FCAT

Florida Comprehensive Assessment Test®

# WRITING+

# Technical Report
# For 2006 FCAT Test Administration

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

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

Harcourt

**San Antonio, TX**

**January 2007**

# Table of Contents

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

# List of Tables

# List of Figures

# INTRODUCTION AND OVERVIEW

This report presents summary data of technical information on the measurement characteristics of the three grade-level Writing+ assessments (Grades 4, 8, and 10) included in the Florida Comprehensive Assessment Test® (FCAT) for Spring 2006. These characteristics provide an indication of the current quality of the Writing+ assessments.

## Description of FCAT Writing+

FCAT Writing+ is based on the belief that statewide standardized testing of student writing can benefit all levels of the educational enterprise if designed carefully to achieve specific educational purposes. The design of the program is aimed at producing unique and effective ways to measure writing as a developmental and recursive process. One goal is to measure achievement of the Sunshine State Standards (SSS) and *Blueprint 2000.* Another goal is to achieve the highest possible efficiency in assessment by developing stimuli that produce maximum information in a minimum amount of testing time.

Careful design means assessing writing abilities in the context of material that is developmentally and instructionally appropriate for students at the specified grade, with authentic tasks that reflect the best practices in writing and language instruction, and with different opportunities for the student to demonstrate proficiency on one test. It also means tasks should have a clear purpose for students and should be legitimate, as agreed upon by the educational community.  The tasks should also represent skills deemed valuable by both educators and the community at large, so the results of these tasks can later be used to improve instruction.

One of the best ways to test student writing is to have students write. FCAT Writing+ includes a performance task portion with a 45-minute demand writing assessment. Florida educators set the scoring standards for the performance task, and specially trained and qualified professional scorers use a state-determined rubric to implement those standards. In addition to the performance task, multiple-choice (MC) items are included because there is more to assessing a student's knowledge of writing than essay composition alone.  Both item types are used to test a student's knowledge of critical writing elements.

In both cases (i.e., with performance tasks and MC items), the involvement of state educators with demonstrated expertise and experience in teaching writing is essential to ensure that the tasks facing students are both legitimate and valuable. The educators and community representatives who form the review committees must determine that (a) the test reflects the best practices in instruction, (b) no task or language used in the test disproportionately affects one ethnic or gender group, (c) the tasks presented are worthy of being tested, and (d) the test adheres to the standards set out in the state's documents: the Sunshine State Standards and *Blueprint 2000).*

### *Multiple-Choice Items*

The approach used to assess writing skills on the MC portion of FCAT Writing+ takes writing objectives out of isolation and puts them into a real-life context. It utilizes various writing stimuli, including simulated student writing samples, to represent specific writing tasks a writer must complete for various audiences, occasions, and purposes. The writing stimuli allow for the assessment of the pre-writing, drafting, revising, and editing stages of writing. Each writing stimulus is followed by questions that reflect SSS benchmarks that are classified into four content categories: *focus*, *organization*, *support*, and *conventions*. Items that relate to benchmarks associated with pre-writing skills are classified in the *focus* and *organization* categories. Items related to drafting, revising, and editing benchmarks are classified in one of the four reporting categories. These categories are defined as follows:

- *Focus* - includes planning for writing by grouping related ideas and identifying the purpose for writing and refers to how clearly a central idea (topic), theme, or unifying point is presented and maintained.
- *Organization* - refers to the structure or plan of development and the relationship of one point to another to provide a logical progression of ideas. It also refers to the use of transitional devices to signal both the relationship of the supporting ideas to the central idea, theme, or unifying point and the connections between and among sentences.
- *Support* - refers to the quality of details used to explain, clarify, or define. The quality of the support depends on word choice, specificity, depth, relevance, and thoroughness. Support may be developed through the use of additional details, anecdotes, illustrations, and examples that further clarify meaning.
- *Conventions* - refer to punctuation, capitalization, spelling, usage, and sentence structure.

All of the items on the Writing+ MC portion are associated with a specific writing stimulus. The following writing types are used:

- *A writing plan* - is an outline or graphic organizer that is associated with three items. It is representative of student pre-writing activities and refers to the *focus* or *organization* reporting categories.
- *A writing sample* - is an example of student draft writing with embedded errors that are addressed by items written to measure understanding of the issues associated with the errors. Writing samples are generally 200 to 350 words in length, and they tend to be shorter at the lower grades and longer at the upper grades. On the test form, 5–7 items are associated with each writing sample.
- *A cloze* - is a text with numbered blanks inserted where words have been omitted in order to have students select the correct usage or spelling of each omitted word. Cloze texts are used to test conventions. These texts are not necessarily representative of student writing. Rather, they are texts that are highly interesting to students at the tested grade level. Graphics are added to these texts to enhance student interest and/or understanding. Cloze texts are somewhat shorter than the writing samples at each tested grade level. Typically, cloze-based samples will be 75 to 150 words long and have 3–4 items associated with each sample.
- *A stand-alone* - is an item that serves as a stimulus to measure the *conventions* reporting category. Stand-alone items measure capitalization, punctuation, and sentence

structure.

