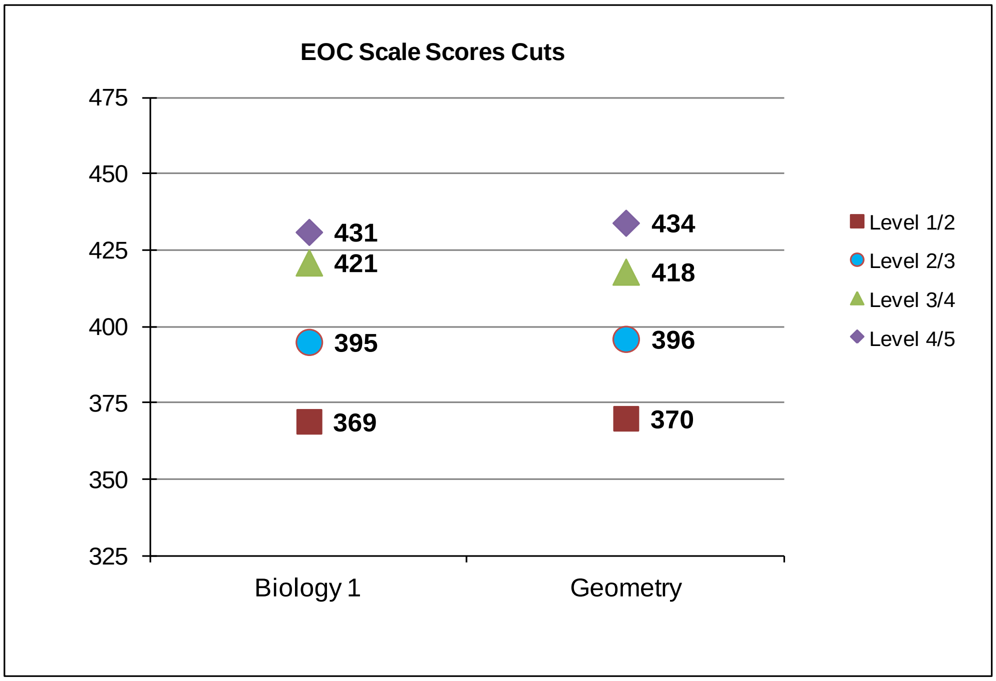

Figure 5-20. Final Approved Cut Scores for Biology 1 and Geometry EOC

The percentages of spring 2012 students grouped into the five achievement levels based upon these final approved cut scores are presented in Table 5-12. Impact data are plotted in Figure 5-21.

Table 5-12. Percentage of Spring 2012 Students in Each Achievement Level Based on Final Approved Cut Scores

Assessment	***Level 1***	***Level 2***	***Level 3***	***Level 4***	***Level 5***
Science Grade 5	22	26	27	12	12
Science Grade 8	22	31	22	13	12
Biology 1 EOC	14	27	37	11	11
Geometry EOC	15	30	30	16	10

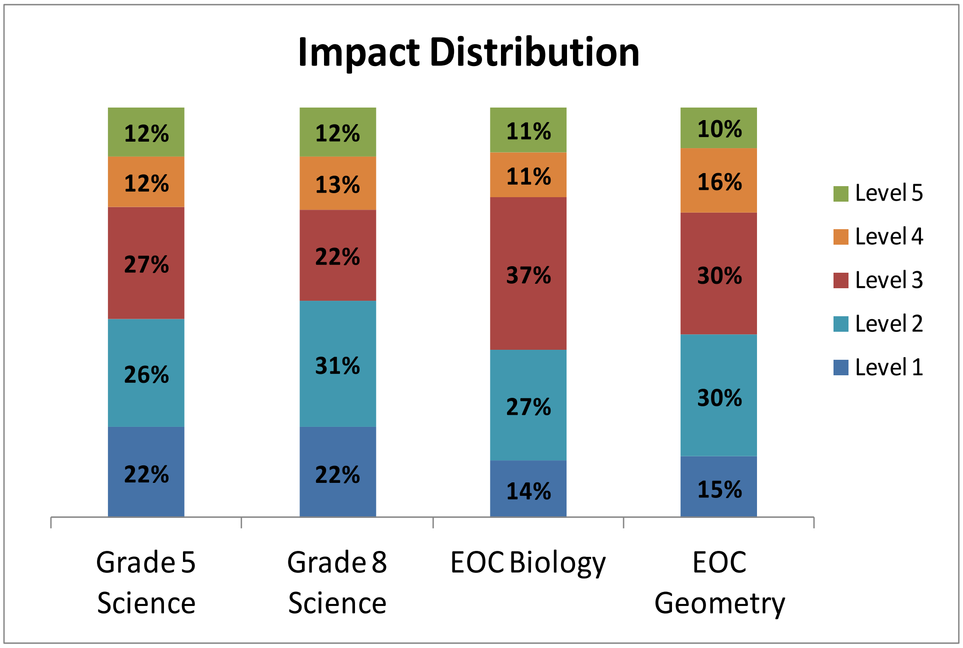

Figure 5-21. Impact Distribution Based on Final Approved Cut Scores

Setting Performance Standards—2013

After the interim standards for the U.S. History EOC Assessment were applied to 2013 results, formal standard setting meetings were undertaken to set cuts that would be applied to the results of the 2014 administration and subsequent administrations. From August 13 through August 16, 2013, several panels of Florida educators—collectively known as the Educator Panel—were convened in Orlando, Florida, to make content-based recommendations for cut scores for the U.S. History EOC Assessment. A separate panel of superintendents, business leaders, and other community stakeholders—Reactor Panel —was convened August 22 through August 23, 2013, in Orlando, Florida, to review the outcomes from the Educator Panel meeting and to form their own set of recommended cut scores for the U.S. History EOC Assessment. A total of 26 Florida educators participated in the Educator Panel meeting, and a total of 15 community stakeholders participated in the Reactor Panel meeting.

The U.S. History EOC Assessment is a statewide test that is administered at the end of the U.S. History (or equivalent) course. This testing program assesses student achievement on the NGSSS. The U.S. History EOC Assessment is comprised entirely of multiple-choice items.

The standard setting for the U.S. History EOC Assessment followed the same process as the standard setting of FCAT 2.0 Science, Biology 1, and Geometry EOC Assessments in 2012 (see the prior section of this chapter for details).

Panelists

There was one committee with 26 Educator Panel Participants. All panelists were asked to provide voluntary demographic information. The Educator Panel participants' professional backgrounds are summarized in Table 5-13.

Table 5-13. Educator Panel: Percentages of Panelists by Professional Background

Grade/Subject	Number of Panelists	TCH (%)	COA (%)	SPC (%)	ADM (%)	OTH (%)
U.S. History EOC	26	65	15	15	23	0

Note. TCH=Teacher, COA=Coach, SPC=Specialist, ADM=Administrator, OTH=Other. Some participants indicated multiple professional backgrounds in their responses, so the percentages may exceed 100%.

Table 5-9 presents a summary of the Educator Panel participants' teaching experience, including both overall teaching experience and experience teaching U.S. History.

Table 5-14. Educator Panel: Teaching Experience

	Less than 1 Year (%)	1-5 Years (%)	6-10 Years (%)	11-15 Years (%)	16-20 Years (%)	More than 20 Years (%)	Not Reported (%)
Teaching	0	8	19	19	31	23	0
Teaching U.S. History	0	42	23	12	19	0	4

Final Cut Score Recommendations

Scale cut scores are plotted for the U.S. History EOC Assessment in **Figure 5-22**. For each cut score in this figure, error bars extending +/- 1 standard error are included as well.

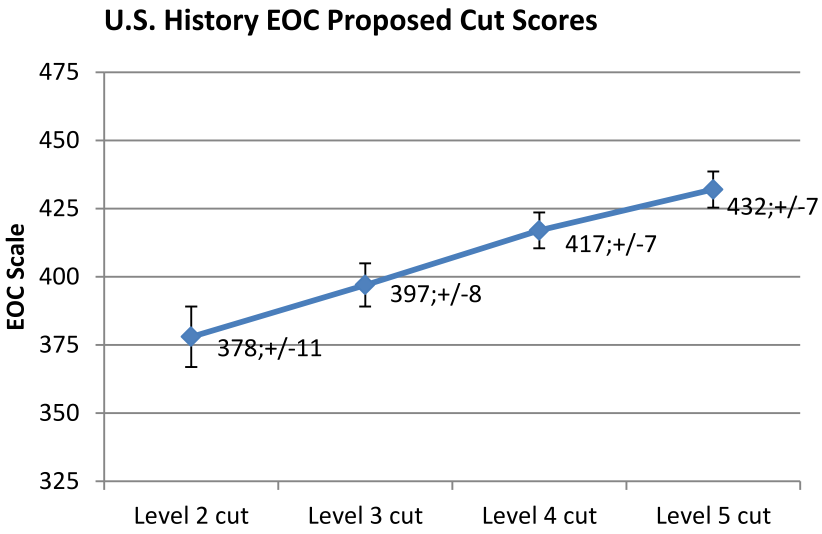

Figure 5-22. Educator Panel: Cut Scores for U.S. History EOC

The percentages of spring 2013 students grouped into the five achievement levels based upon these final cut score recommendations are plotted in Figure 5-23.

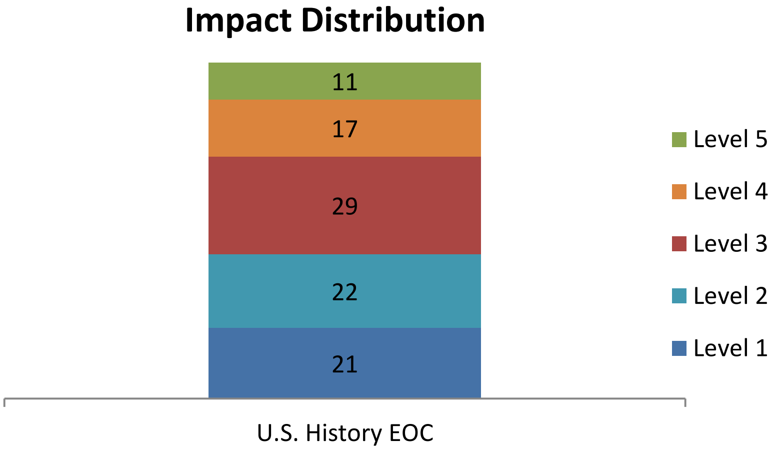

Figure 5-23. Educator Panel: Impact Distribution for U.S. History EOC

Panelist Variability

Table 5-15 below reports *SEM*$_{\text{combined}}$ at each cut point for the U.S. History EOC.

Table 5-15. *SEM*$_{\text{combined}}$ Summary

Assessment	Level 2	Level 3	Level 4	Level 5
U.S. History EOC	11.129	7.880	6.555	6.629

Reactor Panel Meeting—2013

Meeting Overview

A panel of community and business leaders was assembled for one and one half days from August 22 through August 23, 2013, in Orlando, Florida, to review the Educator Panel's standard setting process, the recommended cut scores, impact data, and external test data, and to provide feedback and recommendations regarding the Educator Panel recommended cut scores. The Reactor Panel meeting followed the same process as was used in 2012.

Members of the Reactor Panel were provided with comparisons between the Florida assessments and several external tests (e.g., NAEP U.S. History, Advanced Placement [AP] U.S. History, Scholastic Aptitude Test [SAT] U.S. History), highlighting relevant interpretations and the unique information that each set of data provides. Historical Florida trend data for the Grade 10 FCAT 2.0 Reading assessment and other Florida EOC Assessments were also reviewed by the Reactor Panel as part of evaluating the reasonableness of the cut scores.

Results

All the cut scores recommended by the Reactor Panel were the same as the cut scores recommended by the Educator Panel. Cut scores recommended by the Reactor Panel are plotted in Figure 5-24.

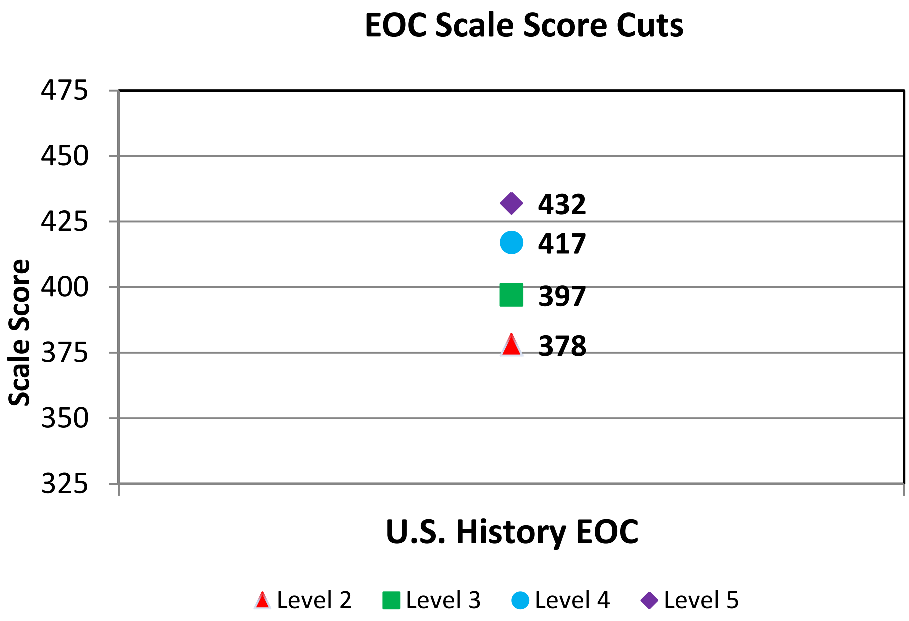

Figure 5-24. Reactor Panel: Cut Scores for U.S. History EOC

The percentages of spring 2013 students grouped into the five achievement levels based upon these final cut score recommendations are plotted in Figure 5-25.

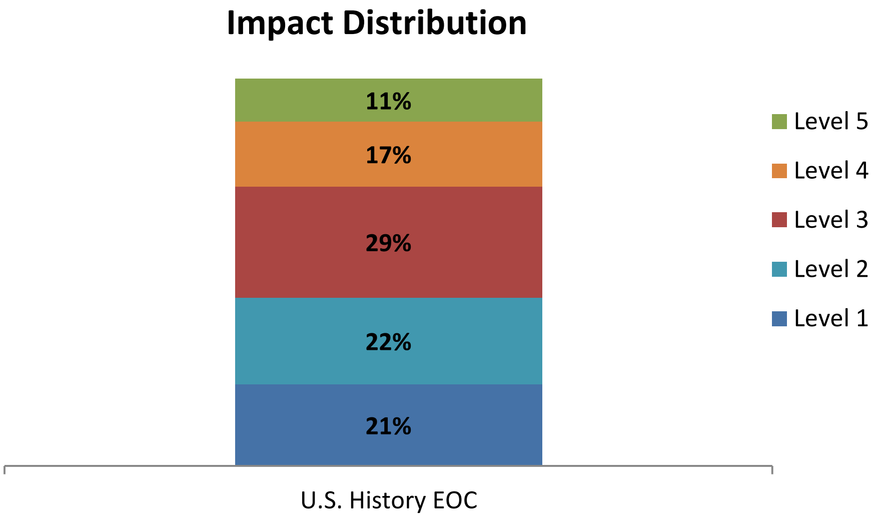

Figure 5-25. Reactor Panel: Impact Distribution

State Board of Education

Following the conclusion of the Educator and Reactor Panel activities, the Commissioner reviewed the two panels' recommended cut scores, which were the same for each of the five Achievement Levels. The Commissioner also reviewed public input received during the rule development workshops held September 3-5, 2013, and recommended that the cut scores recommended by both the Educator and Reactor Panels be implemented in rule.

The State Board of Education convened on October 15, 2013, and was briefed on the standard setting methodology. The State Board of Education approved the Commissioner's recommended cut scores for the U.S. History EOC Assessment. Final approved cut scores are shown in Table 5-16 and Figure 5-26.

Table 5-16. Final Approved Cut Scores

Assessment	Level 2	Level 3	Level 4	Level 5
U.S. History EOC	378	397	417	432

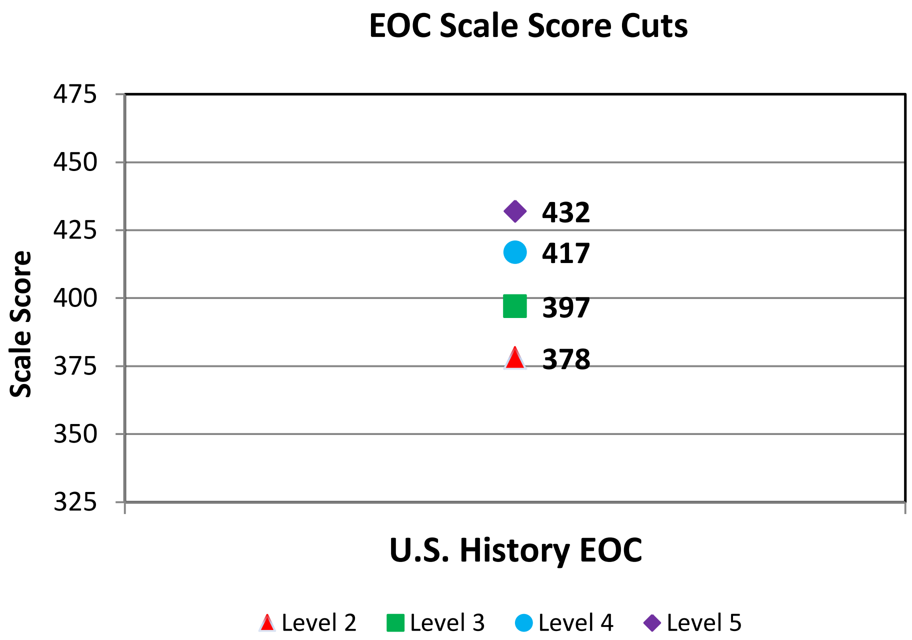

Figure 5-26. Final Approved Cut Scores for U.S. History EOC

The percentages of spring 2013 students grouped into the five achievement levels based upon these final approved cut scores are presented in Table 5-17. Impact data are plotted in Figure 5-27.

Table 5-17. Percentage of Spring 2013 Students in Each Achievement Level Based on Final Approved Cut Scores

Assessment	Level 1	Level 2	Level 3	Level 4	Level 5
U.S. History EOC	21	22	29	17	11

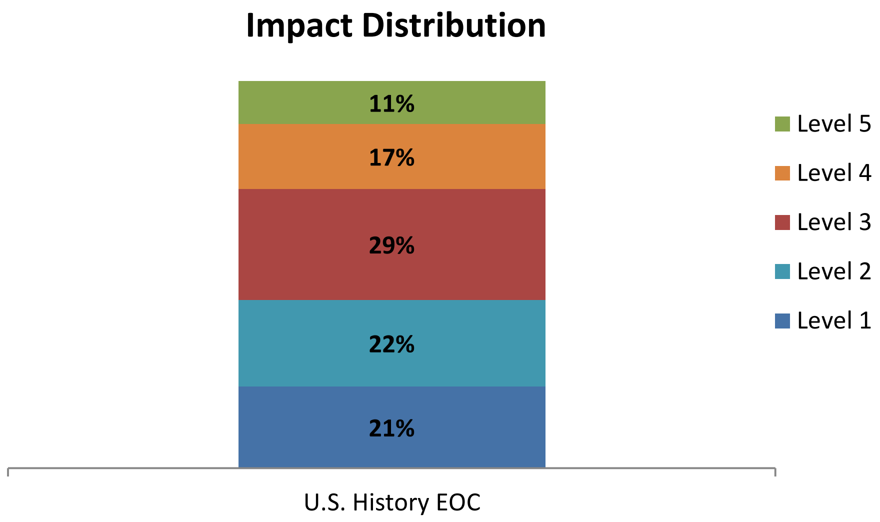

Figure 5-27. Impact Distribution Based on Final Approved Cut Scores

Chapter 6. Scaling

FCAT 2.0 and EOC assessments are standards-based assessments that have been constructed to align rigorously to the Next Generation Sunshine State Standards (NGSSS), as defined by FDOE and Florida educators. For each subject and grade level, the content standards specify the subject matter students should know and the skills they should be able to perform. In addition, performance standards specify how much of the content standards students need to master in order to achieve proficiency. Constructing tests to content standards enables the tests to assess the same constructs from one year to the next. However, it is inevitable that successive test forms will vary slightly in overall difficulty or in other psychometric properties, even though they all measure the same content standards. To ensure scale comparability, a set of derived, or "scale," scores for each test form is calculated. The process for accomplishing this multistep task is described in detail in this chapter and in "Chapter 7. Equating."

Rationale

Scaling is the process whereby we associate student performance with some ordered value, typically a number. The most common and straightforward way to score a test is to simply use the sum of points a student earned on the test, namely, raw score. Although the raw score is conceptually simple, it can be interpreted only in terms of a particular set of items. When new test forms are administered in subsequent administrations, other types of derived scores must be used to compensate for any differences in the difficulty of the items and to allow direct comparisons of student performance between administrations. Typically, a scaled metric is used, on which test forms from different years are equated.

Measurement Models

Item response theory (IRT) is used to derive the scale scores for all of the Florida statewide assessments except FCAT Writing, which reports on the raw score metric only. IRT is a general theoretical framework that models test responses resulting from an interaction between students and test items. The advantage of using IRT models in scaling is that all of the items measuring performance in a particular content area can be placed on the same scale of difficulty. Placing items on the same scale across years facilitates the creation of equivalent forms each year.

IRT encompasses a number of related measurement models, such as the Rasch partial credit (RPC; Masters, 1982), the two-parameter logistic model (2PL; Lord & Novick, 1968), the three-parameter logistic model (3PL; Lord & Novick, 1968), the generalized partial credit model (GPC; Muraki, 1992), as well as many others. A good reference text that describes commonly used IRT models is van der Linden and Hambleton (1997). These models differ in the types of items they can describe. Models designed for use with test items scored as right/wrong, for example, are called dichotomous models. Dichotomous models are used with multiple-choice (MC), gridded-response (GR), and fill-in response (FR) items. For the Florida statewide assessments, the 3PL model is used for MC items, and the 2PL model is used for GR and FR items. Both 3PL and 2PL models are described in the following section.

3PL/2PL Models

This section discusses two IRT measurement models: the 3PL model and the 2PL model. Both models are used with dichotomous items, such as MC, GR, and FR items. The 3PL model is mathematically defined as the probability of person *i* correctly answering item *j*:

$$P_{ij} = c_j + \frac{1 - c_j}{1 + \exp\left[-1.7a_j(\theta_i - b_j)\right]}, \qquad (6.1)$$

where a_j, b_j, and c_j are the item's slope (discrimination), location (difficulty), and lower asymptote/pseudo guessing parameters, respectively, and θ_i is the ability parameter for the person (Lord, 1980). The 2PL model can be defined by setting the *c* parameter to zero:

$$P_{ij} = \frac{1}{1 + \exp\left[-1.7a_j(\theta_i - b_j)\right]}. \qquad (6.2)$$

In Equations 6.1 and 6.2, student ability is represented by the variable θ (theta) and item difficulty by the model parameter *b*. Both θ and *b* are expressed on the same metric, ranging over the real number line, with greater values representing either greater ability or greater item difficulty. This metric is called the θ metric or θ scale.

The 3PL and 2PL models also permit variation in the ability of items to distinguish low-performing and high-performing students. This capability is quantified through a model parameter, usually referred to as the *a* parameter. Traditionally, a measure of an item's ability to separate high-performing students from low-performing students has been labeled the "discrimination index" of the item, so the *a* parameter in IRT models is often called the discrimination parameter.