All items in the MC portion have either three or four options. The items utilizing a 3-option format are cloze-based or stand-alone and address only the *conventions* reporting category, whereas the items utilizing a 4-option format are either sample-based (measuring *focus*, *organization*, and *support*) or plan-based (measuring only *focus* and *organization*).

Both 3- and 4-option item formats appear in groups that are mixed within a test form. To avoid potential confusion for the test taker, each of the two formats appears in blocks of at least 3 items. The answer document displays as many bubbles for an item as there are options, and item formats appear in the same positions across test forms within a grade level. These requirements unify the response format and allow a single answer document to be used for all forms within a grade.

At each grade level, the MC portion of Writing+ includes 44 items that are scored (core items) and 10 additional items that are not scored. There are 23 different forms for the MC portion for each grade level. The core items are located in the same positions across forms and are used to compute student scores. However, the forms do vary with respect to the non-scored items. In 2006, all of the non-scored items were field-test items being tested for potential future use as core items.

## *Performance Task*

In addition to the MC items, students also demonstrate their writing ability. A *writing prompt* serves as the stimulus for the FCAT Writing+ performance task. Each student has 45 minutes to respond to one writing prompt. The prompt presents a topic in a format that serves to encourage, stimulate, and evoke a written response. The format of the prompt is designed to appeal to the greatest number of students possible. The prompt identifies the intended mode or purpose for writing.

Prompts elicit writing for specific purposes, or modes: *narrative* or *expository* writing is used at Grade 4, and *expository* or *persuasive* writing is used at Grades 8 and 10. Within each of the tested grades, half of the students respond to one writing mode, and the other half respond to the other mode. Each prompt serves as a writing stimulus by suggesting that students think about some aspect of the topic’s central theme. The subject matter must be grade-level appropriate, and the wording of each prompt is checked for clarity and readability.

The prompt provides the student with the subject (topic) and purpose for writing. Prompts have two basic components: the writing situation and the directions for writing. The writing situation orients students to the subject, and the directions for writing set the parameters, such as identifying the audience to whom the writing should be directed. In Grades 8 and 10 the prompt components include the headings “Writing Situation” and “Directions for Writing.” In Grade 4 these headings are omitted.

Writing prompt responses are scored using a holistic method. Trained scorers evaluate the overall quality of students’ writing by using a six-point rubric. This rubric allows the readers to consider the integration of the four writing elements. Two readers score each student response.

### *Writing+ Scores*

Writing+ performance is reported on an overall scale described more completely in the item response theory (IRT) scaling section later in this report. In addition, raw scores are reported from the MC component for *focus*, *organization*, *support*, and *conventions*. Finally, writing prompt raw scores are reported as the average of the two hand-scored ratings. The number of items and possible scores are shown in Table 1.

**Table 1.** Number and Possible Scores by Reporting Category for Items in Writing+

| Writing Element | Number of Items[a] | Possible Score |
|---|---|---|
| MC- Focus | 10–13 | 10–13 |
| MC-Organization | 8–10 | 8–10 |
| MC-Support | 8–10 | 8–10 |
| MC-Conventions | 15–16 | 15-16 |
| MC-Total | 44 | 44 |
| Writing Prompt | 1[b] | 1–6[c] |
| Overall Scale Score | 44+1 | 100–500 |

[a] Varies by grade.

[b] Two different prompts are developed for each grade, but each student responds to only one prompt.

[c] The final score for the prompt is the average of two raters' scores, so 0.5 values are possible.

## Report Content

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

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

FCAT is a continuous assessment system. While the essential structure and focus of the FCAT tests remain fairly fixed over time and student achievement results maintain a level of comparability across testing years, specific questions on a test administered in any given year may vary. In addition to the variability of test questions administered on the "core" portion of the test (i.e., the portion of the test that actually contributes to students' reported scores), students will also answer some items on the test that do not count toward their ultimate scores. In 2006, all of the non-scored items were field-tested for potential future use. In future years, "anchor" items will also be included as non-scored items in some forms. Anchors are items repeated from prior years to establish comparability of scores across years.