In addition, the 3PL model also includes a lower asymptote (*c* parameter) for each item. The lower asymptote represents the minimum expected probability an examinee has of correctly answering a multiple-choice item. With multiple-choice items it is assumed that, because of guessing, examinees with minimal proficiency have a probability greater than zero of responding correctly to an item. For items scored right/wrong that are not multiple-choice, such as gridded-response items, the 2PL model is appropriate. As stated before, the 2PL model is equivalent to fixing the lower asymptote of the 3PL model to zero.

Examples of 3PL model item-response functions are presented in Figure 6-1. The *x*-axis is the θ scale, and the *y*-axis is the probability of a correct answer for the item. The solid curve represents an item with a *b*-value of -0.5, and the dotted curve represents an item with a *b*-value of 0.3. A distinguishing characteristic of the 3PL model whose discrimination parameters allow the slopes of the curves to vary is that the item-response functions of two items may cross. Figure 6-1 shows the effect of crossing curves. For students in the central portion of the θ distribution, sample item 2 is expected to be more difficult than sample item 1. However, students with $\theta > 1.0$ or $\theta < -3.0$ have a higher expected probability of getting item 2 correct.

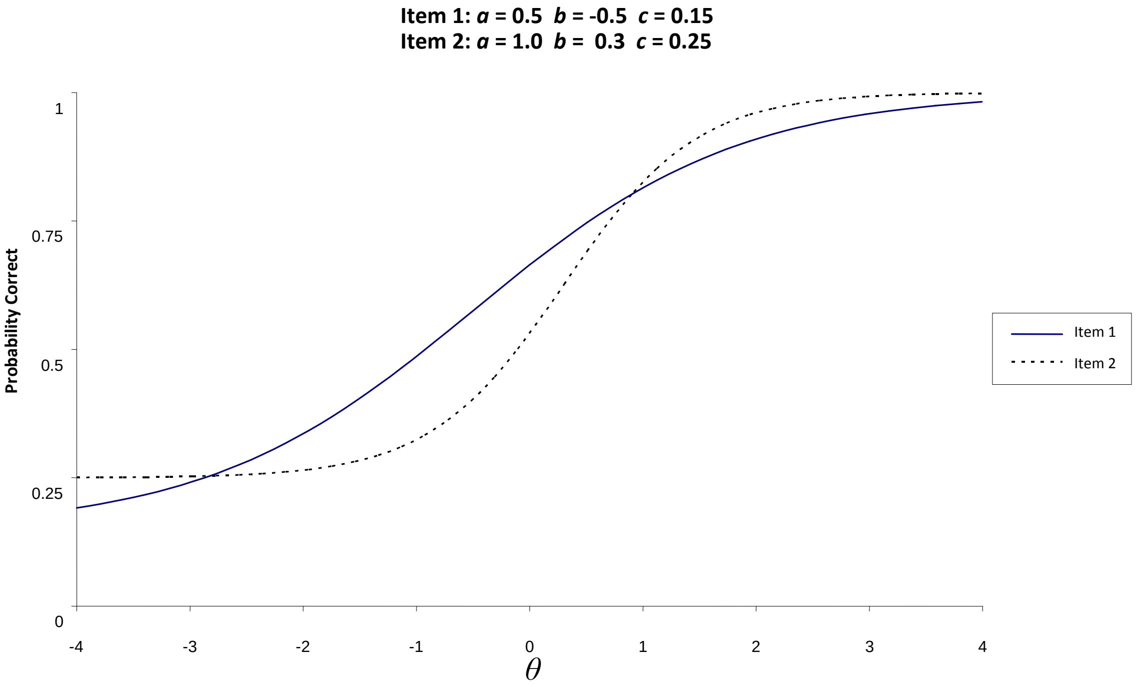

Figure 6-1. Item Response Functions for Two Sample Dichotomous Items

The figure also shows that item 2 clearly has a nonzero asymptote (*c* = .25). Item 1 also has a nonzero asymptote (*c* = .15). However, because of the relatively mild slope of the curve, the asymptote is reached only for extreme negative θ values that are outside the graphed range. In both 3PL and 2PL models, the *b* parameter specifies the inflection point of the curve and is a good overall indicator of item difficulty.

Calibration of items for the 3PL/2PL models is achieved using the computer program MULTILOG 7 (Thissen, 2003), which estimates parameters simultaneously for dichotomous and polytomous items via a statistical procedure known as marginal maximum likelihood. Simultaneous calibration of these items automatically puts their parameter estimates on the same scale. That scale is created on the assumption that test takers have a mean theta of approximately zero and a standard deviation of approximately one. Summaries of the item parameters can be found in the summary statistics reports of the yearbook.

Model Fit

IRT models provide a basis for the Florida reporting and test construction activities, e.g., the selection of test items, the equating procedures, and the scaling procedures. The utility of an IRT model depends on the extent to which it effectively reflects the data, so it is necessary that model data fit be evaluated before an IRT model is applied.

IRT scaling algorithms attempt to find item parameters (numerical characteristics) that create a match between observed patterns of item responses and theoretical response patterns defined by the selected IRT models. The Q_1 statistic (Yen, 1981) is used as an index for how well theoretical item curves match observed item responses. Q_1 is computed by first conducting an IRT item parameter estimation, then estimating students' achievement using the estimated item parameters, and, finally, using students' achievement scores in combination with estimated item parameters to compute expected performance on each item. Differences between expected item performance and observed item performance are then compared at 10 selected equal intervals across the range of student achievement. Q_1 is computed as a ratio involving expected and observed item performance. Q_1 is interpretable as a chi-square (χ^2) statistic, which is a statistical test that determines whether the data (observed item performance) fit the hypothesis (the expected item performance). Q_1 for each item type has varying degrees of freedom because the different item types have different numbers of IRT parameters. Therefore, Q_1 is not directly comparable across item types. An adjustment or linear transformation (translation to a *Z*-score, Z_{Q_1}) is made for different numbers of item parameters and sample size to create a more comparable statistic.

Q_1 can be expressed in the following terms:

$$Q_{1_j} = \sum_{i=1}^{I} \frac{N_{ji}\left(O_{ji} - E_{ji}\right)^2}{E_{ji}\left(1 - E_{ji}\right)} \qquad (6.3)$$

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

$$E_{ji} = \frac{1}{N_{ji}} \sum_{a e i}^{N_{ji}} P_j\left(\hat{\theta}_a\right) \qquad (6.4)$$

The Q_1 values are transformed into the statistic Z_{Q_1} :

$$Z_{Q_1} = \frac{Q_1 - df}{\sqrt{2df}} \qquad (6.5)$$

where *df* is the degree of freedom for the statistic (*df* = 10−the number of parameters estimated. That is, *df* = 7 for MC items and *df* = 8 for GR/FR items). FDOE has set criteria for a minimum Z_{Q_1} value standard for an item to have acceptable fit (FDOE, 1998).[3] Poor fit is indicated where Z_{Q_1} is greater than the critical value.

In 2014, no core items were flagged for poor fit for the FCAT 2.0 Mathematics and Science, which suggests that the 3PL and 2PL models applied to those Florida assessments fit the

[3] If Z_{Q_1} > (sample size × 4)/1500, then item fit is rated as "poor."

response data well. Grades 9 Reading had one item with poor fit. For EOC assessments, misfit was present in Algebra 1 EOC (one item out of 132), Biology 1 EOC (one out of 146), Geometry EOC (two out of 138). U.S. History and Civics EOC did not have any core items flagged for poor fit. A summary of the distributions of Z_{Q_1} and the number of poorly fitting items by item type can be found in the dimensionality reports of the yearbook.

Achievement Scale Unidimensionality

By fitting all items simultaneously to the same achievement scale, IRT is operating under the assumption that there is a strong, single construct that underlies the performance of all items. Under this assumption, item performance should be related to achievement (as depicted by Figure 6-1), and additionally, any relationship of performance between pairs of items should be explained, or accounted for, by variance in students' levels of achievement. This is the "local item independence" assumption of unidimensional IRT and suggests a relatively straightforward test for unidimensionality, called the Q_3 statistic (Yen, 1984).

Computation of the Q_3 statistic starts with expected student performance on each item, which is calculated using item parameters and estimated achievement scores. Then, for each student and each item, the difference between expected and observed item performance is calculated. The difference can be thought of as what is left in performance after accounting for underlying achievement. If performance on an item is driven by a single achievement construct, then not only will the residual be small (as tested by the Q_1 statistic), but the correlation between residuals of the pair of items also will be small. These correlations are analogous to partial correlations, which can be interpreted as the relationship between two variables (items) after the effects of a third variable (underlying achievement) is held constant or "accounted for." The correlation among IRT residuals is the Q_3 statistic.

When calculating the level of local item dependence for two items (*i* and *j*), the Q_3 statistic is

$$Q_3 = r_{d_i d_j}. \qquad (6.6)$$

A correlation between d_i and d_j values is a correlation of the residuals—that is, the difference between expected and observed scores for each item. For test taker *k*,

$$d_{ik} = u_{ik} - P_i(\theta_k), \qquad (6.7)$$

where u_{ik} is the score of the *k*th test taker on item *i* and $\mathrm{P}_i(\theta_k)$ represents the probability of test taker *k* responding correctly to item *i*.

With *n* items, there are $n(n-1)/2$ Q_3 statistics. For example, reading has 45 items and 990 Q_3 values. The Q_3 values should all be small. Summaries of the distributions of Q_3 are provided in the dimensionality report of the yearbook. Specifically, Q_3 data are summarized by minimum, 5th percentile, median, 95th percentile, and maximum values for each FCAT 2.0 grade and subject combination as well as EOC assessments. To add perspective to the meaning of Q_3

distributions, the average zero-order correlation (simple intercorrelation) among item responses is also shown. If the achievement construct is "accounting for" the relationships among the items, Q_3 values should be much smaller than the zero-order correlations. The Q_3 summary tables in the dimensionality reports of the yearbook indicate that for all grades and subjects, at least 90% (between the 5th and 95th percentiles) of the items are expectedly small. These data, coupled with the Q_1 data above, indicate that the unidimensional IRT model provides a very reasonable solution for capturing the essence of student achievement defined by the carefully selected set of items for each grade and subject.

Scale Scores

Basing scores on raw scores is easy to understand and to explain. However, test forms will undoubtedly vary slightly in difficulty across years; thus, a statistical equating process is typically used to ensure the forms yield scores that are comparable. The purpose of the scale score system is to convey accurate information about student performance from year to year.

The spring 2011 administration is the baseline year for FCAT 2.0 Reading grades 3–10 and Mathematics grades 3–8 as well as for the Algebra 1 EOC Assessment. In administrations after 2011, the scale score metric is equated to the 2011 administration for these grades and subjects. For FCAT 2.0 Science grades 5 and 8 and the Biology 1 and Geometry EOC Assessments, the spring 2012 administration is the baseline year. In administrations after 2012, the scale score metric is equated to the 2012 administration for these grades and subjects. The baseline year for the U.S. History EOC Assessment is the spring 2013 administration, and the scale score metric after 2013 will be equated to the 2013 scale. The baseline year for the Civics EOC Assessment is the spring 2014 administration.

Latent-Trait Estimation

After the item parameters are calibrated, equated, verified, and approved by the FDOE, examinee scale scores are produced. FCAT 2.0 and EOC scale scores are derived using an algorithm commonly referred to as pattern scoring, which is used in many educational testing programs across the country. This algorithm uses numerical methods to find the maximum likelihood score estimate of individual examinees given their pattern of responding.

In pattern scoring, the entire pattern of correct and incorrect student responses is taken into account. Unlike in number-correct scoring, where students who get the same number of dichotomously scored questions correct receive the same score, in pattern scoring, students rarely receive the same score, as even students getting the same number-correct score typically differ in the particular items they get correct or incorrect. Because pattern scoring utilizes information from the entire student response pattern, this type of scoring produces more reliable scores than does number-correct scoring. Generally, examinees who correctly answer harder questions receive higher scores than examinees who correctly answer only easier questions.

A software program developed by Pearson, IRT Score Estimation (ISE; Chien, Hsu, & Shin, 2006), was used to conduct pattern scoring. The program has been extensively tested and compared to commercially available software programs (e.g., MULTILOG, PARSCALE; Tong, Um, Turhan, Parker, Shin, Chien, & Hsu, 2007). The report indicates that with normal cases the ISE program is able to replicate MULTILOG and PARSCALE theta estimates.

However, "in problem cases, such as monotonically decreasing likelihood functions, in which MULTILOG and PARSCALE both produced theta estimates, ISE was able to produce the estimates that yielded the largest likelihood function, in alignment with the definition of the maximum likelihood algorithm" (p. 9). In addition, "with problem cases in which MULTILOG and PARSCALE failed to produce theta estimates, ISE was able to produce an estimate that yielded the largest likelihood from the likelihood function of a given response pattern" (p. 9). With regard to the CSEM, ISE produced similar results to MULTILOG. Details on the ISE program can be found in the user manual (Chien, Hsu, & Shin, 2006).

The theta scale is not often used for reporting because of interpretation issues arising from a scale with values ranging from -3.0 to +3.0, for FCAT 2.0 and EOC, for example. Therefore, following the calibration and equating phases, the resulting theta values are transformed to a reporting scale that can be more meaningfully interpreted by students, teachers, and other stakeholders. The scaling constants used for the transformation onto the FCAT 2.0 and EOC operational scales are described below.

Scaling Formulas

As discussed previously, the student theta estimates obtained through pattern scoring need to be transformed to scale scores for ease of interpretation. For FCAT 2.0 Mathematics and Reading, the transformation involves applying scaling constants obtained from the vertical scaling studies conducted for 2011. A vertical or growth scale links tests in the same subject area across grade levels. With a vertical scale, the gain in knowledge from one year to the next can be measured for each student. An accurate measure of student growth is valuable information for users of test scores.

Vertical scale scores (a.k.a. developmental scale scores) are reported for FCAT 2.0 Reading grades 3–10 and Mathematics grades 3–8 on individual student reports. The vertical scale is formed by linking across grades using common items in adjacent grades. The vertical scale allows a student's scores across time to be compared on the same scale and thus allows student performance on the FCAT 2.0 to be tracked as the student progresses from grade to grade. The actual development process used to form the IRT vertical scale is described in the "Vertical Scale" section. The following describes how the IRT vertical scale score is obtained.

Each student's vertical IRT scale score is computed using the student's theta estimate $\hat{\theta}_{EQ}$ (baseline for the first year of operation and post-equated starting the second year), the scaling constants given in the tables below, and the following formula:

$$\hat{\theta}_{VS} = A \cdot \hat{\theta}_{EQ} + B, \qquad (6.8)$$

where *A* represents the slope and *B* represents the intercept. The constants *A* and *B* for each grade of Reading and Mathematics are presented in Table 6-1 and Table 6-2. The minimum and maximum values from these tables represent the lower and upper limits for the vertical IRT scale score. For non-integer values, the vertical score value is rounded to the nearest integer value.

Table 6-1. Reading Vertical Scaling Constants and Scale Range

Grade	*A*	*B*	Minimum Vertical Scale Score	Maximum Vertical Scale Score[4]
3	20.000000	200.000000	140	260
4	19.187200	211.328800	154	269
5	19.371973	219.066806	161	277
6	19.414785	224.988624	167	283
7	19.640579	230.322448	171	289
8	20.247669	235.515613	175	296
9	20.668618	239.626498	178	302
10	18.822290	244.870126	188	302

Table 6-2. Mathematics Vertical Scaling Constants and Scale Range

Grade	*A*	*B*	Minimum Vertical Scale Score	Maximum Vertical Scale Score
3	20.000000	200.000000	140	260
4	19.310400	212.979200	155	271
5	19.341297	221.196548	163	279
6	19.156007	227.015183	170	284
7	18.797598	235.707221	179	292
8	18.565448	242.724553	187	298

There is no vertical scale for the EOC and FCAT 2.0 Science assessments. For the transformation from theta scores to scale scores, the same scaling formula is used as illustrated in Equation 6.8. The slope constant *A* and intercept constant *B* for EOC and Science can be found below in Table 6-3.

Table 6-3. FCAT 2.0 Science and EOC Scaling Constants

Subject	*A*	*B*	Minimum	Maximum
EOC	25	400	325	475
Science	20	200	140	260

Vertical Scale

The first version of the Florida Comprehensive Assessment Test (FCAT) program for Reading and Mathematics had a developmental scale (i.e., vertical scale) that allowed educators to

[4] A correction was applied to grade 10 Reading scores to eliminate a reversal between grades 9 and 10 on the Reading vertical scale.

compare the performance of students at adjacent grade levels with their performance from the previous instructional years. The implementation of NGSSS created a need for new test blueprints to reflect the changes, and as a result, all elements of the assessment tools and products had to be updated. For that reason, a new vertical scale was developed in 2011 for the new assessment program, FCAT 2.0, in Reading (grades 3–10) and Mathematics (grades 3–8).

Data from the spring 2011 administration were used to express all test scores across grades on the same reporting scale and to facilitate comparisons between grades in any given year. This requires the creation of a single score scale for the state assessment spanning the entire range of student performance for grades 3 through 10 for Reading and grades 3 through 8 for Mathematics. A scale of this kind is called a developmental scale or vertical scale.

Vertical scaling refers to the process of placing test scores that measure similar content domains at different educational levels onto a common scale. In K–12 assessments, a commonly used method of collecting data for vertical scaling is a common-item design (Kolen & Brennan, 2004). Under this design, tests at adjacent grade levels share common items, also known as anchor items, which do not count toward student total scores. Nevertheless, the performance of students on these common items does allow a link to be established between the scales for adjacent grade levels. Through this linking chain, test scores across grade levels can be placed onto a common scale.

Vertical scaling is both similar to and materially different from the calibration procedures used to place scores at a specific grade level on a common metric from year to year. The similarities may be characterized as procedural and the principal differences may be characterized as substantive. That is, many, but not all, of the statistical procedures that have been proposed for use in vertical scaling are similar to the more established procedures applied to horizontal scale calibration. Horizontal calibration puts scores from test forms designed to have the same level of difficulty on a common scale, and the procedures for horizontal calibration come from the closely related field of test equating.

One common objective of both horizontal and vertical processes is the development of a mathematical equation for transforming test scores from one scale to another; such transformation equations are often linear, and the procedures for deriving them are similar. Two notable, substantive differences include differences in the difficulty of test forms and the general ability levels of students at different grades. In the horizontal calibration process, different test forms are similar in content and difficulty level, and form-to-form differences in the overall ability levels of students in a given grade are small. In contrast, vertical scaling deals with student populations that differ substantially in ability from grade to grade, and with tests that differ accordingly in difficulty. Another important substantive difference is that the tests designed for the assessment of students at different grades necessarily differ in content, reflecting grade-to-grade differences in curriculum. Curriculum and content differences between nonadjacent grades may be profound; this poses a problem for the enterprise of vertical scaling that generally does not occur in within-grade scale calibration.

Data Collection Design

The data collection design for the FCAT 2.0 vertical scaling study utilized the embedded field-test positions on the assessments during the operational 2011 FCAT 2.0 administration. This design for the vertical scaling required the use of 4 forms (for the lowest and highest grades) or 8 forms (for the intermediate grades). Each form contains up to 8 vertical scaling items. Using the embedded field-test positions was desirable because students would have no knowledge of whether an item is an operational or a vertical scaling anchor item. The motivation factor will be the same for all the items. Therefore, a vertical scale developed using embedded items was deemed to be superior to one developed from a special study that requires a separate administration of vertical linking items.

The operationalized data collection design is based on a common-item non-equivalent groups design with dual-grade common items for both FCAT 2.0 Mathematics and Reading tests. The actual models for FCAT 2.0 Mathematics and Reading are depicted in Figure 6-2 and Figure 6-3, respectively. The data collection design required that the vertical linking items be administered at the same or similar positions in both on-grade-level and off-grade-level tests. As indicated in a recent study (Turhan, Courville, & Keng, 2009), controlling item position effect during vertical scale development results in a defensible vertical scale, with relatively larger mean differences between adjacent grade levels. For FCAT 2.0 vertical scaling, all of the vertical linking items were administered at comparable positions in both adjacent grade levels.