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

## ITEM PREPARATION AND TEST ASSEMBLY

Item preparation and review procedures are described in the *FCAT Writing+ Test Item Specifications* documents, which are actually three separate documents for the three Writing+ grades. These documents are available at http://fcat.fldoe.org/fcatis01.asp.

The most fundamental requirement of any test is that it is valid for its intended use. For an achievement test, the focus in on the test's *content validity*. In other words, do test scores indicate the extent to which students have mastered the knowledge and skills as described by a well-defined content domain? Such validity is built into Writing+ by having each and every test item explicitly tied to Florida's Sunshine State Standards (SSS).

The SSS are a hierarchical catalog of the knowledge and skills students should acquire in all grades. Strands in the SSS, including Writing, are first divided into *standards*. Standards are then divided into *benchmarks*, and benchmarks are further delineated by one or more *benchmark clarifications*. Writing is one of the strands (main divisions) within Language Arts. The *Test Item Specifications* define writing knowledge and skills in terms of the critical writing elements of *focus*, *organization*, *support*, and *conventions* with guidelines for test item content. Within any grade for Writing, the *Test Item Specifications* include four levels with detailed explanations about the knowledge and skills required for content mastery. Furthermore, the *Test Item Specifications* include style and universal design guidelines[1] to ensure that items are well-crafted and focus exclusively on the intended content. Sample items are also included to show how and why various item types are developed. All of this information guides item development to ensure content validity for the assessment.

With the SSS and *Test Item Specifications* as guides, items must go through a three-phase development process. In the first phase, education professionals draft items that are then

[1] Universal design guidelines are intended to make items as fully accessible as possible to all students, including English Language Learners and students with disabilities.

subjected to a critical content and editorial review. These items are then forwarded to content staff at the Test Development Center (TDC) in Florida, where they receive an additional review. Typically, any item submitted may have 1 of 3 fates: (a) it is accepted with no (or minor) edits, (b) it is rejected as inappropriate for the FCAT, or (c) it is returned to the contractor with comments requesting revisions in style or focus, so the item can be returned to the review process. Ongoing dialogue between the contractor and TDC staff on the "accept with revisions" items assures that both the contractor and the TDC staff reach agreement on all items deemed appropriate for use on the Writing+ assessment.

In the second phase of item development, FCAT items go through a rigorous review process before they can be field tested. Item reviews are conducted by the following groups: (a) FDOE for content, sensitivity and bias, match to benchmark, and FCAT style; (b) community sensitivity committees; (c) bias committees, with representatives from a variety of cultural backgrounds; and, (d) content committees. The procedures used for item review for the FCAT 2006 field-test items are described in *Analysis of the FCAT Test Item Review Conducted by the Florida Department of Education and Harcourt Educational Measurement* (FDOE, May 2001).

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

Both Harcourt and TDC staff use the field-tested and previously-used (core) items contained in the FCAT item bank to build test forms through a multi-step process (FDOE, 2004). Typically, Harcourt content and psychometric staff propose draft test forms for each grade and subject for TDC to review prior to test construction. These draft forms are assembled according to (1) content coverage guidelines documented in test blueprints that define how the SSS benchmarks will be included in the test, and (2) statistical guidelines documenting how well the proposed tests (i.e., whole tests, as well as reportable strands/clusters) match the characteristics of previously administered versions of the FCAT.

# CONSTRUCTED-RESPONSE SCORING PROCEDURES

## Scorer Training

For writing prompts, students provide handwritten responses which are then scored by individual human scorers. Because essay scoring requires considerable skill and experience, special attention is paid to training scorers. This is accomplished through the use of FDOE-approved training materials that are developed during the "Rangefinder Review" sessions held with state educators and FDOE staff.

Potential scorers are given an overview of the project along with FDOE expectations and guidelines. To ground them in the rules of scoring, potential scorers are shown several sets of training papers. They are then given "qualification sets" to ensure that a minimum agreement percentage is met. Only after the successful completion of the qualifying process are scorers allowed to rate actual student responses. To ensure consistency between training sessions, papers are presented in the same order with the same comments for each group of scorers trained. This is done so that each group of scorers will complete training with the same rules and information.

## Maintaining Consistent, Reliable Scoring

Every student essay is scored by two independent scorers using a rubric that spans the range between 1 and 6 points. Maximal difference between two scores given to the same essay is one point. If the difference is greater than one point, the essay is read by a third scorer for a resolution. In cases where the third scorer gives a score that substantially differs from the previous two scores, the essay is read by a fourth and final scorer, who is usually a scoring room director.

Other methods implemented to maintain reliable scoring are those used to control scorer drift. One daily process is to have team members review their rangefinder and horizontal training sets for 15 minutes or longer, if needed. This helps to keep all scorers and team leaders grounded in the rules and guidelines laid out in training. All of the validity and reliability reports, along with calibration sets, also help prevent scorer drift.

As a monitoring tool, a validity report shows how frequently a scorer agrees with the “true score” given to pre-selected and expert-scored validity responses. By accessing validity reports, the scoring director can see which validity papers are being missed, which scorers are missing validity papers, and which scorers are scoring the papers too high or too low.

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

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

# 2006 FCAT STATISTICS

This section of the report presents psychometric analyses of the 2006 FCAT core assessment for Writing+. Traditional item analyses and IRT analyses for the initial reporting period were conducted using a special calibration sample of students because of the requirements for rapid turnaround in score reporting. A set of schools was chosen specifically for this purpose, and those schools returned their students’ responses on an early timeline. The general strategy was to select schools that would provide a sample of students who were representative of the State’s regions, ethnic diversity, and achievement scores in past years. Only standard curriculum students were used in the analyses; exceptional students and students in the limited English proficiency (LEP) program for two years or less were excluded. In addition, students in the calibration sample had to meet criteria indicating they had attempted the test.[2] More details about

[2] Test scores are only computed for students who meet the “attemptedness” criteria. The criteria specify that a

the selection of this sample appear in *Plan for Selecting the Calibration Sample for the 2006 FCAT Administration* (FDOE, November, 2005).

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

## Calibration Sample

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

In order to accomplish this goal in a timely fashion, enrollment and scoring information from the previous administration are analyzed. The analysis establishes a target range of characteristics the schools selected need to meet in order to provide a good model that reflects the attributes of Florida's geographic regions.

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

### *Characteristics*

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

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

a. The sample should maintain the same geographic region distribution, plus or minus 200 students.
b. The number of schools selected should maintain the same geographic region distribution, plus or minus three schools.
c. The sample must include at each grade level a school from each of the largest six divisions in the state.
d. The percentage of the four major ethnic groups (White, African American, Hispanic, and Other, which includes Asian, American Indian and Multiracial students) should maintain the same ratio as the state and within each geographic region (northern, central, and

student must have at least 6 non-blank answers in each of 2 sessions for the multiple-choice section of the assessment and must have attempted a response to the writing prompt.

southern), plus or minus 5 percent.

e. The standard deviation unit (computed by dividing the absolute value of the difference between the sample mean and the state mean by the standard deviation of the state) must be 0.2 or less.

f. The standard deviation ratio (computed by dividing the standard deviation of the sample by the standard deviation of the state) must be between 0.9 and 1.1.

Because 2006 was the first year of Writing + live test administration, the final number of students in the calibration sample was twice as much as Reading and Mathematics.

### *Evaluation of Representativeness*

Tables 2 through 10 on the following pages compare each grade/subject calibration sample with other statewide sets of students. One set of comparison students, labeled “Total,” includes all students with FCAT records for 2006.[3] Some of these students, however, did not receive FCAT scores because they failed the attemptedness criteria. A second set of students includes all standard curriculum students, again including those that did not receive test scores because of failing the attemptedness criteria. These two sets of students provide a basis for comparing the gender and ethnicity distributions of the calibration samples. Note that the number of students across the respective categories does not sum to the total listed because of missing ethnicity and gender information (i.e., some students did not provide this information).

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

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

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

**Table 2.** Grade 4 Writing+ Frequency Distributions for Different Student Groups by Ethnicity

| | Asian | African American | Hispanic | American Indian | Multi-racial | White | Total[a] |
|---|---|---|---|---|---|---|---|
| **Calibration Sample** | 377<br>(1.95%) | 4,948<br>(25.56%) | 4,338<br>(22.40%) | 66<br>(0.34%) | 644<br>(3.33% | 8,907<br>(46.00%) | 19,362 |
| **Standard Curriculum Students** | 3,859<br>(2.41%) | 34,981<br>(21.87%) | 36,156<br>(22.61%) | 495<br>(0.31%) | 6,005<br>(3.75%) | 78,381<br>(49.01%) | 159,922 |
| **All Students** | 4,338<br>(2.29%) | 42,036<br>(22.19%) | 45,937<br>(24.25%) | 568<br>(0.30%) | 6,829<br>(3.60%) | 89,706<br>(47.35%) | 189,463 |