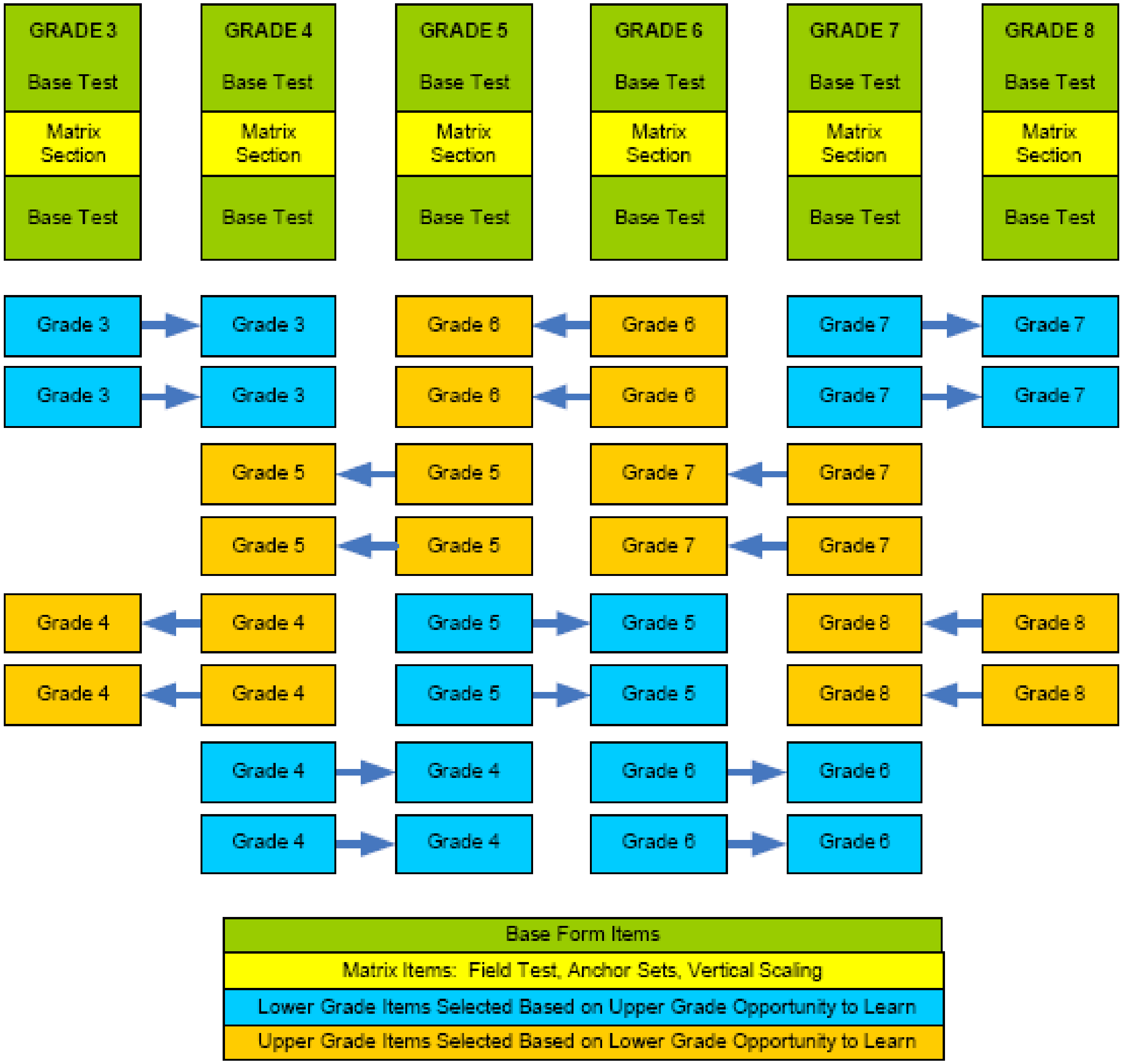

Figure 6-2. FCAT 2.0 Mathematics Vertical Scale Data Collection Design

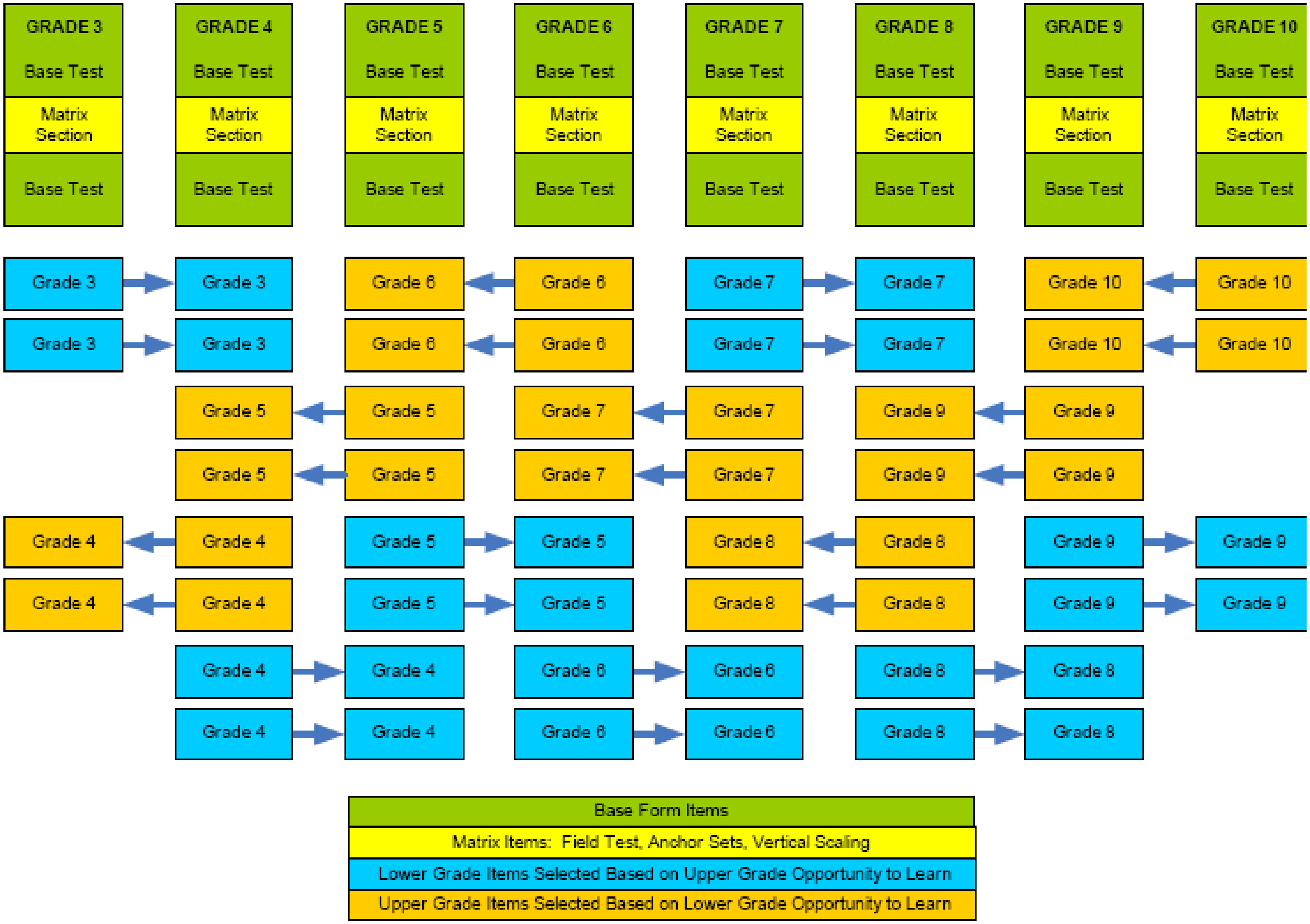

Figure 6-3. FCAT 2.0 Reading Vertical Scale Data Collection Design

Development of the Vertical Scale

The development of the FCAT 2.0 vertical scale involves multiple steps—including item calibration, diagnostic analysis of anchor items, and computation of scaling constants.

Calibration of vertical linking items

The item parameters for vertical scaling were estimated using the 3PL IRT model for multiple-choice (MC) items and the 2PL IRT model for gridded-response (GR) items. The maximum likelihood estimation procedure implemented within MULTILOG 7 (Thissen, 2003) was used. The following standard steps were performed to estimate the item parameters at each grade level:

1. Conduct a separate calibration of the operational items administered to each grade. This method was intended to create the reference scale for a given grade level with operational/core items only so that the reference scale would not be contaminated with the off-grade-level items.
2. Perform a second calibration including the operational and vertical linking items in the unscored item positions by form.
3. Link all of the items with the Stocking and Lord procedure (Stocking & Lord, 1983) through operational items only to put vertical scale items on the scale of a given grade level.
4. Apply linking constants to the vertical scale items to put them on the scale of the current grade level.

Diagnostic analysis of vertical scaling anchor items

Item fit was evaluated using Yen's (1981) Q_1 statistic for both on-grade-level and off-grade-level items. Fit statistics were used to evaluate the goodness-of-fit of IRT item parameter estimates to the actual performance of students. Details about the calculation of Q_1 and identification of poorly fitted items are discussed in the model fit section previously in this chapter. Items showing poor model fit (either on-grade level or off-grade level) were excluded from the computation of vertical linking constants.

In addition, vertical linking item statistics, such as item *p*-value, item discrimination, and IRT *b* parameter estimate from on-grade and off-grade levels, were plotted and any outliers were evaluated. A few items showed negative growth; that is, the *p*-value at the lower adjacent grade level is higher than the *p*-value at the higher adjacent grade level. These items displaying negative growth were also removed from the computation of vertical linking constants.

Defining the base grade level

The grade 3 level was used as the base grade in the development of the FCAT 2.0 vertical scale. A review of vertical scale slopes indicates no significant scale shrinkage and expansion across grades after the removal of items showing poor fit and negative growth.

Finding vertical linking constants for adjacent grade levels

Vertical scaling was conducted using the Stocking and Lord procedure. The Stocking and Lord procedure finds the slope (*A*) and intercept (*B*) equating constants by minimizing the difference between the test characteristic curves (TCCs) of the two forms. The relationships between estimated achievement across two adjacent grade levels was defined as

$$\theta_{Y_i} = A\theta_{X_i} + B, \tag{6.9}$$

where *A* represents the slope, *B* represents the intercept, and θ_{Y_i} and θ_{X_i} are achievement estimates for person *i* from the vertically linked scale *Y* and from the original scale *X,* respectively. The relationships between item parameters of two adjacent grade levels of the same test are represented by

$$a_{Y_j} = \frac{a_{X_j}}{A},$$

$$b_{Y_j} = Ab_{X_j} + B, \tag{6.10}$$

and

$$c_{Y_j} = c_{X_j},$$

where a_{Y_j}, b_{Y_j}, and c_{Y_j} are the item discrimination, item difficulty, and lower asymptote parameters for item *j* on form *Y,* and a_{X_j}, b_{X_j}, and c_{X_j} are the item parameters for item *j* on the reference scale *X* (Kolen & Brennan, 2004). The vertical scaling coefficients were estimated using the Stocking and Lord procedure (1983) with STUIRT (Kim & Kolen, 2004).

After adjacent grade level linking constants were computed, the relationship between the base grade and other grade levels was defined as provided in Table 6-1 and Table 6-2. These formulations are demonstrated as follows:

For example, linking equations between grades 3 and 4 and grades 4 and 5 are

$$Y_3 = A_{34}X_4 + B_{34} \tag{6.11}$$

and

$$Y_4 = A_{45}X_5 + B_{45} \tag{6.12}$$

If we replace X_4 with Y_4, after some simplifications, we can show that the relationship between grades 3 and 5 is

$$Y_3 = A_{34}A_{45}X_5 + (A_{34}B_{45} + B_{34}). \tag{6.13}$$

The product of two slopes (A_{34} and A_{45}) is the slope that relates the grade 5 achievement scale to the grade 3 achievement scale, and the term computed in parentheses is the intercept for this function. By repeatedly applying this algebraic manipulation, functions were created that placed each grade onto the grade 3 achievement scale. More details of the development of the FCAT 2.0 vertical scale can be found in *2011 FCAT 2.0 Vertical Scaling Report* (FDOE, 2011).

Vertical scaling constants were computed after items showing poor fit and negative growth were removed from the anchor set. This solution creates a smooth transition from one grade level to the next, with the intercept consistently increasing with grade level.

Properties of the Vertical Scale

The FCAT 2.0 vertical scale has a mean of 200 and a standard deviation of 20 at grade 3, the base grade. The mathematics scale ranges from 140 to 298 across grades 3–8. The reading scale ranges from 140 to 302 across grades 3–10. The scaling constants can be found in Table 6-1 and Table 6-2, along with the minimum and maximum scale score for each grade and subject.

Scale Drift

Scale drift refers to “a change in the interpretation that can be validly attached to scores on the score scale” (Haberman & Dorans, 2009). Although equating is used to maintain the score scale over time, the cumulative effects of changes might make the scores from one administration yield different meaning from scores on earlier administrations. The sources of changes may include population shifts, inconsistent or poorly defined test-construction practices, estimation error associated with small samples of examinees, and inadequate anchor tests. A series of sound equatings can also produce non-random drift (Haberman & Dorans, 2009). Scale drift studies will be conducted for FL EOC assessments in 2015.

Chapter 7. Equating

Equating is a procedure that allows tests to be compared across years. The procedures are generally thought of as statistical processes applied to the results of a test. Yet, successful equating requires attention to comparability throughout the test construction process. This chapter provides some insight into these procedures as they are applied to the Florida statewide assessments.

Rationale

In order to maintain the same performance standards across different administrations of a particular test, it is necessary for every administration of the test to be of comparable difficulty. Comparable difficulty should be maintained from administration to administration at the total test level and, as much as possible, at the subscore level. Maintaining test form difficulty across administrations is achieved through a statistical procedure called equating. Equating is used to transform the scores of one administration of a test to the same scale as the scores of a second administration of the test. Although equating is often thought of as a purely statistical process, a prerequisite for successful equating of test forms is that the forms are built to the same content and psychometric specifications. Without strict adherence to test specifications, the constructs measured by different forms of a test may not be the same, thus compromising comparisons of scores across test administrations.

For the Florida statewide assessments, a two-stage statistical process with pre- and post-equating is used to maintain comparable difficulty across administrations. This equating design is commonly used in state testing. In the pre-equating stage, item parameter estimates from prior administrations (either field-test or operational) are used to construct a form having difficulty similar to previous administrations. This stage is possible because of the embedded field-test design that allows for the linking of the field-test items to the operational form.

In the post-equating stage, all items are recalibrated, and the test is equated to prior forms through embedded linking items. Linking items are items that have previously been operational or field-test items, and whose parameters have been equated to the base year operational test metric. The performance of the linking items is examined for inconsistency with their previous results. If some linking items are found to behave differently, they are deleted from the linking set used for equating, but still count towards student total score if they are part of the operational test forms.

FDOE strives to use the pre- and post-equating design for all applicable testing programs so that the established level for any performance standard on the original test is maintained on all subsequent test forms. The pre- and post-equating design is fully described in the sections that follow.

In some cases, it may be desirable to compare the scores of tests that have been built to different specifications. A transformation can be made to place two different forms or tests on the same scale, but when the forms or tests in question are built to different specifications, the process is called linking. The term linking is used in place of equating to emphasize the more tenuous relationship created between scores on different tests. Although equating and linking create a relationship between different forms or tests, the strength or quality of the relationship depends on the degree to which the forms or tests measure the same constructs. Discussions on linking are given in Mislevy (1992), Linn (1993), and Kolen and Brennan (2004).

Pre-Equating

Test construction involves selecting items from the item pool meeting the content specifications of the subject tested and targeted psychometric properties. The construction process is an iterative one involving both content and psychometric staff from Pearson and FDOE.

The intent of pre-equating during test construction is to produce a test that is psychometrically equivalent to those used in prior years. Some of the psychometric properties targeted for the tests to be built include test difficulty, precision, and reliability. Since the item response theory (IRT) item parameters for each item in the item bank are on the same scale as the base scale test forms, direct comparisons of test characteristic functions, test information functions, and standard error of measurement can be made to ascertain whether the test has similar psychometric properties as the target form. Existing item parameters in the item bank are used to produce these curves. If the curves are not aligned with the targets, the newly constructed form will be modified until the alignment is satisfactory. Details about test construction of FCAT 2.0 and EOC Assessments can be found in “Chapter 2. Development.”

Post-Equating

Post-equating activities for 2014 FCAT 2.0 Reading, Mathematics, and Science, as well as the Algebra 1, Geometry, Biology 1, and U.S. History EOC Assessments, include calibration, post-equating, and scaling, which are described in “Chapter 6. Scaling.” Meanwhile, the Civics EOC Assessment had calibration and scaling (T-scale) only. Equating was not necessary because 2014 is the year in which new scale was established for the U.S. History EOC Assessment. Post-equating after 2014 for the Civics EOC will involve equating to the 2014 base scale.

Item Sampling for Equating

In order to achieve a successful equating of forms or tests, a solid statistical link must exist between the forms or tests. Typically, this means two forms or tests being equated or linked must have a set of items in common. It is important that the set of linking items be representative of the construct being measured by the tests and have the same approximate difficulty and spread of information as the tests that are being linked.

FCAT 2.0 assessments included up to 32 anchor items for post-equating. These anchor items were administered by utilizing 3 or 4 forms. FCAT 2.0 Mathematics and Science assessments used up to 32 external anchor items for the post-equating purposes. In contrast, FCAT 2.0 Reading assessments utilized up to 16 external anchor items and up to 16 internal anchor items. For EOC assessments, 24 internal anchor items were used across the four core forms. As mentioned previously, no equating occurred in 2014 for the Civics EOC Assessment. The internal linking items for the Civics EOC Assessment were used to put the four forms of the assessment on the same scale.

Student Sampling for Equating

Prior to planning the activities for 2011, FDOE requested that a comparative review of the expected Florida representativeness of the sample of student records be conducted when 60%, 65%, 70%, and 75% had been scanned and scored during the annual processing of FCAT documents. The goal of this review was to identify a threshold for the percentage of statewide processing that would be needed to obtain a reasonable representation of state performance for the purpose of conducting psychometric work. The conclusion was that once 65% of the state documents had been processed that the results were highly representative, and highly predictive, of the final state outcomes. Using this information, FDOE made a policy change in the annual distribution and collection of documents for the purposes of conducting psychometric work.

Before 2011, the previous strategy identified dozens of Florida schools to participate in an early return sample of documents. These schools were required to test in time to return documents several days earlier than other schools. The number of schools selected was based on the requirement of obtaining a sample of at least 16,000 students for each grade and subject. Beginning with the 2011 administration, data for calibration is extracted when at least 65% of the student records have been scanned and scored. In addition, Miami Dade, as the largest school district, needs to have at least 90% of the student data scored before the calibration can start.

Some student data, however, are excluded from the post-equating calibration of items. If the student score is considered invalid and not reported, then data from that student are excluded from the calibration data set. In addition, the student responses from the Department of Juvenile Justice, McKay and Corporate Tax Scholarship, Ahfachkee School, and Home Education Program are excluded from the calibration data set.

Operational Equating Procedures

Once the state-wide data file has been edited for exclusions, a statistical review of all operational items is conducted before beginning item response theory (IRT) calibration. A multiple-choice (MC) item is flagged for further review if it has a very low or very high mean score, a low item-total correlation, an unusually attractive incorrect option or a mean score on any one form that differs substantially from all the other forms. Specifically, MC items with one of the following characteristics are flagged:

- P-value < 0.15

- P-value > 0.90
- Item-total score correlation < 0.20
- Incorrect option selected by 40% or more students
- P-value on any one form differs from the population p-value by |0.08| for operational items which appear on multiple forms

Any flagged items are reviewed to ensure the item was correctly printed. Also, flagged items have keys checked by Pearson and FDOE content staff to certify the key is the correct answer. Gridded-response and fill-in response items are also reviewed for any additional plausible answer other than those already recognized as correct answers.

Calibration of FCAT 2.0 and EOC items is achieved using the computer program MULTILOG 7 (Thissen, 2003), which estimates the item parameters via a statistical procedure known as marginal maximum likelihood. As explained in Chapter 6, the three-parameter logistic model (3PL; Lord & Novick, 1968) is used for multiple-choice items, and the two-parameter logistic model (2PL; Lord & Novick, 1968) is used for gridded-response and fill-in response items. Calibration of FCAT 2.0 operational items is conducted by subject and grade. For EOC assessments, items from all four operational forms, including 24 items in common, are calibrated concurrently, so that item parameters of all forms were placed onto the same scale.

All operational items and anchor items for a test are calibrated simultaneously. After obtaining the anchor item parameter estimates on the current administration's operational scale, another scaling is performed to place the current operational scale on the base year scale. Scaling constants used to transform the current year scale to the base year scale are obtained by using the Stocking-Lord procedure (Stocking & Lord, 1983).

Once the anchor items have been equated to the original scale, a comparison of the item response functions is made to determine whether the anchor items are functionally the same across the two administrations. Substantial deviations in the item response functions of an item indicate that students responded differently to the anchor item as it appears in the current form than did students who took the item in a previous administration. If the item response function is substantially different for the two administrations, a decision may be made to discard the item from the linking set. The scaling process is then continued with the reduced linking set.

Once a satisfactory anchor item set and transformation equation have been determined, the same constants used to transform the anchor items to the base scale are applied to all the operational items of the current administration. With the current administration equated, student scores can be placed on the reporting metric as described in "Chapter 6. Scaling."

Development Procedure for Future Forms

Placing Field-Test Items on Operational Scale

The next step in the equating process is to place the item parameter estimates for the field-test items onto the same scale as the equated operational test items. All items, operational and field-test, are calibrated simultaneously. The Stocking-Lord procedure is used to find the scaling constants to transform the operational item parameter estimates of the combined calibration to the equated item scale. These same constants are then applied to the field-test items.

Item Pool Maintenance

The next step is to update the item pool with the new statistical information. The new item parameter estimates for the operational test items are added to the item pool database, as are the item statistics and parameter estimates for the field-test items. In this way, the item pool contains the parameter values from the most recent administration in which the item appeared.

Sources of Error Variances to Score Instability

Score fluctuations have been observed in many annual state test scores. According to Gary Phillips (2011), such fluctuations may be caused by: (1) sampling error arising from the cluster sampling used in field testing and other psychometric operations, and (2) equating error. A panel of nationally recognized experts reviewed his study, concluding that Phillips has identified “sources of significant and often unrecognized error variance” (Council of Chief State School Officers [CCSSO], 2011, p.1). This panel echoed the concerns raised by Phillips, recommending additional specific changes of psychometric practices. In this section, the sources of error variances identified by Phillips and the additional recommendations by CCSSO are discussed in relation to psychometric practices currently implemented in Florida as well as those planned for the future of the Florida statewide assessments.

Sampling Error

Samples of students are typically used in large-scale assessments to field test items and to equate or link test forms. Phillips illustrated the impact of two aspects of sampling design on sampling error: sampling selection and spiraling of forms. In regard to sampling of students, Phillips stated that cluster sampling (e.g., at the school level) reduces the effectiveness of sample size due to the similar characteristics of students within the cluster. For that reason, spiraling forms at the cluster level instead of the student level also reduces the effective sample size. Phillips recommended using simple random sampling, or conducting power analysis to find an effective sample size for cluster sampling if the random sampling is not feasible. He also suggested spiraling forms at the student level.

For FCAT 2.0 and EOC assessments, all test forms, including anchor and field-test forms, are spiraled at the student level throughout the state. Furthermore, at least 65% of the student population (about 130,000 students) is used to calibrate the operational items.

Field-test items are embedded into the operational test forms and calibrated based on 5,000 students selected throughout the state by spiraling the forms at the student level. These practices implemented in Florida comply with the recommendations by Phillips with respect to minimizing error variances due to sampling design.

Equating Error

Gary Phillips considered equating error as a major source of score instability. The magnitude of equating error depends on measurement errors in the old and new forms, equating sample size, number and quality of linking items, as well as error variances associated with examinees and items. Phillips recommended minimizing the equating error by adopting a statistically sound sampling method. He also suggested using the appropriate error variance calculation for all statistics.

Linking forms for FCAT 2.0 and EOC assessments are spiraled along with field-test forms at the student level throughout the state. This spiraling helps to minimize the sampling error and, therefore, the equating error. To quantify the equating error associated with the sampling of students, psychometricians from Pearson and FDOE employ a technique referred to as ”bootstrapping” (Bradley, 1979) for FCAT 2.0 and EOC assessments. Essentially, this technique is a computer-based method for estimating measures of accuracy of sample estimates, which are the transformation coefficients and their standard errors in the equating sense. The results of the bootstrapping procedures can be found in the yearbook.

To quantify the equating error associated with selecting anchor items, the “jackknife” method (Wu, 1986) is employed. This jackknife method systematically removes one anchor item at a time from an anchor set and recalculates the new transformation coefficients. The resulting transformation coefficient pairs are plotted on a graph to identify the outlier anchor items and monitor standard errors of equating coefficients due to the selection of anchor items.

Other Possible Sources of Error

In addition to the recommendations by Phillips, the expert panel who reviewed his paper has made additional recommendations as follows:

1. Ensure that the final set of anchor items is highly representative of the whole test.
2. Avoid stand-alone field testing and employ embedded field testing.
3. Verify assumptions of local item independence and unidimensionality.
4. Establish two links in common item design, linking either to two forms or two item banks.
5. Do not set standards with field-test results.

Florida currently practices most of the recommendations provided by the expert panel. For example, a set of anchor items for equating is pulled during the test construction such that the anchor form is a “miniature” version of the whole test in terms of both content and psychometric characteristics. Field-test forms are also embedded and spiraled at student level along with anchor forms. The item response theory assumptions of local item independence and unidimensionality are evaluated and reported in “Chapter 6. Scaling.” The standards for FCAT 2.0 and EOC assessments are set based on the student performance in the first operational year of assessments.

Chapter 8. Reliability

Reliability is the consistency of the results obtained from a measurement. When a score is reported for a student, there is an expectation that if the student had instead taken a different but equivalent version of the test, a similar score would have been achieved. A test that does not meet this expectation (i.e., a test that does not measure student ability and knowledge consistently) has little or no value. Furthermore, the ability to measure consistently is a prerequisite to making appropriate (valid) interpretations of scores on the measure. However, a reliable test score is not necessarily a valid one. And a valid and reliable test score for one purpose may not be valid for other purposes. A measure can be consistent and support certain score interpretations but still not support all the inferences a user of the test wishes to make. The concept of test validity is discussed in “Chapter 9. Validity.”