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

**Table 3** Grade 4 Writing+ Frequency Distributions for Different Student Groups by Gender

| | Female | Male | Total[a] |
|---|---|---|---|
| **Calibration Sample** | 10,130<br>(52.32%) | 9,146<br>(47.24%) | 19,362 |
| **Standard Curriculum Students** | 82,856<br>(51.81%) | 77,026<br>(48.16%) | 159,922 |
| **All Students** | 93,580<br>(49.39%) | 95,838<br>(50.58%) | 189,463 |

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

**Table 4.** Grade 4 Writing+ Mean Scale Scores for Different Student Groups

| | Calibration Sample | | | Standard Curriculum Students | | | All Students | | |
|---|---|---|---|---|---|---|---|---|---|
| | **M** | **SD** | **N** | **M** | **SD** | **N** | **M** | **SD** | **N** |
| **All** | 302.64 | 67.11 | 19,362 | 307.39 | 64.64 | 159,922 | 296.12 | 70.67 | 189,463 |
| **Female** | 313.51 | 65.78 | 10,130 | 317.43 | 63.62 | 82,856 | 309.20 | 68.52 | 93,580 |
| **Male** | 290.84 | 66.42 | 9,146 | 296.61 | 63.99 | 77,026 | 283.36 | 70.40 | 95,838 |
| **African American** | 282.77 | 61.99 | 4,948 | 285.31 | 60.47 | 34,981 | 274.66 | 66.05 | 42,036 |
| **Hispanic** | 306.69 | 65.19 | 4,338 | 306.76 | 62.07 | 36,156 | 290.98 | 71.29 | 45,937 |
| **White** | 310.35 | 67.96 | 8,907 | 315.94 | 64.91 | 78,381 | 306.83 | 69.75 | 89,706 |

M = Mean
SD = Standard Deviation
N = Sample Size

**Table 5.** Grade 8 Writing+ Frequency Distributions for Different Student Groups by Ethnicity

| | Asian | African American | Hispanic | American Indian | Multi-racial | White | Total[a] |
|---|---|---|---|---|---|---|---|
| **Calibration Sample** | 422<br>(1.90%) | 5,948<br>(26.75%) | 3,999<br>(17.98%) | 57<br>(0.26%) | 479<br>(2.15%) | 11,228<br>(50.49%) | 22,238 |
| **Standard Curriculum students** | 3,969<br>(2.35%) | 38,069<br>(22.57%) | 37,010<br>(21.94%) | 514<br>(0.30%) | 4,462<br>(2.65%) | 84,553<br>(50.13%) | 168,680 |
| **All Students** | 4,393<br>(2.22%) | 45,709<br>(23.06%) | 45,529<br>(22.97%) | 585<br>(0.30%) | 5,006<br>(2.53%) | 96,918<br>(48.89%) | 198,247 |

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

**Table 6.** Grade 8 Writing+ Frequency Distributions for Different Student Groups by Gender

| | Female | Male | Total[a] |
|---|---|---|---|
| **Calibration Sample** | 11,534<br>(51.87%) | 10,587<br>(47.61%) | 22,238 |
| **Standard Curriculum Students** | 87,270<br>(51.74%) | 81,310<br>(48.20%) | 168,680 |
| **All Students** | 97,686<br>(49.27%) | 100,456<br>(50.67%) | 198,247 |

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

**Table 7.** Grade 8 Writing+ Mean Scale Scores for Different Student Groups

| | Calibration Sample | | | Standard Curriculum Students | | | All Students | | |
|---|---|---|---|---|---|---|---|---|---|
| | **M** | **SD** | **N** | **M** | **SD** | **N** | **M** | **SD** | **N** |
| **All** | 300.97 | 58.17 | 22,238 | 306.21 | 55.22 | 168,680 | 294.61 | 62.68 | 198,247 |
| **Female** | 309.98 | 56.87 | 11,534 | 315.18 | 53.80 | 87,270 | 306.94 | 59.67 | 97,686 |
| **Male** | 291.67 | 57.77 | 10,587 | 296.65 | 55.06 | 81,310 | 282.66 | 63.20 | 100,456 |
| **African American** | 276.84 | 52.31 | 5,948 | 284.04 | 51.10 | 38,069 | 271.75 | 58.77 | 45,709 |
| **Hispanic** | 291.21 | 56.69 | 3,999 | 296.86 | 54.73 | 37,010 | 282.29 | 63.72 | 45,529 |
| **White** | 315.85 | 55.78 | 11,228 | 318.65 | 53.12 | 84,553 | 309.20 | 59.51 | 96,918 |

**Table 8.** Grade 10 Writing+ Frequency Distributions for Different Student Groups by Ethnicity

| | Asian | African American | Hispanic | American Indian | Multi-racial | White | Total[a] |
|---|---|---|---|---|---|---|---|
| **Calibration Sample** | 394<br>(1.84%) | 5,395<br>(25.22%) | 3,620<br>(16.92%) | 55<br>(0.26%) | 358<br>(1.67%) | 11,442<br>(53.48%) | 21,395 |
| **Standard Curriculum Students** | 4,237<br>(2.58%) | 35,468<br>(21.64%) | 36,131<br>(22.04%) | 456<br>(0.28%) | 2,846<br>(1.74%) | 84,701<br>(51.67%) | 163,930 |
| **All Students** | 4,608<br>(2.48%) | 41,513<br>(22.09%) | 42,881<br>(22.82%) | 521<br>(0.28%) | 3,151<br>(1.68%) | 95,172<br>(50.64%) | 187,939 |

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

**Table 9.** Grade 10 Writing+ Frequency Distributions for Different Student Groups by Gender

| | Female | Male | Total[a] |
|---|---|---|---|
| **Calibration Sample** | 11,126<br>(52.00%) | 10,046<br>(46.95%) | 21,395 |
| **Standard Curriculum Students** | 86,246<br>(52.61%) | 77,594<br>(47.33%) | 163,930 |
| **All Students** | 95,145<br>(50.93%) | 92,700<br>(49.32%) | 187,939 |

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

**Table 10.** Grade 10 Writing+ Mean Scale Scores for Different Student Groups

| | Calibration Sample | | | Standard Curriculum Students | | | All Students | | |
|---|---|---|---|---|---|---|---|---|---|
| | **M** | **SD** | **N** | **M** | **SD** | **N** | **M** | **SD** | **N** |
| **All** | 302.25 | 59.32 | 21,395 | 305.54 | 57.64 | 163,930 | 295.47 | 63.90 | 187,939 |
| **Female** | 310.04 | 58.59 | 11,26 | 313.09 | 57.20 | 86,246 | 305.48 | 62.36 | 95,145 |
| **Male** | 294.44 | 58.88 | 10,046 | 297.19 | 56.95 | 77,594 | 285.23 | 63.83 | 92,700 |
| **African American** | 278.73 | 55.33 | 5,395 | 278.01 | 52.98 | 35,468 | 266.88 | 59.82 | 41,513 |
| **Hispanic** | 289.29 | 60.20 | 3,620 | 292.74 | 57.00 | 36,131 | 280.93 | 63.69 | 42,881 |
| **White** | 316.83 | 56.83 | 11,442 | 321.19 | 53.82 | 84,701 | 312.94 | 59.39 | 95,172 |

# FCAT 2006 Item Analysis

This section contains classical item analysis statistics for difficulty and item-total correlations. For each of the items on the three Writing+ tests (Grades 4, 8, and 10), item difficulties (*p*-values), item-total correlations, and correlations between the item and reporting categories within each of the subject areas were computed.

Tables 11–13 summarize the item analysis results by presenting the minimum, 25th percentile, 50th percentile, 75th percentile, and maximum values for each grade's test (across all core items).

## *Item Difficulty Summary*

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

Table 11 illustrates the distribution of *p*-values for all Writing+ items. For a test to be effective, *p*-values should show that the items vary in difficulty, but they should not be too high (e.g., above 0.90) or too low [e.g., 0.20 (near chance for MC items) or less than 0.10 for the other item types]. Table 11 shows that there were some high *p*-values monitored during IRT processing, but generally, the item *p*-values are dispersed across a sufficient range to establish satisfactory measurement reliability for a wide range of achievement.