A Mathematical Definition of Reliability

Classical test theory provides a mathematical definition of reliability wherein all measures consist of an accurate or “true” part and some inaccurate or “error” component (Feldt & Brennan, 1989). This axiom is commonly written as:

$$\text{Observed Score} = \text{True Score} + \text{Error}. \qquad (8.1)$$

Errors occur as a natural part of the measurement process and can never be eliminated entirely. For example, uncontrollable factors such as differences in administration conditions and changes in examinee disposition may work to increase error and decrease reliability. In classical test theory, error is typically assumed to be the result of random, unsystematic influences. If there are systematic influences contributing to the error term, then derived reliability indices are likely to be inaccurate. For example, if a test is administered under conditions of very poor lighting, the results of the test are likely to be biased against the entire group of students taking the test under the adverse conditions. From Equation 8.1, it is apparent that scores from a reliable test generally have little error and vary primarily because of true score differences. One way to calculate reliability is to define reliability as the variance of the students’ true scores divided by the variance of their observed scores (see Equation 8.2).

$$\text{Reliability} = \frac{\sigma_T^2}{\sigma_O^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = 1 - \frac{\sigma_E^2}{\sigma_O^2}, \qquad (8.2)$$

where σ_T^2 is the true score variance, σ_O^2 is the variance of the observed score and σ_E^2 is the error variance. When there is no error, true score variance divided by true score variance is unity; in other words, the measure is perfectly reliable. However, as more error influences the measure, the error component in the denominator of the ratio increases and the reliability decreases.

To calculate reliability, as defined in Equation 8.2, two of the three variances must be known. On any assessment, the variance of observed scores can be calculated. However, true scores and error is unknown so it is not immediately obvious how to calculate these variances. Several reliability estimates are provided below. These estimates provide different ways of using two or more scores on classically parallel forms to estimate the error variance and use it, along with the observed score variance, to estimate reliability.

Estimating Reliability

There are a number of different approaches taken to estimate reliability of test scores. Discussed below are test-retest, alternate forms, and internal consistency methods.

Test-Retest Reliability Estimation

Reliability can be estimated by calculating the correlation coefficient between scores from a test given on one occasion with scores from the same test given on another occasion to the same students. Essentially, the test is acting as its own parallel form. Using the test-retest reliability method has potential pitfalls. A long interval between testing sessions may result in student growth in knowledge of the subject matter or forgetting of the subject matter, while a short interval increases the chance students will remember and repeat answers from the first session. In addition, the test-retest approach requires the same students to take a test twice. For these reasons, test-retest reliability estimation is not used on Florida statewide assessments.

Alternate Forms Reliability Estimation

Alternate forms reliability is similar to test-retest, except that instead of repeating the identical test, two equivalent forms of the test are administered to the same students. The accuracy of the alternate forms coefficient greatly depends on the degree to which the two forms are equivalent (classically parallel). For Florida statewide assessments, alternate forms reliability estimation is not possible because no student takes more than one form of the test during any test administration.

Internal Consistency Reliability Estimation

Internal consistency methods use a single administration to estimate test score reliability. For state assessments where student testing time is at a premium, internal consistency procedures have a practical advantage over reliability estimation procedures requiring multiple tests. Probably the most frequently used internal consistency reliability estimate is coefficient alpha (Cronbach, 1951). Coefficient alpha is based on the assumption that inter-item covariances constitute true-score variance and the fact that the average true score variance of items is greater than or equal to the average inter-item covariance. The formula for coefficient alpha is

$$\alpha = \left(\frac{N}{N-1}\right)\left(1 - \frac{\sum_{i=1}^{N} s_{Y_i}^2}{s_X^2}\right), \qquad (8.3)$$

where N is the number of items on the test, $s_{Y_i}^2$ is the sample variance of the ith item (or component) and s_X^2 is the observed score variance for the test. Coefficient alpha is appropriate for use when the items on the test are reasonably homogenous. Evidence for the homogeneity of Florida tests is obtained through a dimensionality analysis. Dimensionality analyses results are discussed in "Chapter 9. Validity."

The reliability and classification accuracy reports in the yearbook provide up to four estimates of internal consistency reliability, depending on grades and subjects, coefficient alpha, Feldt-Raju (Feldt & Brennan, 1989), stratified α (Qualls, 1995), and IRT model-based or "marginal reliability" (Thissen, 2003) for the total tests. Marginal reliability is described as "an average reliability over levels of θ or theta" (Thissen, 1990). Marginal reliability may be reproduced by squaring and subtracting from 1 each of the 31 "posterior standard deviations" (SEMs) in the MULTILOG output file. Since the variance of the population is 1, each of these values represents the reliability at each of the 31 θs. Marginal reliability is the average of these computations weighted by the normal probabilities for each of the 31 quadrature intervals. The formula for marginal reliability is

$$\bar{\rho} = \frac{s_\theta^2 - E(SEM_\theta^2)}{s_\theta^2}, \qquad (8.4)$$

where s_θ^2 is the variance of a given θ (is 1 for standardized θ) and $E(SEM_\theta^2)$ is the average error variance (a.k.a. the mean of the squared posterior standard deviations by weighting population density). Marginal reliability can be interpreted in the same way as traditional internal consistency reliability estimates (such as coefficient alpha).

Feldt-Raju and stratified α reliability coefficients are appropriate for estimating the reliability of FCAT 2.0 Mathematics and EOC assessments because the assessments are composed of multiple item formats. Different item formats are unlikely to have equivalent distributions of true/observed scores and error variance. Thus, accurate estimation of the reliability of a test composed of multiple item formats requires consideration of the fact that the effective weights of the items on the total test variance differ (Qualls, 1995). Feldt-Raju and stratified α coefficients require dividing the total test into "sub-tests" based on item format. The Feldt-Raju coefficient assumes that the sub-tests measure the same trait and have measurement errors proportional to the number of items. The formula for Feldt-Raju is

$$_{\textit{Feldt-Raju}}\ \rho_{XX'} = \frac{s_X^2 - \sum s_{Y_J}^2}{(1-\sum \hat{\lambda}_j^2)s_X^2}, \tag{8.5}$$

where s_X^2 is the observed score sample variance for the total test, $s_{Y_J}^2$ is the observed sub-test variance, and $\hat{\lambda}_j^2$ is the estimated squared functional weight for sub-test *j*.

Stratified *α* is appropriate when sub-tests measure different traits, but requires that each sub-test be composed of at least two items. The formula for stratified *α* is

$$_{\textit{strat }\alpha}\rho_{XX'} = 1 - \frac{\sum s_{X_i}^2 (1 - {}_{\alpha}\rho_{X_i X_i'})}{s_X^2}, \tag{8.6}$$

where $s_{X_i}^2$ is the observed sub-test variance, s_X^2 is the observed score variance for the total test, and ${}_{\alpha}\rho_{X_i X_i'}$ is the estimated reliability of the sub-test.

Additional reliabilities were calculated on various demographic subgroups[5] using the entire population of students (see reliability and classification accuracy reports in the yearbook). Within each table, coefficient alpha estimates are provided for the entire test, as well as each major subscale. Included with coefficient alpha in the tables is the number of students responding to the test, the mean score obtained by this group of students, and the standard deviation of the scores obtained for this group.

Subscore reliability will generally be lower than total score reliability because reliability is influenced by the number of items (as well as their covariation). In some cases, the number of items associated with a subscore is small (ten or fewer). Results involving subscores must be interpreted carefully, as in some cases these measures have low reliability due to the limited number of items attached to the score.

Standard Error of Measurement

A reliability coefficient expresses test score consistency in terms of variance ratios. In contrast, the standard error of measurement (SEM) expresses score inconsistency (unreliability) in terms of the reported score metric. The SEM is an estimate of how much error there is likely to be in an individual's observed score, or alternately, how much score variation would be expected if the individual were tested multiple times with equivalent forms of the test. The standard error of measurement is calculated using the following formula:

[5] The subgroups are male/female, white/African American/Hispanic/Asian/American Indian/multiracial, economically disadvantaged, English language learners, students with disabilities, and migrants.

$$SEM = s_x\sqrt{1-\rho_{XX'}}\ , \qquad (8.7)$$

where s_x is the standard deviation of the total test (standard deviation of the raw scores) and $\rho_{xx'}$ is the reliability estimate for the total test.

Use of the Standard Error of Measurement

The SEM is used to quantify the precision of a test in the metric in which scores will be reported. The SEM can be helpful for quantifying the variability in student scores due to random factors that are unrelated to student knowledge and skills. A standard error of measurement band placed around the student’s true score would result in a range of values most likely to contain the student’s observed score. The observed score may be expected to fall within one SEM of the true score 68 percent of the time, assuming that measurement errors are normally distributed.

For example, if a student has a true score of 40 on a test with reliability of 0.93 and a standard deviation of 9.45, the SEM would be

$$SEM = 9.45\sqrt{1-0.93} = 2.50. \qquad (8.8)$$

Placing a ±1 SEM band around this student’s true score would result in a score range of 37.50 to 42.50 (that is, 40 ± 2.50). Furthermore, if it is assumed the errors are normally distributed and if this procedure were replicated across repeated test administrations, this student’s observed score would be expected to fall within the ±1 SEM band 68 percent of the time (assuming no learning or memory effects). Thus, the chances are better than 2 out of 3 that a student with a true score of 40 would have an observed score within the interval 37.50–42.50. This interval is called a confidence interval or confidence band. By increasing the range of the confidence interval, one improves the likelihood the confidence interval includes the observed score; an interval of ± 1.96 SEMs around the true score covers the observed score with 95 percent probability and is referred to as a 95 percent confidence interval.

Conversely, it is *not* the case that a ±1 SEM band around the *observed score* will include the true score 68% of the time (Dudek, 1979). Whereas true and error scores are uncorrelated, observed and error scores *are* correlated, as error is a component of observed score. Thus, observed score is a biased estimator of true score, and the correct approach to constructing a confidence band for true score requires centering the confidence band on the observed score adjusted for unreliability. Still, it is common practice to use a confidence band around the observed score as a rough approximation to the true score range.

The SEM is reported by subject and grade for Florida assessments in the test summary reports of the yearbook.

Conditional Standard Error of Measurement

Although the overall SEM is a useful summary indicator of a test's precision, the measurement error on most assessments varies across the score range. This means the measurement accuracy of a test is likely to differ for students depending on their score. To formalize this notion, classical test theory postulates that every student has a true score. This is the score the student would receive on the test if no error were present. The standard error of measurement for a particular true score is defined as the standard deviation of the observed scores of students with that true score. This standard deviation is called the conditional standard error of measurement (CSEM). The reasoning behind the CSEM is as follows: if a group of students all have the same true score, then a measure without error would assign these students the same score (the true score). Any differences in the scores of these students must be due to measurement error. The conditional standard deviation defines the amount of error.

True scores are not observable. Therefore, the CSEM cannot be calculated simply by grouping students by their true score and computing the conditional standard deviation. However, item response theory (IRT) allows for the CSEM to be estimated for any test where the IRT model holds. To compute CSEM from IRT models, it is necessary to compute the test information function since CSEM is the inverse of the test information function. The standard error of measurement for a given θ can be estimated by using the following formula (Hambleton, Swaminathan & Rogers, 1991):

$$CSEM(\theta) = \frac{1}{\sqrt{I(\theta)}}. \tag{8.9}$$

The test information function is the sum of the item information functions (Thissen and Orlando, 2001). The test information is calculated using the following formula:

$$I(\theta) = \sum_{i=1}^{n} \frac{P_i^{'}(\theta)^2}{P_i(\theta)Q_i(\theta)}, \tag{8.10}$$

where $P_i(\theta)$ is the probability of an examinee responding correctly to item *i* given an ability of θ, $Q_i(\theta)$=1-$P_i(\theta)$, and $P_i^{'}(\theta)$ is the first derivative of $P_i(\theta)$. CSEM would be different for each scale score since every item's contribution to the test information depends on the item's parameters.

The conditional standard errors of the scale scores are provided in the achievement level reports of the yearbook. The conditional standard error values can be used in the same way to form confidence bands as described for the traditional test-level SEM values.

Measurement Error for Groups of Students

As is the case with individual student scores, district, school and classroom averages of scores are also influenced by measurement error. Averages, however, tend to be less affected by error than individual scores. Much of the error due to systematic factors (that is, bias) can be avoided with a well-designed assessment instrument that is administered under appropriate and standardized conditions. The remaining random error present in any assessment cannot be fully eliminated, but for groups of students much of the random error is expected to cancel out (that is, average to zero). Some students score a little higher than their true score, while others score a little lower. The larger the number in the group, the more the cancelling of errors tends to occur. The degree of confidence in the average score of a group is almost always greater than for an individual score.

Standard Error of the Mean

Confidence bands can be created for group averages in much the same manner as for individual scores, but in this case, the width of the confidence band varies due to the amount of *sampling error*. Sampling error results from using a sample to infer characteristics of a population, such as the mean. Sampling error will be greater to the degree the sample does not accurately represent the population as a whole. When samples are taken from the population at random, the mean of a larger sample will generally have less sampling error than the mean of a smaller sample.

A confidence band for group averages is formed using the standard error of the mean. This statistic, s_e, is defined as

$$s_e = \frac{s_x}{\sqrt{N}}, \qquad (8.11)$$

where s_x is the standard deviation of the group's observed scores and N is the number of students in the group.

As an example of how the standard error of the mean might be used, suppose that a particular class of 20 students had an average scale score of 215 with a standard deviation equal to 10. The standard error would equal

$$s_e = \frac{10}{\sqrt{20}} = 2.2. \qquad (8.12)$$

A confidence band around the class average would indicate that one could be 68 percent confident that the true class average on the test was in the interval 215 ± 2.2 (212.8 to 217.2).

Scoring Reliability for Written Compositions

Reader Agreement

Human raters score the Florida's writing assessments that involve writing a composition in response to a prompt. In order to monitor the reliability of the written composition scores, 100 percent of the compositions are scored independently by two raters. The reliability of these scores is estimated using between-rater agreement. Rater agreement data show the percentage of perfect agreement of each rater against all other raters.

However, reader agreement data do not provide a mechanism for monitoring drift from established criteria by all raters at a particular grade level. Thus, an additional set of data, resulting from a procedure known as validity scoring, is collected daily to check for reader drift and rater consistency in applying scoring criteria to student responses.

When Pearson's scoring supervisors identify ideal student responses (i.e., ones that appear to be exemplars of a particular score value), they route these to the scoring directors for review. Scoring directors examine the responses and choose appropriate papers for validity scoring. Validity responses are usually solid score point responses. The scoring directors confirm the scores assigned to the validity responses, and, upon approval by TDC staff, those responses are entered into the validity scoring pool. Raters are assigned validity responses through the image-based scoring system in the same way they receive other student responses. They do not know when they are scoring a validity response. Results of validity scoring are analyzed regularly by scoring directors, and appropriate actions are initiated as needed, including the retraining or termination of scorers.

The rater agreement reports in the yearbook give the score frequency distribution for each prompt for grades 4, 8, and 10. Also presented is the percent agreement among raters. As mentioned above, checking the consistency of scores that raters provide for the same composition is one form of inter-rater reliability. Rater agreement is categorized as perfect agreement (no difference between raters), adjacent agreement (one score point difference), non-adjacent agreement (two score point difference), or non-agreement (more than two point score difference). Another index of inter-rater reliability reported in the tables is the correlation of ratings from the first and second rater.

Score Appeals

A district may appeal the score assigned to any student's composition about which a question has been raised. In these instances, an individual analysis of the composition in question is provided.

Student Classification Accuracy and Consistency

Students are classified into one of five performance levels based on their FCAT 2.0 scale scores. It is important to know the reliability of student scores in any examination, but assessing the reliability of the classification decisions based on these scores is of even greater importance. Evaluation of the reliability of classification decisions is performed through estimation of the probabilities of correct and consistent classification of students. Procedures were used from Livingston and Lewis (1995) and Lee, Hanson, and Brennan (2000) to derive measures of the accuracy and consistency of the classifications. A brief description of the procedures used and the results derived from them are presented in this section.

Accuracy of Classification

According to Livingston and Lewis (1995, p.180), the accuracy of a classification is "the extent to which the actual classifications of the test takers . . . agree with those that would be made on the basis of their true scores, if their true scores could somehow be known." Accuracy estimates are calculated from cross-tabulations between "classifications based on an observable variable (scores on . . . a test) and classifications based on an unobservable variable (the test takers' true scores)." True score is also referred to as a hypothetical mean of scores from all possible forms of the test if they could be somehow obtained (Young & Yoon, 1998). Since these true scores are not available, Livingston and Lewis provide a method to estimate the true score distribution of a test and create the cross-tabulation of the true score and observed score classifications. An example of the 5 × 5 cross-tabulation of the true score versus observed score classifications for FCAT 2.0 grade 3 Reading is given in Table 8-1. It shows the proportions of students who were classified into each performance category by the actual observed scores and by estimated true scores. The results in the accuracy table below are similar to the ones calculated in 2014. The current accuracy tables for FCAT 2.0 and for the EOC assessments are provided in the yearbook. The example discussed in the following tables—pulled from one subject and grade of the 2014 administration cycle—is provided in the technical report only to explain in a detailed way the reasoning behind the tables produced in the yearbook.

Table 8-1. 2014 FCAT 2.0 Grade 3 Reading True Scores vs. Observed Scores Cross-Tabulation (Accuracy Table)

True Score	Observed Score					
	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5	Total
LEVEL 1	0.150	0.031	0.001	0.000	0.000	0.180
LEVEL 2	0.036	0.160	0.054	0.003	0.000	0.255
LEVEL 3	0.001	0.050	0.128	0.049	0.000	0.226
LEVEL 4	0.000	0.003	0.050	0.161	0.030	0.236
LEVEL 5	0.000	0.000	0.000	0.021	0.072	0.103
Total	0.187	0.244	0.232	0.235	0.102	1.000

Note: Columns and row totals are computed from non-rounded values. Shaded cells are used for computing overall accuracy index.

Consistency of Classification

Consistency is “the agreement between classifications based on two non-overlapping, equally difficult forms of the test” (Livingston & Lewis, 1995, p. 180). Consistency is estimated using actual response data from a test and the test’s reliability in order to statistically model two parallel forms of the test and compare the classifications on those alternate forms. The example of 5 × 5 cross-tabulation between two forms of FCAT 2.0 grade 3 Reading is given in Table 8-1 The table shows the proportions of students who were classified into each performance category by the actual test and by expected scores. The accuracy table compares classifications based on two different types of scores. Also, note that agreement rates are lower in the consistency table because both classifications contain measurement error; whereas, in the accuracy table, true score classification is assumed to be errorless. The results in the accuracy table below are similar to the ones calculated in 2011, 2012, and 2013. The current consistency tables for FCAT 2.0 and for the EOC assessments are provided in the yearbook.

Table 8-2. 2014 FCAT 2.0 Grade 3 Reading True Scores vs. Observed Scores Cross-Tabulation (Consistency Table)

Expected Scores	Observed Scores					
	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5	Total
LEVEL 1	0.143	0.049	0.005	0.000	0.000	0.198
LEVEL 2	0.040	0.127	0.064	0.012	0.000	0.243
LEVEL 3	0.004	0.056	0.096	0.054	0.002	0.212
LEVEL 4	0.000	0.012	0.063	0.130	0.032	0.237
LEVEL 5	0.000	0.000	0.003	0.039	0.068	0.110
Total	0.187	0.244	0.232	0.235	0.102	1.000

Note: Columns and row totals are computed from non-rounded values.

Accuracy and Consistency Indices

There are three types of accuracy and consistency indices that can be generated from these tables: *overall, conditional-on-level,* and *cut point*. In order to facilitate their interpretations, a brief outline of computational procedures used to derive accuracy indices will be presented using the example of the FCAT 2.0 grade 3 Reading test. The *overall accuracy* of performance level classifications is computed as a sum of the proportions on the diagonal of the joint distribution of true score and observed score levels, as indicated by the shaded area in Table 8-1 **Error! Reference source not found.**. Actually, it is a proportion (or percentage) of correct classifications across all the levels. In this particular example, the overall accuracy index for FCAT 2.0 grade 3 Reading test equals 0.67 (67 percent). It means that 67% of students are classified in the same performance categories based on their observed scores as they would have been classified based on their true scores if they could be known.

The overall consistency index is analogously computed as a sum of the diagonal cells in the consistency table. Using the data from Table 8-2, it can be determined that the overall consistency index for the FCAT grade 3 Reading test equals 0.56 (56 percent). In other words, 56% of grade 3 students would be classified in the same performance levels based on the alternate form if they had taken it. Another way to express overall consistency is to use Cohen's Kappa (κ) coefficient (Cohen, 1960). The overall coefficient Kappa when applying all cutoff scores together is

$$\kappa = \frac{P - P_c}{1 - P_c}, \qquad (8.13)$$

where P is the probability of consistent classification, and P_c is the probability of consistent classification by chance (Lee, 2000). The P is the sum of the diagonal elements and P_c is the sum of the squared row totals.

Kappa is a measure of "how much agreement exists beyond chance alone" (Fleiss, 1973), which means that it provides the proportion of consistent classifications between two forms after removing the proportion of consistent classifications expected by chance alone.

Using the data from Table 8-2, it was computed that Cohen's κ for FCAT 2.0 grade 3 Reading equals 0.446, which is a moderate agreement (Viera & Garrett, 2005). Compared to the previously described overall consistency estimate, Cohen's κ has lower value because it is corrected for chance.

Consistency conditional-on-level is computed as the ratio between the proportion of correct classifications at the selected level (diagonal entry) and the proportion of all the students classified into that level (marginal entry). In Table 8-2, the row LEVEL 4 is outlined and corresponding cells are shaded. The ratio between 0.13 (proportion of correct classifications) and 0.237 (total proportion of students classified into LEVEL 4) yields 0.548, which represents the index of consistency of classification for FCAT grade 3 Reading that is conditional on LEVEL 4. It indicates that 54.8% of all the students whose performance is classified as LEVEL 4 would be classified in the same level based on the alternate form if an alternate form were taken.

Accuracy conditional-on-level is analogously computed. The only difference is that in the consistency table both row and column marginal sums are the same; whereas, in the accuracy table, the sum that is based on true status is used as a total for computing accuracy conditional on level. For example, in Table 8-1 the proportion of agreement between true score status and observed score status at LEVEL 1 is 0.150 whereas the total proportion of students with true score status at this level is 0.180 (the row total). The accuracy conditional on level is equal to the ratio between those two proportions, which yields 0.83. It indicates that 83% of the students estimated to have true score status on LEVEL 1 are correctly classified into that category by their observed scores in the FCAT 2.0 grade 3 Reading test.

Perhaps the most important indices for accountability systems are those for the accuracy and consistency of classification decisions made at specific cut points. To evaluate decisions at specific cut points, the joint distribution of all the performance levels is collapsed into a dichotomized distribution around that specific cut point. For example, the dichotomization at the cut point that separates LEVEL 1 through LEVEL 3 (combined) from LEVEL 4 and LEVEL 5 (combined) for FCAT 2.0 grade 3 Reading is depicted in Table 8-3. The proportion of correct classifications below that particular cut point is equal to the sum of the cells in the upper left shaded area (0.610), and the proportion of correct classifications above the particular cut point is equal to the sum of the cells in the lower right shaded area (0.284). The accuracy index at cut point is computed as the sum of the proportions of correct classifications around a selected cut point.

Table 8-3. Cut Point Accuracy for 2013 FCAT 2.0 Grade 3 Reading True Scores vs. Observed Scores Cross-Tabulation

True Score	Observed Score					
	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5	Total
LEVEL 1	0.150	0.031	0.001	0.000	0.000	0.180
LEVEL 2	0.036	0.160	0.054	0.003	0.000	0.255
LEVEL 3	0.001	0.050	0.128	0.049	0.000	0.226
LEVEL 4	0.000	0.003	0.050	0.161	0.030	0.236
LEVEL 5	0.000	0.000	0.000	0.021	0.072	0.103
Total	0.187	0.244	0.232	0.235	0.102	1.000

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

In our example from Table 8-3, the sum of both shaded areas (upper left shaded areas added to lower right shaded areas) equals 0.894, which means that 89.4% of students were correctly classified either above or below the particular cut point. The sum of the proportions in the upper right non-shaded area (0.053) indicates false positives (i.e., 5.3% of students are classified above the cut point by their observed score but fell below the cut point by their true score), and the sum of the lower left non-shaded area (0.053) is the proportion of false negatives (i.e., 5.3% of students are observed below the cut point level whose true level is above the cut point).