**Table 11.** Proportional[a] *p*-value Summary Data for All Writing+ Items

| Grade | Reporting Category | No. of Items | Minimum | 25th Percentile | 50th Percentile | 75th Percentile | Maximum |
|---|---|---|---|---|---|---|---|
| 4 | Total | 46[b] | 0.403 | 0.592 | 0.681 | 0.769 | 0.889 |
| | Prompt | 2[b] | 0.613 | 0.613 | 0.657 | 0.701 | 0.701 |
| | Focus | 10 | 0.434 | 0.559 | 0.699 | 0.768 | 0.857 |
| | Organizations | 10 | 0.504 | 0.582 | 0.681 | 0.752 | 0.826 |
| | Support | 8 | 0.514 | 0.630 | 0.698 | 0.877 | 0.889 |
| | Conventions | 16 | 0.403 | 0.576 | 0.628 | 0.792 | 0.854 |
| 8 | Total | 46[b] | 0.452 | 0.572 | 0.661 | 0.785 | 0.942 |
| | Prompt | 2[b] | 0.664 | 0.664 | 0.671 | 0.679 | 0.679 |
| | Focus | 13 | 0.488 | 0.593 | 0.638 | 0.807 | 0.900 |
| | Organizations | 6 | 0.718 | 0.785 | 0.819 | 0.841 | 0.858 |
| | Support | 10 | 0.487 | 0.544 | 0.606 | 0.679 | 0.869 |
| | Conventions | 15 | 0.452 | 0.495 | 0.612 | 0.761 | 0.942 |
| 10 | Total | 46[b] | 0.319 | 0.563 | 0.669 | 0.751 | 0.911 |
| | Prompt | 2[b] | 0.666 | 0.666 | 0.670 | 0.673 | 0.673 |
| | Focus | 11 | 0.502 | 0.618 | 0.729 | 0.760 | 0.884 |
| | Organizations | 8 | 0.596 | 0.617 | 0.707 | 0.821 | 0.911 |
| | Support | 9 | 0.472 | 0.597 | 0.643 | 0.683 | 0.896 |
| | Conventions | 16 | 0.319 | 0.417 | 0.609 | 0.709 | 0.889 |

[a] Mean score divided by total possible score.

[b] Note that although two writing prompts exist, each student receives only one.

### *Pearson Item-Total Correlations*

Table 12 shows the distribution of item-total raw score correlations and correlations between items and reporting category scores. These are computed as Pearson correlations.[4] The total score is the sum of all item points. The reporting category score is the sum of points from items in that category. Distributions for the item-reporting category include only correlations of items from that category.

The most important criterion for the correlation statistics is that they are not negative nor near zero. Items with negative correlations should not be used in IRT processing. As seen in Table 12, negative correlations were not observed.

**Table 12.** Item-Total Correlation Summary by Cluster: Writing+ Core Items

| Grade | Reporting Category | No. of Items | Minimum | 25th Percentile | 50th Percentile | 75th Percentile | Maximum |
|---|---|---|---|---|---|---|---|
| 4 | Total | 46[a] | 0.155 | 0.302 | 0.369 | 0.423 | 0.565 |
| | Prompt | 2[a] | 0.558 | 0.558 | 0.562 | 0.565 | 0.565 |
| | Focus | 10 | 0.198 | 0.368 | 0.386 | 0.412 | 0.435 |
| | Organizations | 10 | 0.272 | 0.382 | 0.430 | 0.456 | 0.492 |
| | Support | 8 | 0.239 | 0.362 | 0.371 | 0.403 | 0.427 |
| | Conventions | 16 | 0.155 | 0.270 | 0.305 | 0.341 | 0.394 |
| 8 | Total | 46[a] | 0.198 | 0.317 | 0.377 | 0.414 | 0.668 |
| | Prompt | 2[a] | 0.651 | 0.651 | 0.660 | 0.668 | 0.668 |
| | Focus | 13 | 0.227 | 0.316 | 0.373 | 0.390 | 0.489 |
| | Organizations | 6 | 0.352 | 0.356 | 0.399 | 0.405 | 0.437 |
| | Support | 10 | 0.287 | 0.356 | 0.416 | 0.464 | 0.501 |
| | Conventions | 15 | 0.198 | 0.250 | 0.329 | 0.377 | 0.414 |
| 10 | Total | 46[a] | 0.141 | 0.297 | 0.370 | 0.441 | 0.717 |
| | Prompt | 2[a] | 0.679 | 0.679 | 0.698 | 0.717 | 0.717 |
| | Focus | 11 | 0.274 | 0.307 | 0.407 | 0.467 | 0.526 |
| | Organizations | 8 | 0.256 | 0.344 | 0.386 | 0.460 | 0.496 |
| | Support | 9 | 0.317 | 0.346 | 0.391 | 0.400 | 0.501 |
| | Conventions | 16 | 0.141 | 0.232 | 0.296 | 0.374 | 0.469 |

[a]Note that although two types of writing prompts exist per grade, each student receives only one prompt.

### *Biserial Item-Total Correlations*

The point-biserial correlations produced for dichotomous items are restricted in possible range to the extent that the items are either very easy or very difficult. The biserial correlation may be understood as an estimate of the correlation that would have been obtained if the dichotomous item had actually been a normally distributed continuous measure. It will always be larger than

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

the corresponding point biserial. In fact, if the total score on the test is not normally distributed, then the biserial correlation can nonsensically exceed 1 (Cohen & Cohen, 1975). The performance task items are not included in the calculation of the biserial correlation, which are found in Table 13.