The consistency at cut point is obtained in an analogous way. For example, if the distribution at the cut point between LEVEL 1 and all other levels combined in Table 8-2 is dichotomized, it can be determined that the proportion of correct classifications around that cut point equals 0.901. This means that 90.1% of students would be classified by alternate form (if they had taken it) in the same two categories (LEVEL 1 or LEVEL 2 through LEVEL 5 combined) as they were classified by the actual form taken.

Accuracy and Consistency Results

Detailed tables with accuracy and consistency cross-tabulations, dichotomized cross-tabulations, overall indices, indices conditional-on-level, and indices by cut points are presented in the reliability and classification accuracy reports of the yearbook.

Chapter 9. Validity

Validation is the process of collecting evidence to support inferences from assessment results. A prime consideration in validating a test is determining if the test measures what it purports to measure. During the process of evaluating if the test measures the construct of interest, a number of threats to validity must be considered. For example, the test may be biased against a particular group, test scores may be unreliable, students may not be properly motivated to perform on the test, the test content may not span the entire range of the construct to be measured, etc. Any of these threats to validity could compromise the interpretation of test scores.

Beyond ensuring that the test is measuring what it is supposed to measure, it is equally important that the interpretations made by users of the test's results are limited to those that can be legitimately supported by the test. The topic of appropriate score use is discussed in "Chapter 4. Reports" (in the section "Cautions for Score Use") and "Chapter 6. Scaling."

Demonstrating that a test measures what it is intended to measure and that interpretations of the test's results are appropriate requires an accumulation of evidence from several sources. These sources generally include expert opinion, logical reasoning, and empirical justification. What constitutes a sufficient collection of evidence in the demonstration of test validity has been the subject of considerable research, thought, and debate in the measurement community over the years. Several different conceptions of validity and approaches to test validation have been proposed, and as a result the field has evolved.

This chapter begins with an overview of the major historical perspectives on validity in measurement. Included in this overview is a presentation of a modern perspective that takes an argument-based approach to validity. Following the overview is the presentation of validity evidence for the FCAT 2.0 and EOC assessments.

Perspectives on Test Validity

The following sections discuss some of the major conceptualizations of validity used in educational measurement.

Criterion Validity

The basis of criterion validity is demonstration of a relationship between the test and an external criterion. If the test is intended to measure mathematical ability, for example, then scores from the test should correlate substantially with other valid measures of mathematical ability. Criterion validity addresses how accurately criterion performance can be predicted from test scores. The key to criterion-related evidence is the degree of relationship between the assessment tasks and the outcome criterion (Cronbach, 1990). In order for the observed relationship between the assessment and the criterion to be a meaningful indicator of criterion validity, the criterion should be relevant to the assessment and be reliable. Criterion validity is typically expressed in terms of the product-moment correlation between the scores of the test and the criterion score.

There are two types of criterion-related evidence: concurrent and predictive. The difference between these types lies in the procedures used for collecting validity evidence. Concurrent evidence is collected from both the assessment and the criterion at the same time. An example might be found in relating the scores from a district-wide assessment to the ACT assessment (the criterion). In this example, if the results from the district-wide assessment and the ACT assessment were collected in the same semester of the school year, this would provide concurrent criterion-related evidence. On the other hand, predictive evidence is usually collected at different times; typically the criterion information is obtained subsequent to the administration of the measure. For example, if the ACT assessment results were used to predict success in the first year of college, the ACT results would be obtained in the junior or senior year of high school, whereas the criterion (e.g., college grade point average) would not be available until the following year.

In ideal situations, the criterion validity approach can provide convincing evidence of a test's validity. However, there are two important obstacles to implementing the approach. First, a suitable criterion must be found. Standards-based tests like the FCAT 2.0 and EOC assessments are designed to measure student achievement on Florida's NGSSS. Finding a criterion representing achievement on the standards may be difficult to do without creating yet another test. It is possible to correlate performance on the FCAT 2.0 and EOC assessments with other types of assessments, such as the ACT or school assessments. Strong correlations with a variety of other assessments would provide some evidence of validity for the FCAT 2.0 and EOC assessments, but the evidence would be less compelling if the criterion measures are only indirectly related to the standards.

A second obstacle to the demonstration of criterion validity is that the criterion may need to be validated as well. In some cases, it may be more difficult to demonstrate the validity of the criterion than to validate the test itself. Further, unreliability of the criterion can substantially attenuate the correlation observed between a valid measure and the criterion.

Criterion-related validity evidence on the FCAT 2.0 and EOC assessments will be collected and reported in an ongoing manner. These data are most likely to come from districts conducting program evaluation research, university researchers and special interest groups researching topics of local interest, as well as the data collection efforts of FDOE.

Content and Curricular Validity

Content validity is a type of test validity that addresses whether the test adequately samples the relevant domain of material it purports to cover (Cronbach, 1990). If a test is made up of a series of tasks that form a representative sample of a particular domain of tasks, then the test is said to have good content validity. For example, a content-valid test of mathematical ability should be composed of tasks allowing students to demonstrate their mathematical ability.

Evaluating content validity is a subjective process based on rational arguments. Even when conducted by content experts, the subjectivity of the method remains a weakness. Also, content validity only speaks to the validity of the test itself, not to decisions made based on the test scores. For example, a poor score on a content-valid mathematics test indicates that the student did not demonstrate mathematical ability. But from this alone, one cannot conclusively determine that the student has low mathematical ability. This conclusion could only be reached if it could be shown or argued that the student put forth his or her best effort, the student was not distracted during the test, and the test did not contain a bias preventing the student from scoring well.

Generally, achievement tests such as the FCAT 2.0 and EOC assessments are constructed so that they have strong content validity. As documented by this report, tremendous effort is expended by FDOE, the content vendor (Pearson), and the educator committees to ensure the FCAT 2.0 and EOC assessments are content-valid. Although content validity has limitations and cannot serve as the only evidence for validation, it is an important piece of evidence for the validation of the FCAT 2.0 and EOC assessments.

Construct Validity

The term construct validity refers to the degree to which the observed test score is a measure of the underlying characteristic (i.e., the latent construct) of interest. A construct is an individual characteristic assumed to exist in order to explain some aspect of behavior (Linn & Gronlund, 1995). When a particular individual characteristic is inferred from an assessment result, a generalization or interpretation in terms of a construct is being made. For example, problem solving is a construct. An inference that students who master the mathematical reasoning portion of an assessment are “good problem-solvers” implies an interpretation of the results of the assessment in terms of a construct. To make such an inference, it is important to demonstrate this is a reasonable and valid use of the results.

Construct-related validity evidence can come from many sources. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) provides the following list of possible sources:

- High inter-correlations among assessment items or tasks attest that the items are measuring the same trait, such as a content objective, sub-domain, or construct;
- Substantial relationships between the assessment results and other measures of the same defined construct;

- Little or no relationship between the assessment results and other measures that are clearly not of the defined construct;
- Substantial relationships between different methods of measurement regarding the same defined construct; and
- Relationships to non-assessment measures of the same defined construct.

Messick (1989) describes construct validity as a “unifying force” in that inferences based on criterion evidence or content evidence can also be framed by the theory of the underlying construct. From this point of view, validating a test is essentially the equivalent of validating a scientific theory. As Cronbach and Meehl (1955) first argued, conducting construct validation requires a theoretical network of relationships involving the test score. Validation not only requires evidence supporting the notion that the test measures the theoretical construct, but it further requires evidence be presented that discredits every plausible alternative hypothesis as well. Because theories can only be supported or falsified, but never proven, validating a test becomes a never-ending process.

Kane (2006) states that construct validity is now widely viewed as a general and all-encompassing approach to accessing test validity. However, in Kane’s view there are limitations of the construct validity approach, including the need for strong measurement theories and the general lack of guidance on how to conduct a validity assessment.

Validity Argument Evidence for the Florida Assessments

The following sections present a summary of the validity argument evidence for each of the four parts of the interpretive argument: scoring, generalization, extrapolation, and implication. Much of this evidence is presented in greater detail in other chapters in this report. In fact, the majority of this report can be considered validity evidence for the FCAT 2.0 and EOC assessments (e.g., item development, performance standards, scaling, equating, reliability, performance item scoring, quality control). Relevant chapters are cited as part of the validity evidence given below.

Scoring Validity Evidence

Scoring validity evidence can be divided into two sections. These sections are the evidence for the scoring of performance items and the evidence for the fit of items to the model.

Scoring of Performance Items

The scoring of constructed response items on Florida assessments is a complex process that requires its own chapter to describe fully. “Chapter 10. Constructed-Response Scoring” gives complete information on the careful attention paid to the scoring of performance items. The chapter’s documentation of the processes of range-finding, rubric review, recruiting and training of scorers, quality control, appeals, and security provides some of the evidence for the validity argument that the scoring rules are appropriate. Further evidence comes from the tables found in the yearbook that report inter-rater agreement and inter-rater reliabilities. The results in those tables show both of these measures are generally high for FCAT Writing.

Model Fit and Scaling
Item response theory (IRT) models provide a basis for the FCAT 2.0 and EOC assessments. IRT models are used for the selection of items to go on the test, the equating procedures, and the scaling procedures. A failure of model fit would undermine the validity of these procedures. Item fit is examined during test construction. Any item displaying misfit is carefully scrutinized before a decision is made to place the item on the test. However, the vast majority of items display good model fit.

Principal Components Analysis
Further evidence of the fit for the IRT model comes from dimensionality analyses. IRT models for the FCAT 2.0 and EOC assessments assume the domain being measured by the test is relatively unidimensional. To test this assumption, principal components analysis (Jolliffe, 2002) is performed. The scree plots for the principal component analyses for each subject and grade are provided in the dimensionality reports of the yearbook. The topography of the scree plot indicates the dimensionality of the items constituting the test. The point at which the scree plot becomes flat is indicative that adding further dimensions to the analysis would become irrelevant. The scree plots presented in the yearbook show the first dimension is dominant in terms of explaining item variance, with markedly less item variance explained by additional dimensions. This type of result in a scree plot is evidence that items appearing on the FCAT 2.0 and EOC assessments measure a single dimension.

Confirmatory Factor Analyses (CFA)
Second-order confirmatory factor analyses (Bollen, 1989) were conducted for the FCAT 2.0 and the EOC Assessments. The goal of the analyses was to investigate whether or not performance on the items in each assessment reflects a single underlying construct. The findings from the analyses also could be used to establish whether the unidimensional model-based IRT used to calibrate the FCAT 2.0 and EOC Assessments items was appropriate.

Mplus [version 6.11] (Muthén & Muthén, 2007) was used to calculate matrices of tetrachoric correlations[6] between the items included in each analysis. Mplus was also used to fit CFA models to the data using WLSMV (weighted least squares mean and variance adjusted) estimation. In the CFA model for each grade and subject, one second-order CFA model was fitted to the data: the factors at the first level are the factors defined by the items in each reporting category, and the second-order factor was defined by all of the first level factors (see Figure 9-1). For example, in FCAT 2.0 Reading, four first-order constructs measuring (1) Vocabulary, (2) Reading Application, (3) Literary Analysis (Fiction and Nonfiction), and (4) Informational Text and Research

[6] A correlation coefficient computed for two dichotomous variables that are assumed to have continuous and normal distributions.

Process were specified, and the second-order factor was specified as reading ability. Models with three first-order factors were specified for FCAT 2.0 Mathematics (except grade 7, which has 4 reporting categories) and the EOC assessments.

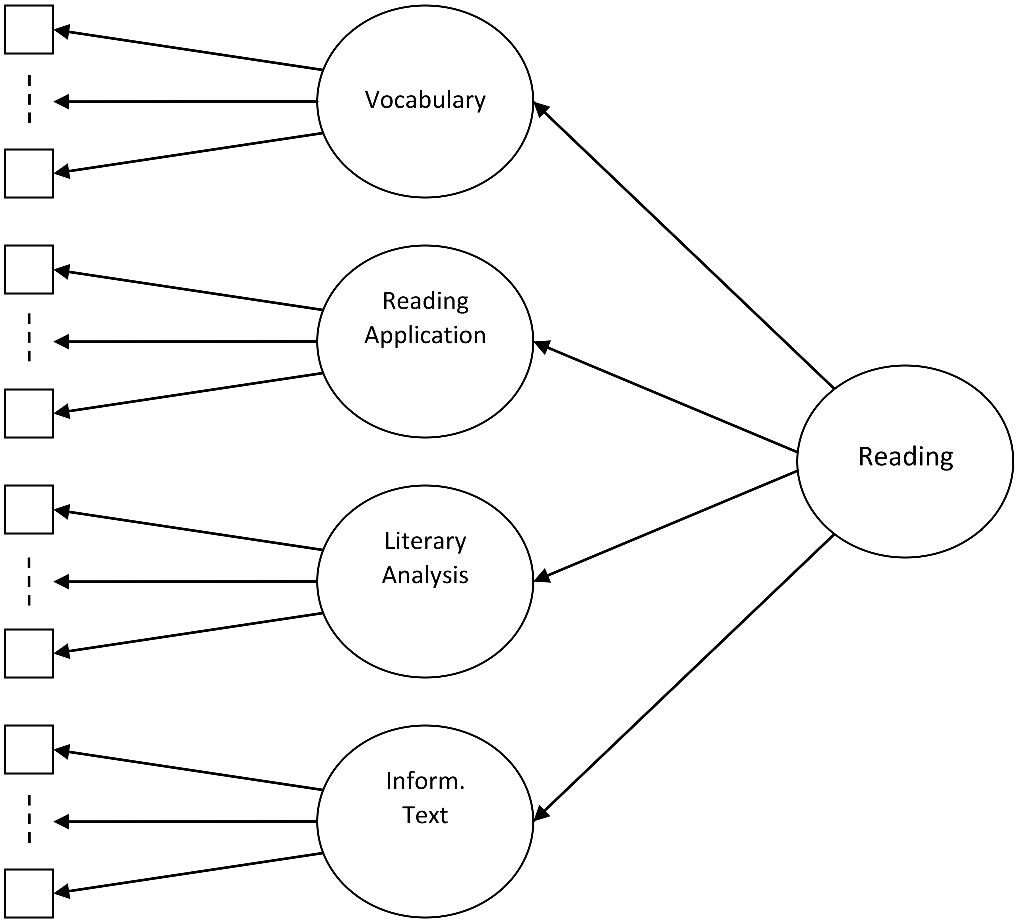

Figure 9-1. Simplified Representation of Fitted Reading Second Order CFA model.

Three fit statistics were used to interpret the model fit. Two relative fit indices and one absolute fit index were investigated to assess model fit. The Tucker-Lewis index (TLI) (Tucker & Lewis, 1973) and the comparative fit index (CFI) are both relative fit indices. Relative fit indices compare fit of the hypothesized model with that of a null or “independence” model, in which all covariances are equal to zero (i.e., the model with the worst possible fit). Relative fit indices can be roughly interpreted to represent the percentage of improvement in model fit when comparing the hypothesized model to the null model. Both CFI and TLI range between 0 and 1, with values greater than or equal to 0.95 indicating close fit (Hu & Bentler, 1999). The root mean square error of approximation (RMSEA) is an absolute fit index (i.e., an index that does not compare back to a null model) with a model parsimony correction factor included. RMSEA can be interpreted as representing the amount of misfit per degree of freedom in the model.

RMSEA values theoretically range between 0 and infinity, but observed RMSEA values greater than 1.00 are rare; values of 0 indicate perfect fit and values less than 0.05 indicate close fit (Brown, 2006).

Table 9-1 shows the results of the analyses. The three fit statistics indicated that the specified model (see Table 9-1) by first order and second order fit the data well in all content areas. These findings provide evidence that the tests for each content area measure a single dimension.

Table 9-1. CFA Model Fit Summary

Grade/Form	Subject	Convergence	RMSEA	CFI	TLI
3/Core	Reading	Yes	0.022	0.982	0.982
4/Core	Reading	Yes	0.019	0.981	0.980
5/Core	Reading	Yes	0.019	0.979	0.978
6/Core	Reading	Yes	0.016	0.987	0.986
7/Core	Reading	Yes	0.019	0.982	0.981
8/Core	Reading	Yes	0.018	0.981	0.980
9/Core	Reading	Yes	0.017	0.984	0.983
10/Core	Reading	Yes	0.015	0.986	0.985
3/Core	Mathematics	Yes	0.030	0.970	0.969
4/Core	Mathematics	Yes	0.029	0.972	0.970
5/Core	Mathematics	Yes	0.031	0.970	0.969
6/Core	Mathematics	Yes	0.032	0.965	0.963
7/Core	Mathematics	Yes	0.025	0.979	0.978
8/Core	Mathematics	Yes	0.024	0.978	0.976
EOC/Form 1	Algebra 1	Yes	0.025	0.962	0.960
EOC/Form 2	Algebra 1	Yes	0.026	0.962	0.960
EOC/Form 3	Algebra 1	Yes	0.028	0.955	0.953
5/Core	Science	Yes	0.014	0.987	0.986
8/Core	Science	Yes	0.015	0.986	0.986
EOC/Form 1	Geometry	Yes	0.027	0.972	0.971
EOC/Form 2	Geometry	Yes	0.033	0.964	0.963
EOC/Form 3	Geometry	Yes	0.035	0.952	0.950
EOC/Form 4	Geometry	Yes	0.033	0.959	0.958
EOC/Form 1	Biology 1	Yes	0.017	0.975	0.974
EOC/Form 2	Biology 1	Yes	0.015	0.983	0.982
EOC/Form 3	Biology 1	Yes	0.016	0.982	0.981
EOC/Form 4	Biology 1	Yes	0.016	0.981	0.980
EOC/Form 1	U.S. History	Yes	0.014	0.984	0.983
EOC/Form 2	U.S. History	Yes	0.013	0.985	0.984
EOC/Form 3	U.S. History	Yes	0.013	0.987	0.986
EOC/Form 4	U.S. History	Yes	0.014	0.985	0.985
EOC/Form 1	Civics	Yes	0.012	0.975	0.974
EOC/Form 2	Civics	Yes	0.011	0.978	0.977
EOC/Form 3	Civics	Yes	0.011	0.979	0.978
EOC/Form 4	Civics	Yes	0.014	0.967	0.965

Item-Level Analyses

Another check for unidimensionality can be made at the item level. The content measured by each item on the test should have a strong relationship with the content measured by the other items. An item-total correlation (also called a point-biserial correlation when items are dichotomously scored) is the correlation between an item and the total test score. Conceptually, if an item has a high item-total correlation (that is, 0.30 or above), it indicates that students who performed well on the test answered the item correctly and students who performed poorly on the test answered the item incorrectly; the item did a good job of discriminating between high-achieving and low-achieving students. Assuming the total test score represents the extent to which a student possesses the construct being measured by the test, high item-total correlations indicate the items on the test require this construct to be answered correctly. The summary statistics report of the yearbook present summaries of item-total correlations by subject and grade. For the FCAT 2.0 and EOC assessments, item-total correlations are generally high.

Justification for the scaling procedures used for the FCAT 2.0 and EOC assessments can be found in “Chapter 6. Scaling.”

Generalization Validity Evidence

There are two major requirements for validity that allow generalization from observed scale scores to universe scores[7]. First, the items administered on the test must be representative of the universe of possible items. Evidence regarding this requirement comes from content validity. Content validity is documented through evidence that the test measures the content standards and benchmarks. The second requirement for validity at the generalization stage is that random measurement error on the test is controlled. Evidence that measurement error is controlled comes largely from reliability and other psychometric measures. These sources of evidence are reported in the sections that follow.

Evidence of Content Validity

The FCAT 2.0 and EOC assessments are based on content standards and benchmarks along with extensive content limits that help define what is to be assessed. Committees of educators collaborate with item development experts, assessment experts, and FDOE staff annually to review new and field-test items so that each test adequately samples the relevant domain of material the test is intended to cover. These review committees participate in this process to verify the content validity of each test.

[7] Universe score is defined as the expected value of a person’s observed scores over all observations in the universe of generalization, which is analogous to a person’s “true score” in classical test theory (Shavelson & Webb, 2006).

The sequential committee review process is outlined in “Chapter 2. Development.” In addition to providing information on the difficulty, appropriateness, and fairness of items and performance tasks, committee members provide a check on the alignment between the items and the benchmarks measured. When items are judged to be relevant, that is, representative of the content defined by the standards, this provides evidence to support the validity of inferences made regarding knowledge of this content from the results. When items are judged to be inappropriate for any reason, the committee can either suggest revisions (e.g., rewording an item or reclassifying the item to a more appropriate benchmark) or elect to eliminate the item from the field-test item pool. Items approved are later embedded in live forms to allow for the collection of performance data. In essence, these committees review and verify the alignment of the test items with the content standards and measurement specifications so that the items measure the appropriate content. The nature and specificity of these review procedures provide strong evidence for the content validity of the test.

Skilled professionals are also involved in establishing evidence of content validity in other ways. Item writers must have at least three years teaching experience in the subject areas for which she/he will be creating items and tasks or two years of experience writing or reviewing items for the subject area. Each team is composed of qualified professionals who also have an understanding of psychometric considerations and sensitivity to racial/ethnic, gender, religious, and socioeconomic issues. Pearson is responsible for identifying a team of commissioned reading passage authors. A sufficient number of passage authors is included so that there are a variety of high-quality passages. The passage authors have been previously published in a critically-reviewed publication and must have their résumés approved by FDOE.

Using a varied source of item writers provides a system of checks and balances for item development and review, reducing single-source bias. Since many different people with different backgrounds write the items, it is less likely items will suffer from a bias that might occur if items were written by a single author. The input and review by these assessment professionals provide further support of the item being an accurate measure of the intended content domain.

The test summary reports of the yearbook contain tables showing the number of assessment components, tasks, or items matching each content standard. Test specifications are posted on the FDOE website and can be found at the following locations:

- FCAT 2.0: http://fcat.fldoe.org/fcat2/itemspecs.asp
- EOC: http://fcat.fldoe.org/eoc/itemspecs.asp

Evidence of Control of Measurement Error

Reliability and the standard error of measurement (*SEM*) are discussed in “Chapter 8. Reliability.” The yearbook has tables reporting the conditional *SEM* and coefficient alpha reliability, broken down by gender and ethnic groups. As discussed in Chapter 8, these measures show the FCAT 2.0 and EOC assessments to be reliable.

Further evidence is needed to show the IRT model fits well. Item-fit statistics and tests of unidimensionality apply here, as they did in the section describing evidence argument for scoring. As described above, these measures indicate good fit of the model.

Validity Evidence for Different Student Populations

It can be argued from a content perspective that the FCAT 2.0 and EOC assessments are not more or less valid for use with one subpopulation of students relative to another. The FCAT 2.0 and EOC assessments measure the NGSSS, which are required to be taught to all students. The tests have the same content validity for all students because what is measured on the tests is taught to all students, and all tests are given to all students under standardized conditions.

Great care has been taken so that the items constituting the FCAT 2.0 and EOC assessments are fair and representative of the content domain expressed in the content standards. Additionally, much scrutiny is applied to the items and their possible impact on demographic subgroups making up the population of the state of Florida. Every effort is made to eliminate items that may have ethnic or cultural biases. As described in “Chapter 2. Development,” item writers are trained on how to avoid economic, regional, cultural, and ethnic bias when writing items. After items are written and passage selections are made, committees of Florida educators are convened by FDOE to examine items for potential subgroup bias. As described in “Chapter 7. Equating,” items are further reviewed for potential bias by committees of educators and FDOE after field-test data are collected.

Extrapolation Validity Evidence

Validity for extrapolation requires evidence that the universe score is applicable to the larger domain of interest. Although it is usually impractical or impossible to design an assessment measuring every concept or skill in the domain, it is desirable for the test to be robust enough to allow some degree of extrapolation from the measured construct. The validity argument for extrapolation can use either analytical evidence or empirical evidence. These lines of evidence are detailed below.

Analytic Evidence

The NGSSS create a common foundation to be learned by all students and define the domain of interest. As documented in this report, the FCAT 2.0 and EOC assessments are designed to measure as much of the domain defined by the standards as possible.

The use of different item types also increases the validity of Florida assessments. The combination of multiple-choice, gridded-response, and fill-in response items results in assessments measuring the domain of interest more fully than if only one type of response format was used.

A threat to the validity of the test can arise when the assessment requires competence in a skill unrelated to the construct being measured. For example, students who are English Language Learners (ELL) may have difficulty fully demonstrating their mathematical knowledge if the mathematics assessment requires fluency in English. The use of accommodation avoids this threat to validity by allowing students who are ELL to demonstrate their mathematical ability on a test that limits the quantity and complexity of English language used in the items. The FCAT 2.0 and EOC assessments also allow accommodations for students with vision impairment or other special needs. The use of accommodated forms allows accurate measurement of students who would otherwise be unfairly disadvantaged by taking the standard form. Accommodations are discussed in "Chapter 3. Administration."

Empirical Evidence

Empirical evidence of extrapolation is generally provided by criterion validity when a suitable criterion exists. As discussed before, finding an adequate criterion for a standards-based achievement test can be difficult.

Studies investigating criterion validity have yet to be carried out for the FCAT 2.0 and EOC assessments. Because no other assessment is likely to be found to measure the standards as well as these assessments, the most promising empirical evidence would come from criterion validity studies with convergent evidence. Any test that measures constructs closely related to the standards could potentially serve as a criterion. Although these tests would not measure the content standards as well as the FCAT 2.0 and EOC assessments, they could serve as an external check. If a number of these external tests that are highly correlated with the FCAT 2.0 and EOC assessments could be found, the convergent evidence from them would provide justification for extrapolation.

Implication Validity Evidence

There are inferences made at different levels based on the FCAT 2.0 and EOC assessments. Individual student scores are reported, as well as aggregate scores for schools and districts. Inferences at some levels may be more valid than those at others. For example, the FCAT 2.0 assessments report individual student scores, but some students may feel that few ramifications of the test directly affect them; such students may fail to put forth their full effort. Although this report documents in detail evidence showing that the FCAT 2.0 assessment is a valid measure of student achievement on the standards, individual and school-level scores are not valid if students do not take the test seriously. The incorporation of graduation requirements associated with the grade 10 FCAT 2.0 Reading and EOC Assessments increases the consequences of the test for high school students; this may mitigate concerns about student motivation affecting test validity. Also, as students are made fully aware of the potential No Child Left Behind (NCLB) ramifications of the test results for their school, this threat to validity should diminish.

One index of student effort is the percentage of blank or "off-topic" responses to the writing prompts. Because writing prompts require more time and cognitive energy, low levels of non-response on these items is evidence of students giving their full effort. The rater agreement reports of the yearbook provide the percentage of unscorable responses for grades 4, 8, and 10, which are typically very low.

One of the most important inferences to be made concerns the student's achievement level, especially for accountability tests. Even if the total-correct score can be validated as an appropriate measure of the standards, it is still necessary that the scaling and achievement level designation procedures be validated. Because scaling and standard setting are both critical processes for the success of the FCAT 2.0 and EOC assessments, separate chapters are devoted to them in this report. Chapter 5 discusses the details concerning performance standards, and Chapter 6 discusses scaling. These chapters serve as documentation of the validity argument for these processes.

At the aggregate level (i.e., school, district, or statewide), the implication validity of school accountability assessments can be judged by the impact the testing program has on the overall proficiency of students. Validity evidence for this level of inference will result from examining changes over time in the percentage of students classified as proficient. As mentioned before, there exists a potential for negative impacts on schools as well, such as increased dropout rates and narrowing of the curriculum. Future validity studies need to investigate possible unintended negative effects as well.

Summary of Validity Evidence

Validity evidence is described in this chapter as well as other chapters of this report. In general, validity arguments based on rationale and logic are strongly supported for the FCAT 2.0 and EOC assessments. The empirical validity evidence for the scoring and the generalization validity arguments for these assessments are also quite strong. Reliability indices, model fit, and dimensionality studies provide consistent results, indicating the FCAT 2.0 and EOC assessments are properly scored, and scores can be generalized to the universe score.

Less strong is the empirical evidence for extrapolation and implication. This is due in part to the absence of criterion studies. Because an ideal criterion for the FCAT 2.0 or EOC assessments probably cannot be found, empirical evidence for the extrapolation argument may need to come from several studies showing convergent validity evidence. Further studies are also needed to verify some implication arguments. This is especially true for the inference that the state's accountability program is making a positive impact on student proficiency and school accountability without causing unintended negative consequences.

Chapter 10. Constructed-Response Scoring

During the 2013–2014 academic year, only the FCAT Writing test required scoring of constructed-response items. In this case, student responses to a direct writing prompt were scored through a process that was thoroughly overseen by the FDOE staff.

Scoring Process

The scoring process for FCAT Writing consists of four distinct processes: rangefinding, recruiting of scorers, training, and operational scoring. Each process contains several steps. The processes used by Florida meet or exceed industry best practices (Cohen & Wollack, 2006). The scoring rubrics for FCAT Writing are posted on the FDOE website (http://fcat.fldoe.org/rubrcpag.asp). Student papers are scored 1 through 6, with 6 representing the highest mark. In rare incidents, a student paper cannot be scored and is given a condition code. This includes cases where the student response is: blank (scored “B”), illegible (“IL”), off topic (“OT”), or given in a language other than English (“FL”).

Rangefinding

Two types of rangefinding committees are conducted for FCAT Writing. The primary purpose for both types of meetings is to work with Florida educators and Pearson in the creation of carefully compiled sets of student responses that will be used to train scorers. These scorers are hired to score the field test responses and/or the responses from each year’s operational administration.

After FCAT Writing prompts are field tested every few years, the scoring of a sample of student responses for each prompt is conducted in the spring. Under the supervision of the FDOE/TDC staff, a group of 6-10 Florida educators for each tested grade and purpose for which students are assigned to write (grade 4 – narrative or expository; grades 8 and 10 – persuasive or expository) reviews each response in the sample sets and comes to a consensus concerning its score. The scores on these responses are then considered as archetypal scores during operational scoring of all field test responses for each prompt. These archetypal scores form the basis for building scoring guides and other training materials that inform the raters (scorers) who score student work. The rangefinding committees provide decisions on how to score every score point for each prompt.

The scoring method used to score responses is called holistic scoring, meaning the total piece of writing is judged in terms of four writing elements: focus, organization, support, and conventions. The FCAT Writing rubric for each tested grade provides a scoring description for each of six score points and applies to all writing prompts administered at each grade. Application of the 6-point scoring rubric requires using the rubric in conjunction with an established set of papers that illustrate the range of quality allowed within each score point, i.e., a calibration scoring guide. The rubric provides general statements about how scores should be applied for each attainable level, but holistic scoring assumes that writing skills are closely interrelated, so scorers must consider the integration of the four writing elements in order to determine the score. The purpose of rangefinding is to clarify, for each prompt, the quality of writing allowed within each score point by aligning the response being scored to the responses in the calibration scoring guide so that consistent decisions can be made for all responses.

The rangefinding committees spend several hours first reviewing the responses in the established calibration scoring guides to internalize the characteristics of the examples for each score point. Then, sets of responses are provided to the educators by Pearson and TDC content specialists, who are also present at the rangefinding meetings to listen to the committee discussion and to record scores. It is critical that the committee provide decisions about each score point for each prompt and that these decisions be founded in actual student work.

Following rangefinding, scorer training sets are assembled. These training materials are provided to TDC staff for review and approval. Once the training sets are approved, scorers are trained and all field test responses are officially scored.

After all field test responses are scored, the resulting statistics are provided to FDOE for use in making judgments about the use of each prompt for the spring operational (census) FCAT Writing administration. At the annual prompt selection meeting, FDOE/TDC staff review the statistics as well as the content recommendations from Pearson’s scoring directors to choose the most appropriate prompt for each tested grade, 4, 8, and 10. All students in each tested grade are assigned the grade’s prompt, and the purpose for writing is not announced in advance of the assessment to avoid the possibility of narrowing instruction (teaching to the test).

Once prompts have been approved for the operational administration, Pearson’s content specialists begin preparation for fall operational rangefinding meetings, which follow the same process described above for field test rangefinder meetings. The difference between field test and operational rangefinding meetings is that for operational rangefinding, Florida educators score a sample of student responses for each grade’s operational prompt that will be administered to Florida students the following spring. The rangefinding process is particularly important so that all student papers are judged against a common standard; the scores from each year’s administration then result in a source of statewide information that can be used to characterize the performance of students on the writing assessment over time.

Recruiting Scorers

Recruiting is the responsibility of Pearson, which keeps a database of individuals who have scoring experience, including previous experience scoring FCAT Writing. The recruiting of scorers is done at the various regional sites where scoring is conducted. FDOE requires that each prospective scorer for writing has at least a bachelor’s degree in a related field of English, journalism, history, or psychology or have previous writing scoring experience. The number of scorers recruited for any project is based on the amount of time allocated for the scoring activity and the volume of scores to be assigned. Pearson recruits slightly more scorers than the projected need in order to accommodate for some attrition of scorers during the project.

Training

Scorers complete extensive training before they are allowed to participate in operational scoring. This training includes instruction on how to score FCAT Writing, several rounds of practice scoring, and a minimum of two rounds of qualifying scoring. Once scorers have demonstrated proficiency in scoring, they participate in several days of scoring warm-up, which is referred to as pseudoscoring. All training activities are supervised firsthand by FDOE and TDC staff.

Preparation of Materials

Training materials originate from field testing, rangefinding, and rangefinder review. Several sets of papers are assembled for specific purposes.

- **Instructional Materials:** Documents such as the rubric definitions of the scoring elements described in the rubric – focus, organization, support, and conventions, and an explanation of narrative, expository, and persuasive writing applicable to the prompt being scored.
- **Anchor Sets:** Papers that are labeled with the correct score and annotated for features that scorers must carefully review in order to assign the correct score. Anchor sets represent every score point and the range of quality within the score point. For example, a “high 4” indicates a paper that is a 4 but has some stronger characteristics. These papers help scorers understand the elements of student writing that demarcate the different score points.
- **Practice Sets:** Papers from only a few score points that have specific features that are used to instruct scorers on important aspects of FCAT Writing scoring. Each set has a different purpose.
- **Qualifying Sets:** Papers from all score points that are a randomly representative collection of papers.
- **Calibration Sets:** Papers from a single score point that focus on a very specific feature of student response that groups of scorers need to spend additional time attending to in their ratings.

Every document in the training set is reviewed and approved by FDOE/TDC staff. Creation of the training set may require several revisions, during which time Pearson’s staff implement a version control process.

Training Process

Training for FCAT Writing is conducted in person in a large room at each regional scoring location. The process includes four main steps that are identified in Figure 10-1. The training process begins with formal instruction for FCAT Writing by FDOE and Pearson staff. Prospective scorers are informed that they are not official scorers until they satisfactorily complete the rigorous training activities. The formal training is followed by several practice scoring opportunities where specific features of scoring are presented to scorers as opportunities to practice the concepts they were instructed on. After each round of practice scoring, every item in the practice set is discussed thoroughly in a large group. Scorers have ample opportunity to ask clarifying questions before moving on to additional activities. Pearson's staff members monitor the types of questions asked by scorers and at their discretion, or that of FDOE staff, additional training, or more focused training, can occur before entering the qualifying stage. Up to this point, scorer performance is not monitored.

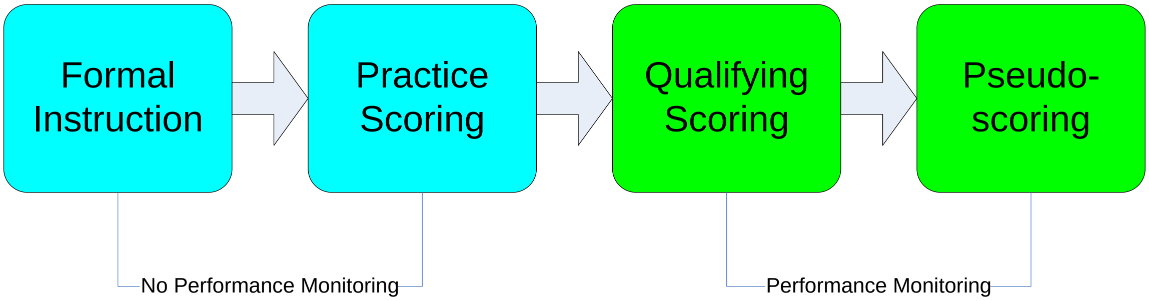

Figure 10-1. Steps in FCAT Writing Training Process

Qualifying

To qualify, scorers must meet a predefined quality level in order to become official FCAT Writing scorers. Each scorer takes up to three qualifying sets, depending on the quality of his or her scoring. Quality in this case is defined as the percentage of agreement between the scorer and the official score for the papers in the qualifying set. Scorers who have very low agreement are dismissed at this point and are not eligible to be FCAT Writing scorers for this administration. The remaining scorers are given a third qualifying set. If they meet the qualifying standards, they are declared "passed"; otherwise, they are dismissed.

Pseudoscoring

The final step in the training process includes several hours of practice scoring using operational papers from the current administration. This setting differs from earlier training activities as the complete operational process is used and most of the papers presented to scorers have not been previously scored either by a committee or Pearson scoring staff. In every way, Pearson attempts to create an environment where the scorers believe they are assigning real scores to student papers. However, the scores assigned are not retained during this phase. For this reason, this step is referred to as pseudoscoring. Pseudoscoring entails the scoring of a minimum number of papers, at which time FDOE reviews the scoring quality by verifying each scorer's agreement on pre-scored responses approved by the TDC content specialists, which are inserted into flow of "live" responses at an interval of 1:7. If initially unsatisfied with the quality of scoring as a whole during the pseudoscoring, FDOE may extend the session until it is satisfied with the quality. At the end of the pseudoscoring session, individual scorers demonstrating low scoring quality will receive some form of remediation. This could range from retraining on a particular scoring feature to constant monitoring by a supervisor until his or her quality numbers reach a certain level or the scorer will be dismissed and all applicable responses are reset during pseudoscoring.

Operational Scoring

During operational scoring each scorer assigns a single score to each student paper the scorer is presented. Scores available include a rubric value, 1–6, or a condition code (blank [B], illegible [IL], off topic [OT], or given in a language other than English [FL]), which is a value given to a paper that could not be assigned a valid score. The order in which the papers are presented to a scorer is partly randomized, with quality monitoring papers (see quality control section) inserted (see Figure 10-2).

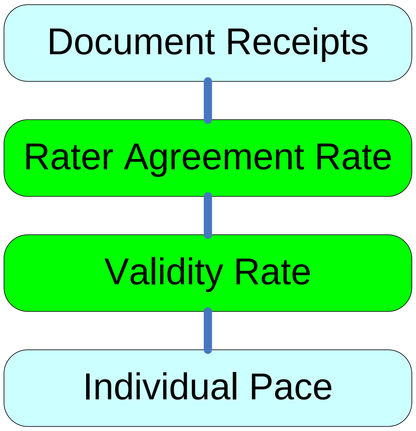

Figure 10-2. Factors Involved in Presentation of Papers to Scorers

In the figure, the light blue elements refer to factors that neither Pearson's scoring supervisors nor FDOE staff can directly regulate. These factors include the time when documents are received, scanned, and inserted into the scoring system. It also includes the pace of an individual scorer. However, if the pace of a particular scorer is falling considerably behind his or her peers, that scorer will be encouraged to work more quickly in a manner that does not affect the quality of his or her scoring. The green elements are quality monitoring tools that are completely within the control of Pearson and FDOE. The rater agreement rate measures how frequently each scorer matches the initial score of a previously scored paper. In this case, the scorer is acting as a second scorer in order to conduct score monitoring activities. The validity rate is the rate at which each scorer is given a paper with a known true score, which is called a validity paper. This validity check is done in order to monitor the ongoing quality of individual scorers. These activities are described in a section below.

Operational scoring does not begin until a large number of documents have been scanned and inserted in the system. Once scoring begins, each scorer is assigned either a paper where they are the first rater, a paper where they are the second rater, or a validity paper. For papers where a scorer is a first or second rater the presentation is made randomly. Validity papers are presented according to the order in which they are inserted into the system.

Scorer Monitoring
Scorer monitoring occurs from the very beginning of operational scoring activities. The performance of individual scorers is observed through various statistics described in the section below, and their scoring rate (number of papers rated per hour). Individuals who do not meet minimum thresholds for these statistics receive warnings, added monitoring, or remediation, depending on the degree by which they miss the mark. If a group of scorers, or the entire team, displays similar problems, Pearson staff may temporarily take the group off operational scoring and conduct a remedial activity. Typical remediation includes a discussion of the rules for scoring that are applicable to the problem area, and it may include scoring of a calibration set. See the section on preparation of materials above.

Quality Control

Several statistics during operational scoring are calculated to establish thresholds for statistical values that are used as triggers to conduct remediation activities. FDOE approves the use of all statistics well in advance of operational scoring, though modifications to the thresholds can be considered during operational scoring as the situation warrants. These decisions leading to the threshold values are informed by refereed literature, industry standards, consultation with FDOE's technical experts, the history of the program, and the field-test outcomes for the particular writing prompt.

Frequency Distribution Compared to Field Test

The distribution of scores on a prompt for the current administration is compared to the field-test administration of the prompt. Over the history of FCAT Writing, statewide operational performance on a prompt is similar to or higher than when the prompt was field tested.

Inter-rater Reliability (IRR)

Inter-rater reliability (also known as rater agreement) is a measure of how consistently the scorers are applying the scoring rules with one another. This is calculated as the percentage of perfect agreement between the scores of two independent scorers of the same paper. In 2014, 100% of student papers for FCAT Writing were scored a second time by a randomly assigned scorer. Since 1999, values ranging from 54% to 65% agreement have been observed for FCAT Writing. Historically, agreement is higher for grade 4 than grade 8, and grade 8 is higher than grade 10.

Validity Agreement

While inter-rater reliability measures the consistency of the scores among first and second scorers, validity measures the consistency of the scorers with a predetermined outcome. These validity papers are carefully selected based on how well they represent targeted features of papers and scoring decisions. To set "valid" scores, dozens of papers are chosen to assign "true" scores that are derived from the consensus score of two scoring supervisors. These “true” scores are then presented to FDOE staff for formal approval. Once formally approved by FDOE, the validity papers are inserted into the scoring system and given to the operational scorers in a blind fashion—meaning the scorers do not know they are reading a validity paper. Validity agreement is computed as a percentage of perfect agreement between the “true” score and the value given by the scorer.

Scorer Monitoring

Beyond operational scoring itself, scorer monitoring is the most significant activity during scoring. The statistics above are computed in real time by Pearson’s scoring systems. If individual scorers drift from the thresholds identified by FDOE, a series of specially designed interventions are implemented to improve the quality of the scores produced by each individual scorer. Monitoring of the scoring statistics can also lead to group-level interventions or even entail an intervention with the entire grade-level scoring team. These interventions range from a private warning given for individuals who may have drifted slightly away from the target in an easily identifiable manner to a retraining activity and even to a dismissal. Nevertheless, because of the demanding qualification process and rapid response to individual scorers who have drifted, dismissals are rare.

Aside from an individual warning, the most common intervention is a remedial training activity. Remedial training activities can be conducted individually or with groups of scorers. In a remedial training activity, the individuals involved are withdrawn, temporarily, from the scoring system and are not allowed to score operational papers again until they satisfactorily complete the training. Less structured remedial training activities may include review of the rubric, scoring notes, anchor papers, and perhaps a few additional papers selected by a scoring supervisor to highlight some concepts that the individual or group of scorers seems to be struggling with. The scoring supervisor discusses the situation with the scorer in question and allows that scorer to return to operational scoring once he or she orally demonstrates understanding of the scoring concepts involved.

If the identified problem is more systematic, then FDOE or Pearson’s scoring leadership may use a more structured approach, which includes the use of a calibration set. A calibration set is a collection of papers that contain a specific feature of scoring about which the group requires a better understanding. For example, perhaps there are more papers than expected where some scorers give *2*s and other scorers give *3*s. The calibration set may include three or more papers receiving both scores, for which the scoring supervisors are able to identify concrete features that should have caused raters to give one point or the other. This remediation is delivered by using the same instructional approach as conducted in the lesser structured activity. However, in this case, to be released back to operational scoring, each scorer must satisfactorily complete the calibration set with a minimal level of perfect agreement.

Security

Pearson's scoring facilities are closed to the general public. Pearson employees and FDOE and TDC staff must wear identification badges. Access is controlled by the security staff, receptionists, and programmed digital locks and/or card readers, as authorized by the site/general manager and specified in the site security plan.

Scorers assigned to the FCAT Writing program must sign a nondisclosure agreement before they can see any FCAT Writing materials. Furthermore, all materials provided to scorers are secured on-site by Pearson.

Finally, all operational scoring is conducted by using Pearson's image-based scoring system. This system is a computer-based application that operates over a secure network. Each scorer must log in with a unique ID and password. Only FCAT Writing scorers have access to the project. The image for scoring presented to scorers does not contain any identifying information about the student or the student's school or district.

Chapter 11. Quality Control Procedures

The FCAT 2.0 and EOC assessment programs and their associated data play an important role in the state accountability system as well as in many local evaluation plans. Therefore, it is vital that quality control procedures are implemented to ensure the accuracy of student-, school-, and district-level data and reports. Pearson has developed and refined a set of quality procedures to help ensure that all of FDOE's testing requirements are met or exceeded. These quality control procedures are detailed in the paragraphs that follow. In general, Pearson's commitment to quality is evidenced by initiatives in two major areas:

- Task-specific quality standards integrated into individual processing functions and services, and
- A network of systems and procedures that coordinates quality across processing functions and services.

Quality Control for Test Construction

Content Development

During the item development process, numerous steps are followed so that items will be developed to meet Florida's high standards. Multiple-choice items go through special steps to verify that they have a single correct answer and plausible alternate options (referred to as "distractors"). Performance tasks for the writing assessment and gridded-response items for mathematics (or fill-in response items for online tests), also entail special steps to verify that the rules for scoring the questions accurately identify all responses that are deserving of either partial or full credit.

A new development cycle is initiated for FCAT 2.0 and EOC assessments each year. Development cycles for writing (developing field-test prompts) are initiated once every two or three years. Before an item is used on an operational test form, it goes through 18–24 months of validation activities, including multiple reviews by content experts and validation by psychometricians. FDOE and TDC staff and Florida educators (in most cases) participate in these reviews. Students are never scored on any item unless Florida educators, FDOE, and TDC have already approved its wording, correct answer, and scoring rules.

Prior to the commencement of each development cycle, the item development plan (IDP) is created. This plan breaks out the number of items by subject, grade, reporting category, and benchmark that will be developed in that given cycle. This document is reviewed and approved by TDC to verify that it aligns with what is needed to build future tests given the content that already exists within the item bank. A variety of content foci are targeted when developing items so that there is diversity in item contexts.

As part of the quality processes, item writer training materials are developed and submitted to FDOE and TDC for review and approval. These materials and item specifications documents are provided to item writers in a yearly live training held by Pearson and overseen by TDC. This training facilitates the initial submission of quality multiple-choice items with solid distractors, or, in the case of constructed- and gridded-response items, clearly articulated scoring rules. All item writers are screened by Pearson and TDC to confirm that they meet the necessary requirements of expertise and experience required to write items for Florida.

Items are submitted by writers via an online item-authoring system that enforces many of the rules governing Florida items. Items are then submitted to a first pass review by Pearson's content specialists. After review, the items are accepted, rejected, or marked for revision and resubmission. Once an item is accepted by the content specialist, the item is reviewed again and revised by TDC's subject-matter experts to make the item adhere to Florida standards for item format (e.g., for multiple-choice questions, a single correct answer must exist and the incorrect options must be plausible given the varying levels of difficulty that the items are required to meet).

Before items are submitted to TDC for review, they go through a series of internal quality control and verification steps conducted by Pearson. Members of Pearson's internal review team generally do not write the initial version of items, which allows them to objectively evaluate the accuracy and quality of each item and its accompanying scoring rules. Items are senior-reviewed by the content lead for overall quality of the item, accuracy of content, benchmark match, grade appropriateness, and scoring rules (e.g., key or rubric). Cognitive levels are addressed in roughly the same proportion that they appear in the test-design summary for each content area.

Items, passages, and context-dependent sets are also sent through two additional groups, research librarian (RL) and universal design review (UDR), for review. The "context" is a rich scenario for which several items are developed. The "set" is the combination of the items and the scenario. Pearson's research librarians investigate the items, verify sources, and ascertain that the content of passages, items, and CD sets are valid. Items, passages, and CD sets are also reviewed by the UDR group to verify that they meet industry best practices for universal design.

Throughout these internal reviews, feedback is captured within the item-development system. Correctable flaws, such as UDR and RL violations, are corrected by the content specialists. Pearson's editorial team reviews items for grammar, punctuation, clarity, consistency, and concision and verifies that items adhere to Florida's requisite style guidelines. These edits are later incorporated by the content specialist if the edits are determined to be logical and do not impact content within the item. No item is submitted to TDC for review until all members of the content review team approve it. Numerous test questions do not pass the scrutiny of this internal review and are never submitted to TDC.

Florida Review Procedures

Once items have been initially accepted using Pearson's internal review process, they are formally submitted to TDC for review within the item-development system. The item-development software provides TDC content staff with secure, web-based access to the items and reduces or eliminates the need to print, package, and securely ship items from Pearson to TDC for review. Training is provided to TDC staff members prior to Pearson's first item submissions. TDC staff members may request additional one-on-one training assistance from Pearson's staff should the need arise. The training includes instructions for accessing, reviewing, and approving items; inputting feedback on items needing revision; and submitting results to Pearson for subsequent action. TDC content teams review items for benchmark match, content and grade appropriateness, single answer, and plausible distractors. Items that require further action are reviewed and resubmitted and subsequently approved or rejected for committee review. TDC editors may also review and comment on items in the item-development system during this review period. Their recommended edits are approved by TDC content staff before edits are incorporated.

FDOE and TDC, with support from Pearson, conduct numerous review meetings each year with Florida educators. The purpose of these meetings is to receive feedback on the quality, accuracy, alignment, and appropriateness of the passages, prompts, scenarios, and test items that are developed for the FCAT 2.0 and EOC assessments. Item review and content advisory committees are composed of Florida educators. The bias and community sensitivity review committees are composed of educators and other Florida citizens selected by the TDC staff. The meetings are held at various venues throughout Florida. The roles of these committees are described in the following sections.

Bias and Community Sensitivity Committees

As described in "Chapter 2. Development," bias committee reviews are performed to identify passages or items that in some way inadvertently demonstrate bias (e.g., racial, ethnic, gender, geographic, etc.). Additionally, community sensitivity committee reviews are performed to check for issues that might be considered especially sensitive, based on the wide range of cultural, regional, philosophical, and religious backgrounds of students throughout Florida.

During these reviews, each book of items and passages must receive a minimal number of reads, which is determined by TDC, based on the volume of material and the number of committee members. The reading assignments for each reviewer are organized so that each set is reviewed by a demographically representative sample of the committee members. Upon completion of the review assignments, reviewers sign in and return their review books, completing an affidavit indicating which sets of passages and items they reviewed.

During each bias and community sensitivity meeting, Pearson staff help track the number of reviews completed for each set of items or passages. An electronic file of reviewer comments is compiled, organized, and reviewed by TDC staff prior to or during content review committees. The data inform decision-making about suitability of passages and items for placement on future tests.

Item Review Committees
Items at each grade level and content area are presented to committees of Florida educators in three-ring binders and in an electronic format projected on a screen. For each item review committee meeting, a member of Pearson's staff keeps an electronic record of decisions made and documents any changes requested to item stems, options, art, or changes made to maintain alignment to the NGSSS. This electronic record is reconciled daily with the written record kept by the TDC staff member in charge of facilitating the meeting. Items are categorized as accepted, accepted with revisions, revised during the meeting, rejected, or marked to move grade (in cases where the item is rejected for the grade submitted but accepted for a grade in which it would fit the content specifications). Some items may be revised by an on-site Pearson staff member during the meeting for re-review by the committee. These revised items will be presented for review and approval to the TDC representative prior to the presentation to committee members unless arrangements are otherwise agreed upon by the TDC content lead and Pearson's content lead. Item binders for committee members do not have correct answers indicated. This allows committee members to "take the test" for each question. This strategy is designed to help identify any items with miskeyed answers, multiple correct answers, or ambiguous wording. TDC and Pearson staff members are provided with the correct answers in their notebooks.

Initial Publishing Activities

After Pearson's content staff have applied corrections to the items as indicated by item review committees, the items are "composed," meaning that the item is moved from the item-development system and changed to a format for online or print publication. An additional review of the items occurs at this time. This review permits Pearson and TDC content specialists and editors to perform a final review of the content, art, correct answer, and plausible distractors. It also allows the content and editorial teams to verify the correct rendering of the content. TDC provides Pearson with official approval of all items, passages, and context-dependent sets once items and passages appear in this composed format.

At this phase, publishing operations also conducts a quality step called "preflight" that electronically verifies that the new items comply with Florida's style requirements. Preflight also verifies that the embedded art objects conform to system requirements so they can be accurately transformed into a publishable format.

Test Construction Process

The selection of passages, context-dependent sets, and test questions appearing on all Florida tests is guided by two documents commissioned and approved by FDOE and

TDC: (1) Florida's test item specifications (available on the FDOE website), which outline the general blueprint for each test; and (2) the *Test Construction Specifications* (a secure document), which specifies the content and psychometric guidelines for test construction. The *Test Construction Specifications* document is drafted by the senior Pearson staff working on the project, including expert assessment content staff and doctoral-level psychometricians. This document extends the test specifications by providing the detailed guidelines, process, and clarification needed to select and place content on each test. The document is reviewed and edited by FDOE leadership, content, and psychometric staff. In the process of building a form, content specialists look at the parameters and then evaluate the existing item bank. Content specialists make sure reporting categories, benchmarks, and content foci are adequately represented within the assessment as outlined by the test blueprint and item development plan. Items are reviewed as part of the proposed form to avoid clueing. Once Pearson's content specialists have built a form, it is evaluated by Pearson's psychometrician. Pearson's content staff and psychometricians then engage in a collaborative and iterative process of refining the test form. Draft forms are not submitted to FDOE and TDC until they are approved by both Pearson's content staff and its psychometricians.

The FDOE/TDC review process is multi-staged in a similar manner to Pearson's process. Iterative reviews take place first between TDC and Pearson, and then between TDC, FDOE psychometric staff, and Pearson. FDOE and TDC provide tentative approval for forms in preparation for FDOE leadership review. A face-to-face review meeting is held to finalize the test build. FDOE and TDC staff—including content specialists, psychometricians, and FDOE leadership—confer with Pearson's staff to complete the test build. Refinement occurs until the best possible test forms are selected. Test forms must satisfy both content and psychometric criteria. All items appearing on a form, the sequence on the form, and the scoring rules are explicitly approved by FDOE and TDC.

Test Map Creation and Management

Throughout this process, Pearson, FDOE, and TDC scrutinize each item for correctness and verify that the correct answers/scoring rules are indicated in the item database. Once tests are approved by FDOE and TDC, Pearson prepares a test map (previously referred to as "test define" by FDOE), which is an electronic record of the items, their position on the test form, the key, and other scoring rules, as well as alignment to Florida's NGSSS. The test map is created by extracting the data elements from the electronic item bank, which is the original source of the scoring and alignment information. The test map is then verified manually against the composed test form.

Once complete, the test map is passed to Pearson's test map team (TMT). The TMT is responsible for inspecting the test maps, normalizing content for use in Pearson's systems, document control, and change management. The test map created for Florida serves as the source document for all Pearson publishing and scoring activities.

Inspection Process

Prior to receipt of the first test map, the TMT interviews Pearson's teams creating and using the test maps to obtain detailed internal requirements about the project. These requirements range from the data elements appearing in the test map, to how test forms will be composed, to the data elements that will be used by the various scoring and reporting systems. This information is used to verify that the test maps contain the information needed for internal users to fulfill their objectives.

Each test map is thoroughly inspected for valid values and the elements needed for internal Pearson groups to conduct their work. Some data elements prescribed by FDOE and TDC to be included in the test map are for Florida use only, and are not used by Pearson's systems. A series of reports are generated during the inspection process that are reviewed by the TMT and given to Pearson content specialists to validate:

> <u>Benchmark Report</u>
> This report lists the alignment codes reported in the test map and the number of items and points associated with those codes. This report is used to verify that the alignment codes conform to the specification, and that items and points conform to the blueprint.
>
> <u>Distinct Values Report</u>
> This report lists every value found in each column of the test map. This is valuable for identifying invalid characters. For example, data such as '1' (one) and 'l' (lower-case letter L) look very similar in some typefaces, and in some cases, '1' may be a valid character while 'l' may not be. If both characters are found in the same column they will both be listed in this report, thus making them obvious for visual inspection. This report is reviewed by the TMT to validate values needed by internal Pearson groups. Pearson's content team uses this report to verify the valid data elements uniquely used by FDOE and TDC.
>
> <u>Repeat Items Report</u>
> This report lists every test item that appears more than one time in the test map. In many cases, it is valid for the item to appear more than once. For example, a field-test item may appear on two forms. The repeat items report is used by the TMT and Pearson's content team to verify that the scoring and alignment information for repeated items is identical in all places. It is also used to confirm that changes made to content occur wherever the content is used.

Document Control and Change Management Process

Each new document that the TMT receives first goes through the inspection process in order to verify the contents. Once the content is verified, the document is made read-only and given a version number. From this point on, the TMT makes all changes to the test map. When a change is requested, a copy of the original version of the test map is saved to an archive folder for historical preservation. The requested changes are made to the original, and an electronic comparison between the original version and the archived copy is made. A change report is generated and inspected to verify that the change was made correctly and to confirm that no other changes were inadvertently made to the test map. The change report is then posted, as is the change request, for a second TMT member to validate. A message is sent to the change requester, providing a link on Pearson's network to the test map and change report so the change requester can verify completion.

Quality throughout Test Map Production

Special quality control steps are conducted at three separate times during test map production. The first of these inspections is detailed in the previous section. The next two events occur after the publishing operations group receives final approval to print the test forms. Step two is an electronic comparison of the test map to a data extract made from the test form. Embedded in the electronic files of Pearson's published content are data that are suppressed from the printed material. These data include several pieces of information about the test items on the form: the sequence number, the unique identifier, and the keyed response. Using a special script, the data are extracted from the electronic files and then compared to the test map. This comparison verifies that the test book and test map are in sync with one another. Any discrepancies are resolved immediately.

The third step is an electronic comparison between the test map and the results of the final "Taking the Test" step from the key verification process. This step verifies that changes made to content that affect the correct answers to the test, if any, are identified before printing, and the test maps are appropriately modified. Discrepancies are immediately resolved.

Only after these steps have been successfully completed are the test maps released for use by Pearson's scoring systems.

Content Monitoring of Quality during Forms Composition

Once items are selected for a test form, Pearson's content specialists review each publication-ready test form and compare the composed item in the test book against the previously approved version of the item and the test map. Pearson's editors also review the composed items against the previously approved items and item bank versions of the items at the initial round and compare the files to the test map to make sure that any changes requested by TDC have been incorporated correctly. In subsequent rounds, Pearson's editors compare edited items to the previous review round and continue verifying changes made to test maps. After each internal review by Pearson, the PDF files of the test forms are posted for TDC staff along with the test map. TDC editorial staff and content specialists review the composed forms along with the test map and mark up changes to items and test maps, as appropriate.

Pearson also sends the test forms to a subcontractor, EdExcel, during the review process so the materials will receive an external review. EdExcel specializes in conducting editorial reviews of assessment materials, and the company has nearly 25 years of assessment review experience. EdExcel performs an unbiased review of all Florida assessment materials, including test forms, interpretive products, and test administration manuals. EdExcel provides feedback about the items or materials and suggests edits based on the review. These edit suggestions are then sent to TDC for review and approval. Upon approval by TDC staff, the edits are incorporated into the forms by TDC editors.

Throughout this process, Pearson and TDC editorial staff perform careful cross-checks to make sure edits have been applied across test forms for items that appear on multiple forms and that edits appear on both the PDF files and test maps, if necessary. TDC's edits are then reviewed by Pearson's content and editorial team, and any outstanding queries are forwarded to TDC content specialists for resolution prior to sending the files to the final publishing preparation. If there are any changes to test maps, these are applied by the TMT at each round during the composition process. This is an iterative process, which does not stop until TDC provides approval for both the test forms and test maps.

Publishing Operations Electronic Checks

At the beginning of forms composition, Pearson's designers (desktop publishing experts) once again execute "preflight" checks that verify correct application of Florida style, such as line weights, RGB color model, and fonts. If any potential quality issues are found, they are corrected using the processes for notification outlined in the *Florida Statewide Assessments Production Specifications*. This guide is created with TDC and FDOE input and updated yearly.

After FDOE and TDC approve a test form to be printed, Pearson's designers create the final print-ready file by running the form through an automated print-ready file creation system that outputs a composite proof, a registration proof, and a separation proof. The publishing operations staff verify that each of these documents match client specifications.

Print Procurement

Quality Performance for Accurate Test Documents

Because of the absolute need for accurate test documents, Pearson conducts an annual review of print vendors. Variances are tracked and reviewed as part of a collaborative support program aimed at maximizing quality, accuracy, and performance.

Security

Pearson has a long-established audit program and partnership with a select pool of print suppliers. These suppliers must meet a series of stringent security protocols in order to work on Pearson's material. All files, film negatives, and plates are maintained in secure locations at the print supplier, with only authorized personnel permitted access to the material. All plates and film negatives are securely destroyed by the print supplier upon completion of a contract. At the end of each day's print run, authorized personnel securely shred all press overages and waste material. Each production run is made under close supervision of the printing supervisor. Test material is kept secure at all times to preserve its integrity.

Quality Assurance
Pearson has specifically chosen suppliers that will enhance the program quality. As a requirement of Pearson, each supplier has implemented several additional quality procedures, including but not limited to inspecting additional sheets/forms off press throughout each print run and installing an electronic signature recognition system into their binding process. These additional quality steps are designed to mitigate defective products by eliminating miscollations. Pearson's internal print facility is ISO certified. Pearson also works with outside print suppliers that are ISO certified.

Printing
Once the printer proofs are approved, another quality check is performed during the creation of press plates and at press to verify that no data have dropped from the document. The press operators are required to pull a specified number of print sheets to verify registration, pagination, print quality, and color consistency. Signatures coming off the press are stacked on a skid and tagged with a color-coded tag that identifies all signatures to verify a positive identification during transport to the next production station.

Bindery
Each print supplier certified by Pearson, including its own internal print facility, has installed an electronic signature recognition system to prevent miscollations within a test booklet. The supplier's system will either electronically read a small bar code on the first page of each signature of a booklet uniquely coded to that specific signature or will read a predetermined image zone of a page to confirm that the correct signature is being processed. Any booklet that contains miscollated pages or a missing signature will cause an automatic shutdown of the bindery equipment for proper corrective action. During the binding process, a set number of books coming off the conveyor belt are pulled to confirm that proper quality control standards are met. As an additional quality measure, corresponding colored tags are placed on each individual pocket on the bindery equipment for visual check for accurate collation of materials. These materials are physically inspected by quality checkers within the manufacturing facility, and a predetermined number of these inspected booklets are sent to Pearson for review and approval. Pearson has developed these quality assurance standards, which exceed standard industry practice, to provide the highest confidence in the quality of Pearson's products.

Key Verification Process

Pearson mandates a four-phase key verification process for all test construction and publishing. These steps are essential to verifying that correct answers are identified in test maps.

Taking the Test during Test Construction
The first verification phase in the process occurs during test construction. While the test questions are being selected, the content staff working on the project scrutinize every selected question for the identified scoring information. The correct answer is confirmed for every question, and each incorrect option is verified as incorrect. If discrepancies are discovered, the issue is first reconciled with the item bank to verify that the comparison data are correct. Additional follow-up is made with FDOE and TDC if the question is flawed. In some cases the question is corrected, while in others, FDOE and TDC decide to change the status of the item to "do not use", if the item is a FT item.

Creating the Test Map from the Item Bank
The second verification phase occurs when the test maps are created. Test maps are created electronically using the master database of information about the test questions rather than key-entering data into a spreadsheet. This reduces or eliminates the opportunity for key entry problems. The test maps are then checked by the TMT and content specialists for accuracy.

Taking the Print-Ready Test
When the electronic publishing files are sent to the printer, examination copies are provided to a team in Pearson's content group. This is the third phase of verification. Two content experts who do not work on the Florida project are provided with a test and an answer document. They read the test questions and respond to them by entering their answers on the answer document. If correct answer discrepancies are found during this process, they are reported immediately to Pearson 's Florida content team and reconciled with the TMT. The TMT conducts an electronic comparison between the test map and the answer document used for this exercise. This is to verify that all issues have been reported. The TMT does not release the test map for scoring until evidence is documented that resolution has taken place.

Statistical Key Check
After the test has been administered, student responses to the items are evaluated statistically to identify the presence of statistical flaws in the outcomes. Details on this fourth phase of the verification process are provided in a subsequent section of this chapter.

Quality Control for Non-Scannable Documents

Pearson contracts with outside vendors for the printing of non-scannable documents because of the large volume of printed materials necessary for the FCAT 2.0 and EOC assessment programs. To ensure the accuracy of these documents, Pearson holds periodic meetings with all of their printers to reiterate the high expectations for printing quality and to remind them of the penalties associated with the failure to perform to standards. The following quality controls are implemented to facilitate the successful performance of outside printing companies.

- Pearson provides design and schedule requirements to printers well in advance of the delivery of copy so that the schedule for printing can be arranged.
- If any changes are made by FDOE and Pearson with regard to a print schedule, the printer is notified immediately.
- Corrections submitted by FDOE are added to any of the corrections Pearson sends to the printer.
- All page proofs, final proofs, and specimens of printed materials are proofread in their entirety by the forms support department and are submitted to FDOE for review.
- Sample printed materials are examined for the required paper type, ink color, collation, and copy. If discrepancies are noted, the printer is notified immediately to make allowances for corrections and reprint where required.
- Whenever possible, electronic transfer of copy is used to minimize human error and to expedite the printing process.

An additional quality check of all outside printing materials is made by Pearson during the packaging operation. Each box of materials is spot-checked to verify printing and collating accuracy.

Quality Control in Data Preparation

To ensure an accurate accounting of the assessment documents that Pearson receives, data preparation staff perform a series of receipt and check-in procedures. All incoming materials are carefully examined for a number of conditions, including damage, errors, omissions, accountability, and secured documents. When needed, corrective action is promptly taken according to specifications developed jointly by Pearson and FDOE.

Quality Control in Production Control

Pearson uses the “batch control” concept for document processing. When documents are received and batched, each batch is assigned an identifying number unique within the facility. This unique identifier assists in locating, retrieving, and tracking documents through each processing step. The batch identifying number also guards against loss, regardless of batch size.

All FCAT 2.0 and EOC assessment documents are continually monitored by Pearson’s proprietary computerized workflow management system (WFM). This mainframe system can be accessed throughout Pearson’s processing facility, enabling its staff to instantly determine the status of all work in progress. WFM efficiently carries the planning and control function to first-line supervisory personnel so that key decisions can be made properly and rapidly. Since WFM is updated on a continuous basis, new priorities can be established to account for documents received after the scheduled due date, late vendor deliveries, or any other unexpected events.

Quality Control in Scanning

Pearson has many high-speed scanners in operation, each with a large per-hour scanning capability. Stringent quality control procedures and regular preventative maintenance ensure that the scanners are functioning properly at all times. In addition, application programs consistently include quality assurance checks to verify the accuracy of scanned student responses.

Through many years of scanning experience, Pearson has developed a refined system of validity checks, editing procedures, error corrections, and other quality controls so that there is maximum accuracy in the reporting of results. During scanning, assessment documents are carefully monitored by a trained scanner operator for a variety of error conditions. These error routines identify faulty documents, torn and crumpled sheets, document misfeeds, and paper jams. In these events, the scanner stops automatically. The operator can easily make corrections in most cases; otherwise, corrections will be made in the editing department.

Expected results are created in the format that is expected to be received from the scanning system. Once the test deck is scanned, expected results are compared to the scan file to verify there are no discrepancies. The following are included in this validation:

1. Every response (bubbles and write-in boxes) on every header and answer document is hand-gridded using pre-defined patterns to verify that all data are captured properly and each response has no impact on other responses.
2. Multiple marks (double-grids) are hand-gridded to verify that the scanners correctly identify when more than one response is received.
3. Pages are extracted for every header and answer document. This verifies that the scanner outputs the data properly when pages are missing.
4. Every header and answer document is scanned with no responses gridded. This ensures that data are captured properly when processing blank documents.
5. Every bar code is scanned for each header and answer document. This verifies that all bar codes are being captured properly.
6. An answer document that cannot be scanned (cut corner of sheet) is included. This verifies that the scanner will recognize when an answer document cannot be scanned.

Quality Control in Editing and Data Input

Data files are created to simulate the data coming out of the scanning system. These data contain both positive and negative (including “edge” conditions) test cases for every edit rule. Expected results are created for each test case. The data files are processed through the editing system and the output is compared to the expected results to verify there are no discrepancies.

Assessment & Information Quality

Test cases are designed to verify processing and editing of paper material is performing as intended. These predefined cases, called mock data, are populated on a sample of answer sheets or scannable books that are hand-gridded. Some of the mock data is auto-gridded on the answer sheets by predefined software. All software and interfaces are utilized and executed in the same manner that will process live data. The data processed through this system are generated from the material distribution phase. The edited data that are generated out of this system are used to test all the downstream systems (scoring and report distribution). The following are within scope of these testing activities:

End-to-End Testing

1. Every field for every answer document and header is hand-gridded with predefined patterns to verify that all data are being captured, edited, and reported accurately and that no field has an impact on another.
2. Every answer document is hand-gridded using the max field lengths (all values populated for each field) to verify that all gridded values are picked up by the scanners, edited, scored, and reported properly.
3. Every answer document is hand-gridded with all item responses to verify that all gridded responses are being captured, edited, scored, and reported properly.
4. Every answer document is gridded with no item responses to verify data are captured, edited, scored, and reported properly when processing blank item responses.
5. Every field is hand-gridded with “edge” values (e.g., 0 and 9; A and Z) to verify proper data capture, editing, scoring, and reporting.
6. Cases to check the document count form processing (covering all grids, including special document types).
7. Cases to check errors on the document count form (blank DCFs, double-grids).
8. Cases where all bubbles are gridded to be sure they are being scanned properly, including each student demographic field, form field, accommodation fields, and item responses (see scannable document configuration chart and examples).
9. Cases to check multi-marks as well as missing data for each field.
10. Cases loaded based on the PreID files as well as cases that are added into the computer-based system by the school coordinators.
11. Cases to check procedures for duplicate testers in the PreID file (a student with the same last name and SID with two answer documents for one subject and grade level).
12. Cases to check all other score flag scenarios.
13. For each grade level and subject, cases to check that the PreID label information, if not blank, is overriding any gridded information for the birth date, gender, race and ethnic codes, primary exceptionality, ELL, and Section 504 data.
14. Cases to check procedures for duplicate testers after processing across both paper and computer versions if both are administered (may be computer generated duplicates).

15. Cases to check RMS rules for not meeting attemptedness, i.e., a student who answers fewer than 6 questions (may be computer generated; Score Flag = 2).
16. Cases for a variety of score ranges including raw scores of zero, perfect scores, and scores on each side of cut scores (may be computer generated).
17. Cases for each achievement level (may be computer generated; 100% in any one achievement level within a school).
18. Cases to check all aspects of reporting student scores and aggregated scores, including all flags used on the data files (may be computer generated, using enough student records that results are displayed and not suppressed on mock reports; more than 9 reportable students in school).
19. Cases to check each item response area, including tracking of changed responses; some items should have erasures with another response gridded and some with just erasures.
20. For each grade level and subject, cases to check that history data are properly merged for each grade level (may be electronically generated; includes grades 3–9 only).
21. For each grade level and subject where the reading answer document is separate from the mathematics answer document, cases to check that the reading and mathematics records are merged properly (see scannable document configuration chart).
22. Cases to check that all PreID and security barcode information are being recorded properly (verifies PearsonAccess is bringing in correct data).
23. Retake cases to check that gridded grade level can override the grade level on a PreID label (check cases where retake reading and retake math grade levels conflict, and where DCF grade level and scannable documents grade levels conflict).
24. For each grade level and subject, cases to check that the aggregated suppression rules are being applied correctly.
25. For each grade level and subject, cases to check proper rounding on aggregated reports (may be computer generated).
26. For each grade level and subject, cases to check that special school types and district numbers are being processed correctly (may be computer generated).
27. For each grade level and subject, cases to check that school data are being properly aggregated to district data (see item 20).
28. When appropriate, cases to check that items answered in the computer-based tests in a non-sequential manner have responses associated with the correct item number in the test.
29. When appropriate, cases to check that paper-based and computer-based records for the same student are merged and reported properly.
30. For retakes, cases to check that the APS file is being used properly. A mock APS file may be generated (may be computer generated).
31. Cases to check all hand-scoring condition codes and rules (may be computer generated).
32. Cases to check hand edits resulting from torn books, unreadable barcodes, missing pages, pages out of order, etc.

33. All scanning and editing outputs are validated against expected results to verify the material is being processed correctly.

Production Validation

1. Once live material is received from the systems, a sample of the material is selected. This sampling contains all answer documents and headers.
2. Once this live material has been scanned, a sampling of each header and answer document (both hand-gridded and PreID) is selected for validation.
3. Every response for every field for each of the samples is manually compared to the scanning output to verify that all data are captured properly.
4. Once the scanning validation is complete, the Assessment & Information Quality tester approves the documents to be edited.
5. Once editing is complete, the edits performed on the headers and answer documents are manually validated to verify that all edits performed are accurately reflected in the edited file.

Quality Control in Performance Scoring Center – Writing Tasks

Quality control permeates all steps of the writing task scoring process. It starts with a scorer recruiting and screening process designed to locate and employ the most highly-qualified individuals available. Scorers receive careful, exacting, and thorough training in the specific items and rubrics at the beginning of each scoring project, regardless of their previous scoring experience. Training is provided by Pearson's staff who, after fulfilling rigorous internal guidelines for presentation skills and knowledge, have become qualified trainers. During scoring, scorers are constantly monitored to ensure they are scoring accurately and consistently. More complete details regarding Performance Scoring Center (PSC) quality control procedures for constructed-response scoring are presented in Chapter 10.

Quality Control for Test Form Equating

As discussed in Chapter 7, calibration, scaling, and equating activities are the scientifically based analyses that lead to reported scores. Pearson's psychometricians create detailed specifications several months in advance in order to facilitate planning and communication among the various participants. These specifications originate from the prior year's documentation and are carefully scrutinized for any special considerations that must be attended to for the upcoming administrations. Details in the specifications include the following: file naming conventions, secure file posting directions, formats for input and output files, detailed directions and formulations for statistical analyses, and the responsibilities of each participant. All parties to the psychometric analyses are given ample opportunity to review and comment on the specifications, with clarifications made as needed. Further planning and preparation does not begin until all parties agree to the processes and details described.

Every process used by Pearson's psychometric team is transparent to the FDOE, as is the source code for software that is uniquely written for Florida to manage and analyze results. Pearson, Pearson's psychometric subcontractor (HumRRO), and FDOE use the commercial statistical software platform SAS to manage and conduct many of the analyses. Both Pearson and its subcontractor provide actual SAS source code to FDOE psychometricians for inspection and validation each year. In addition, the commercially available MULTILOG software is used to estimate (calibrate) Item Response Theory (IRT) item parameter estimates that are used for item analysis and equating.

Each winter, after the specifications have been thoroughly vetted, the calibration teams (Pearson and its subcontractor) deliver preliminary versions of the software to FDOE and all parties engage in a process practice session. This session uses mock student data that are formatted to the FDOE file specifications and generated to be a realistic facsimile of student responses in order to validate the processes and software that will be used in the analysis of the following spring student results. This practice session provides all parties with an opportunity to fine-tune their roles and systems for the upcoming spring and gives FDOE an opportunity to evaluate the effectiveness of the process and each participant. No work with actual student data commences until the parties resolve any discrepancies encountered in the practice session and FDOE has approved the final readiness of the process and software.

In addition to the external replication conducted by the subcontractor, there are three replication/oversight activities conducted. First, an external reviewer (in 2014, the Buros Institute from the University of Nebraska at Lincoln) observes all communication and reviews all documentation, including the specifications and results. The external reviewer provides real-time consultation to FDOE, if needed, and a full report on the equating activities. Second, FDOE psychometricians execute the software provided by the subcontractor and Pearson to verify that it ran correctly, and FDOE psychometricians themselves engage in additional analyses to validate outcomes. Third, Pearson psychometricians conduct a series of replication and diagnostic activities for quality assurance:

- All student responses are rescored by contractor data analyst staff using the source keys. These rescored responses are compared to the official scores. If discrepancies are discovered, they are reconciled or corrected before data are turned over to FDOE or the subcontractor.
- A statistical key check is conducted to identify items that may require closer review. This check is an extension of the key verification process discussed in "Chapter 2. Development." The goal of the statistical key check is to use statistical procedures and generally accepted criteria to identify items that are not performing to expectations. Specifically, MC items with one of the following characteristics are flagged:
    - p-value < 0.15
    - p-value > .90
    - Item-total score correlation < 0.20
    - Incorrect option selected by 40% or more students

  - p-value on any one form differs from the population p-value by |0.08| for operational items which appear on multiple forms
- Any flagged items above are reviewed to ensure the item was correctly printed. Also, flagged items have keys checked by Pearson and FDOE content staff to certify the key is the correct answer. The statistical key check and all follow-up are completed before equating begins.
- Key elements of psychometric processes are conducted independently by two psychometricians from Pearson, and discrepancies, if discovered, are reconciled before results are posted or discussed with FDOE.
- After the program MULTILOG is run, a plot of the empirical item response distribution is overlaid on the plot of the model fitted item characteristic curve. If the two plots are dissimilar, the MULTILOG implementation is checked to verify it was executed correctly. If it was, and very poor model-data fit is encountered, the issue is brought to the calibration team to discuss action.
- The item statistics from the current administration are compared to their item bank values. If differences are large, psychometricians consult with contractor content staff to determine if the question was changed in some way, or if the position of the item is very different between administrations. If issues are encountered, they are shared with the calibration team to discuss action.
- The percentage of students classified in each level of achievement is compared to the historical trend as a final step in validating the results. If the current year appears to be off-trend, the steps are retraced to verify they were conducted correctly, and cohort comparisons are made to determine if the trend is visible from a different data view. If issues are encountered, they are shared with the calibration team to discuss.

Independence and Oversight

The four parties participating in the psychometric analyses work both independently and collaboratively. Final decisions are the exclusive purview of FDOE. The detailed roles of each participant are as follows:

Pearson

1. Writes and maintains specifications and SAS code.
2. Conducts the primary inspection of the data to determine if the information is accurate so that psychometric activities can be initiated.
3. Conducts the statistical key check.
4. Responsible for communication management. Sets up and leads daily conference calls during analysis window. Makes sure all parties are aware of issues and decisions.
5. Posts data to secure FTP, and maintains secure FTP.
6. Conducts and posts all computational results first.
7. Conducts interpretive analysis, and provides professional judgments about various solutions.
8. Investigates anomalous or unexpected results to verify correctness of data or outcomes.

9. Produces the official production scoring documents.
10. Archives all final documents, beginning specifications for next year as needed.

FDOE

1. Oversees all Pearson and subcontractor work; approves all specifications.
2. Evaluates all outcomes. Judiciously replicates computations and explores alternative solutions to validate final decisions.
3. Seeks professional advice from Pearson, the subcontractor (HumRRO), and process reviewer (Buros Institute).
4. Confirms that Pearson's and subcontractor's results match. Oversees resolution process when they do not.
5. FDOE leadership, psychometricians, and content experts comprehensively evaluate all of the scaling and equating solutions, considering numerous factors related to content and statistics, and make a final decision. Verifies that the official scoring documents created by Pearson are accurate and reflect the decisions made by FDOE.
6. Replicates the final reported score computations using Pearson's and subcontractor's scoring programs before scores are reported.
7. Documents rationale for final decisions.

Psychometric subcontractor (HumRRO)

1. Reports directly to FDOE. Provides independent advice, review, and replication of results using independently created systems.
2. Reviews and provides feedback on specifications.
3. Conducts data inspection and statistical key check.
4. Replicates all computational activities and posts after Pearson.
5. Conducts independent professional evaluation and provides professional judgment to FDOE.
6. Compares official scoring files to independently generated files.
7. Replicates the final reported score computations before scores are reported.

Process Reviewer (Buros)

1. Reviews and provides feedback to FDOE on the specifications.
2. Attends conference calls.
3. Evaluates the process, and provides a written report to FDOE about the effectiveness of the process.
4. Replicates the analyses for selected components.

Triangulation and Depth of Investigation

Throughout the process, the psychometric parties provide their own solutions and professional judgment, presenting solutions to each other and to FDOE. Computational procedures are compared and they are all required to meet a demanding level of tolerance. Results not matching are painstakingly explored until identical results are achieved (in almost all cases), or the reason for the mismatch is ruled as immaterial by the entire team (this is a rare outcome). The independence of each of the three parties engaged directly in computational activities provides confidence that the best solutions are actualized. The collegial and collaborative approach that the parties take brings resolutions more quickly and minimizes communication problems.

Through the equating process, when judgments must be made, all participants thoroughly investigate the possible solutions in order to provide FDOE with the most complete information with which to make final decisions.

Quality Control in Scoring and Reporting

Final Score Replication

After final scores are assigned by Pearson's scoring system and validated by the IT Assessment Validation (AV) and Assessment and Information Quality (AIQ) groups, two additional validations are completed by the psychometric groups. First, Pearson's psychometric group reassigns the reported scores using the source final scoring files (files approved for use by FDOE) and PC version of the scoring program used by the scoring system. This provides an internal contractor replication of the reported scores. Third, the psychometric subcontractor receives the same data and reassigns the final scores using their final scoring files, and an independently written scoring program. The psychometric subcontractor results are reported directly to FDOE. Third, FDOE replicates scoring using both Pearson's and the subcontractor's scoring programs. If discrepancies are found, FDOE and the subcontractor confirm findings. The differences in scale score calculations are evaluated by FDOE's psychometric team and resolved before scores are reported.

Scoring and Report Distribution

IT Assessment Validation

Test cases are designed to test each component of scoring independently to verify that all data are captured properly and every scoring rule is tested thoroughly.

Assessment & Information Quality

Test cases are designed to verify that scoring and reporting of testing records are performing as intended. All software and interfaces are utilized and executed in the same manner as used for live data. The data processed through this system are generated from the material distribution and data capture and processing phases. All scoring and reporting outputs (including data files, paper reports, and electronic reports) are validated against expected results to verify scoring and reporting are accurate.

Annotated Table of Contents

Item Response Theory Statistics: Difficulty, Discrimination, and Guessing
Differential Item Functioning (DIF) Statistics

Item Bank

Test Construction *(Standards: 3.1, 4.11)*
Form Design
Guidelines for Item Selection
Application of Item Response Theory

Quality Control in Test Construction *(Standards: 4.11)*

Chapter 3. Administration

Eligibility

Administration Procedures *(Standards: 5.1)*
Administration Schedule
Timed Assessment Sessions

Accommodations *(Standards: 5.1, 5.3, 5.4, 5.5, 5.6, 9.4, 10.1, 10.2, 10.4)*
Students with Disabilities
Special Accommodation for Computer-Based Testing
English Language Learners

Test Security *(Standards: 5.7)*

Data Forensics Program *(Standards: 5.6)*

Chapter 4. Reports

Appropriate Uses for Scores and Reports *(Standards: 1.1, 1.2, 4.1, 4.3, 7.8)*

Reports Provided *(Standards: 1.2, 4.1, 4.3, 5.10, 5.13)*
Individual Student Reports
Parent and Student Report
School Report of Students
District Report of Schools
District Summary Report
State Report of Districts
State Summary Report

Description of Scores *(Standards: 4.1, 4.3, 4.9)*
Subscale Raw Score *(Standards: 4.1, 4.3)*
Scale Score *(Standards: 4.1)*
Achievement Levels *(Standards: 4.9)*
Historical Scores

Appropriate Score Uses *(Standards: 1.1, 1.2, 4.1, 4.3, 7.8)*
Individual Students *(Standards: 1.2, 4.1, 4.3, 7,8)*
Groups of Students *(Standards: 1.2, 4.1, 4.3, 7,8)*

Cautions for Score Use *(Standards: 1.11, 2.1, 4.1, 4.3)*
Understanding Measurement Error *(Standards: 2.1)*
Using Scores at Extreme Ends of the Distribution *(Standards: 1.11, 4.3)*
Interpreting Score Means *(Standards: 4.1, 4.3)*
Using Reporting Category Information *(Standards: 4.1, 4.3, 4.4)*

Scaling Formulas

Vertical Scale *(Standards: 4.2, 4.14)*

Data Collection Design
Development of the Vertical Scale
Properties of the Vertical Scale

Scale Drift *(Standards: 4.17)*

Chapter 7. Equating

Rationale *(Standards: 4.10)*

Pre-Equating *(Standards: 4.11, 4.12)*

Post-Equating *(Standards: 4.11, 4.12)*

Item Sampling for Equating *(Standards: 4.13)*
Student Sampling for Equating *(Standards: 1.5, 3.8)*
Operational Equating Procedures *(Standards: 4.15)*

Sources of Error Variances to Score Instability *(Standards: 4.11, 4.12)*

Sampling Error *(Standards 4.12)*
Equating Error *(Standards 4.11)*
Other Possible Sources of Error

Chapter 8. Reliability

A Mathematical Definition of Reliability

Estimating Reliability *(Standards: 2.1, 2.4, 2.5)*

Test-Retest Reliability Estimation
Alternate Forms Reliability Estimation
Internal Consistency Reliability Estimation *(Standards: 2.6, 2.7, 2.11, 2.12)*

Standard Error of Measurement *(Standards: 2.2, 2.3, 2.4, 2.5)*

Use of the Standard Error of Measurement
Conditional Standard Error of Measurement *(Standards: 2.14)*
Measurement Error for Groups of Students
Standard Error of the Mean *(Standards: 2.19)*

Scoring Reliability for Written Compositions *(Standards: 2.1, 2.4, 2.10)*

Reader Agreement *(Standards: 2.10)*
Score Appeals *(Standards: 8.13)*

Student Classification Accuracy and Consistency *(Standards: 2.1, 2.4, 2.14, 2.15)*

Accuracy of Classification
Consistency of Classification
Accuracy and Consistency Indices
Accuracy and Consistency Results

Chapter 9. Validity

Perspectives on Test Validity *(Standards: 1.1)*

Criterion Validity *(Standards: 1.13, 1.14, 1.16)*
Content and Curricular Validity *(Standards: 1.6)*
Construct Validity *(Standards: 1.2)*

Validity Argument Evidence for the Florida Assessments

Scoring Validity Evidence *(Standards: 1.11)*
Generalization Validity Evidence *(Standards: 1.5, 1.7, 1.12)*
Extrapolation Validity Evidence *(Standards: 1.8, 10.11)*
Implication Validity Evidence
Summary of Validity Evidence *(Standards: 1.3)*

Chapter 10. Constructed-Response Scoring

Scoring Process

Rangefinding *(Standards: 1.7, 3.22)*
Recruiting Scorers *(Standards: 1.7, 3.23)*
Training

Quality Control *(Standards: 2.10, 5.8, 5.9)*

Frequency Distribution Compared to Field Test
Inter-rater Reliability (IRR)
Validity Agreement
Scorer Monitoring

Security *(Standards: 5.7)*

Chapter 11. Quality Control Procedures

Quality Control for Test Construction *(Standards: 3.1)*

Content Development *(Standards: 1.7, 3.6, 3.7, 5.9, 7.4, 7.7)*
Florida Review Procedures *(Standards: 1.7, 3.6, 7.4)*
Initial Publishing Activities
Test Construction Process *(Standards: 3.1, 4.11)*
Test Map Creation and Management
Content Monitoring of Quality during Forms Composition
Publishing Operations Electronic Checks
Print Procurement
Key Verification Process

Quality Control for Non-Scannable Documents

Quality Control in Data Preparation

Quality Control in Production Control

Quality Control in Scanning

Quality Control in Editing and Data Input

Assessment & Information Quality

Quality Control in Performance Scoring Center – Writing Tasks *(Standards: 5.9)*

Quality Control for Test Form Equating *(Standards: 4.10)*

Independence and Oversight
Triangulation and Depth of Investigation

Quality Control in Scoring and Reporting *(Standards: 5.8)*

Final Score Replication
Scoring and Report Distribution

References

AERA/APA/NCME. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* New York: Wiley.

Bradley, E. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26.

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Council of Chief State School Officers (2011). *Addressing two commonly unrecognized sources of score instability in annual state assessments*. Retrieved from: http://www.ccsso.org/Documents/2011/Addressing%20Two%20Commonly%20 Unrecognized.pdf.

Chien, M., Hsu, Y., & Shin, D. (2006). *IRT score estimation program* [computer program]. Iowa City, IA: Pearson.

Cizek, G. J., & Bunch, G. J. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage Publications.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.). *Educational measurement* (4th ed., pp. 355–386). Westport, CT: American Council on Education/Praeger.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–47.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation for the behavioral sciences.* New York: Wiley.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Cronbach, L. J. (1990). *Essentials of Psychological Testing* (5th Ed.), Harper & Row, NY.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dudek, F.J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86*(2), 335–337.

Education Accountability Bill of 2010. S.B. 4, Florida 2010 Legislative Session. (2010).

Education Bill of 2008. S.B. 1908, Florida 2008 Legislative Session. (2008).

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.

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

Florida Department of Education. (2004). *Assessment & Accountability Briefing Book: FCAT—School Accountability—Teacher Certification Tests*. Tallahassee, Florida: Author.

Florida Department of Education. (2007). *Assessment & Accountability Briefing Book: FCAT—School Accountability—Teacher Certification Tests—2007*. Tallahassee, Florida: Author.

Florida Department of Education (2011). *2011 FCAT 2.0 vertical scaling report.* Tallahassee, FL: Assessment and School Performance.

Hambleton, R., & Plake, B. (1997). *An anchor-based procedure for setting standards on performance assessments*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Haberman, S., & Dorans, N. J. (2009). *Scale consistency, drift, stability: Definitions, distinctions and principles.* Paper presented at the National Council on Measurement in Education, San Diego, CA.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Structural Equation Modeling, 6*(1), 1–55.

Jolliffe, I.T. (2002). *Principal Component Analysis* (2nd ed.), Springer, NY.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education and Praeger Publishers.

Kim, S., & Kolen, M. (2004). *STUIRT* [computer program]. Iowa City, IA: The University of Iowa.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer.

Lee, G., & Lewis, D. (2001, April). *A generalizability theory approach toward estimating standard errors of cut scores set using the Bookmark standard setting procedure.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Lee, W., Hanson, B. A., & Brennan, R. L. (2000, October). *Procedures for computing classification consistency and accuracy indices with multiple categories.* (ACT Research Report Series 2000–10). Iowa City, Iowa: ACT, Inc.

Linn, R. L. (1993). Linking results in distinct assessments. *Applied Measurement in Education, 6*(1), 83–102.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessing in teaching* (7th ed.). New Jersey: Prentice-Hall Inc.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 179–197.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Michaelides, P. M. (2003). *Sensitivity of IRT equating to the behavior of test equating item.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service, Policy Information Center.

Muraki, E. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

Muthén, L. K., & Muthén, B. O. (2007). M*plus* (Version 6.11) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Phillips, G. W. (2011). *Score drift: Why the results of large scale testing programs bounce around from year to year*. (Under journal review).

Qualls, L. A. (1995). Estimating the reliability of a test containing multiple item formats, *Applied Measurement in Education, 8*(2), 11–120.

Reckase, M. D. (2010). *Study of best practices for vertical scaling and standard setting with recommendations for FCAT 2.0.* Retrieved from the Florida Department of Education website: http://www.fldoe.org/asp/k12memo/pdf/StudyBestPracticesVerticalScalingStandardSetting.pdf

Shavelson, R. J., & Webb, N. M. (2006). Generalizability theory. In Green, J. L., Camilli, G. & Elmore, P. B. (Eds.), *Complementary methods for research in education* (3rd ed., pp. 309-322). Washington, DC.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210.

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer (Ed.) *Computerized adaptive testing: A primer* (pp. 161–186). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D. (2003). MULTILOG (Version 7.03) [Computer software]. Chicago: Scientific Software International.

Thissen, D., Chen, W-H, & Bock, R.D. (2003). MULTILOG (version 7) [Computer software]. In Mathilda du Toit (Eds.). *IRT from SSI: BILOG-MG MULTILOG PARSCALE TESTFACT.* Chicago: Scientific Software International.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tong, Y., Um, K., Turhan, A., Parker, B., Shin, D., Chien, M., & Hsu, Y. (2007). *Evaluation of IRT score estimation*. Pearson.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–10.

Turhan, A., Courville, T., & Keng, L. (2009). *The effects of anchor item position on a vertical scale design.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

van der Linden, W.J., & Hambleton, R.K. (Eds.) (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

Viera A.J. & Garrett J.M.(2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, 37(5), 360-363.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics, 14*(4), 1261–1295.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*(2), 245–262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125–145.

Young, M. J., & Yoon, B. (1998, April). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment.* (CSE Technical Report 475). Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles, Calif.: University of California, Los Angeles.