**Table 13.** Biserial Correlation Summary by Cluster: Writing+ Core MC Items

| Grade | Reporting Category | No. of Items | Minimum | 25th Percentile | 50th Percentile | 75th Percentile | Maximum |
|---|---|---|---|---|---|---|---|
| 4 | Total | 44 | 0.194 | 0.400 | 0.489 | 0.553 | 0.667 |
| | Focus | 10 | 0.303 | 0.504 | 0.513 | 0.519 | 0.586 |
| | Organizations | 10 | 0.341 | 0.485 | 0.574 | 0.628 | 0.667 |
| | Support | 8 | 0.299 | 0.478 | 0.548 | 0.595 | 0.635 |
| | Conventions | 16 | 0.194 | 0.373 | 0.410 | 0.462 | 0.499 |
| 8 | Total | 44 | 0.270 | 0.409 | 0.504 | 0.553 | 0.653 |
| | Focus | 13 | 0.326 | 0.396 | 0.498 | 0.558 | 0.617 |
| | Organizations | 6 | 0.512 | 0.530 | 0.568 | 0.615 | 0.619 |
| | Support | 10 | 0.362 | 0.482 | 0.544 | 0.582 | 0.653 |
| | Conventions | 15 | 0.270 | 0.339 | 0.451 | 0.504 | 0.543 |
| 10 | Total | 44 | 0.184 | 0.410 | 0.490 | 0.595 | 0.681 |
| | Focus | 11 | 0.373 | 0.439 | 0.510 | 0.608 | 0.679 |
| | Organizations | 8 | 0.395 | 0.452 | 0.569 | 0.615 | 0.664 |
| | Support | 9 | 0.398 | 0.492 | 0.499 | 0.570 | 0.676 |
| | Conventions | 16 | 0.184 | 0.292 | 0.413 | 0.502 | 0.681 |

# Item Response Theory (IRT) Scaling

## *IRT Framework*

FCAT scoring is built on item response theory (IRT). In essence, IRT assumes that test item responses by students are the result of underlying levels of achievement possessed by those students. IRT algorithms search for “item parameters” which capture a nonlinear relationship between achievement and the likelihood of correctly answering each item. Items that fit the IRT model will exhibit a pattern of lower probabilities of correct responses from low-ability students to higher probabilities of correct responses from high-ability students. This is reflected in an “item characteristic curve (ICC), as depicted in Figure 1 for a MC item. Items vary in difficulty such that the position of the point of inflection is higher or lower (i.e., to the right or to the left) along the achievement scale. For example, the point of inflection of the curve for the sample item in Figure 1 is centered at zero, the mean on the achievement index. An efficient test is composed of items with test characteristics similar to that depicted in Figure 1 but with varying difficulties (B parameter) that discriminate achievement along the entire achievement scale, which is typically called “theta.” ICCs also differ in their lower asymptotes, which relate to how easy it is to get the item correct by guessing (C parameter) and the gradient of their slopes at the inflection point (A parameter).

While IRT modeling of performance tasks is conceptually similar, performance tasks require a more complex mathematical treatment. In the end, however, IRT modeling of a performance task captures the expected number of points that students should achieve on the performance task, depending on their achievement level. The result is a curve similar to Figure 1 where the *y*-axis represents the probability of correct response.

The three-parameter logistic (3PL) model (Lord & Novick, 1968) is used to process MC items, and the two-parameter partial credit (2PPC) model (Muraki, 1992) is used to process PT items. Figure 1 depicts an ICC using the 3PL model. For the PT items, student scores could fall into any of several different score categories (1, 2, 3, 4, 5, or 6 for writing prompt items). The 2PPC model captures probabilities for students receiving any of the possible points, depending on differences in their achievement. Figure 2 depicts the probabilities of a correct answer for a writing prompt item. *FCAT 2005 Test Construction Specifications* (FDOE, 2004) presents the technical details of these models more fully. The statistical package MULTILOG (Thissen, 1991) is used for the IRT analyses.

The MC items and ratings for the writing prompts were scaled together in one MULTILOG run per grade. Because of MULTILOG limitations on the number of categories allowed for 2PPC items, each final rating (two per student) was entered separately into the model. The separate ratings were randomly assigned as rating 1 and rating 2, yielding two sets of item parameter estimates. Because the resulting two sets of parameter estimates are replications based on only one response per student, they were averaged for scoring purposes (i.e., geometric mean for the discrimination parameter and arithmetic mean for the category difficulty values).

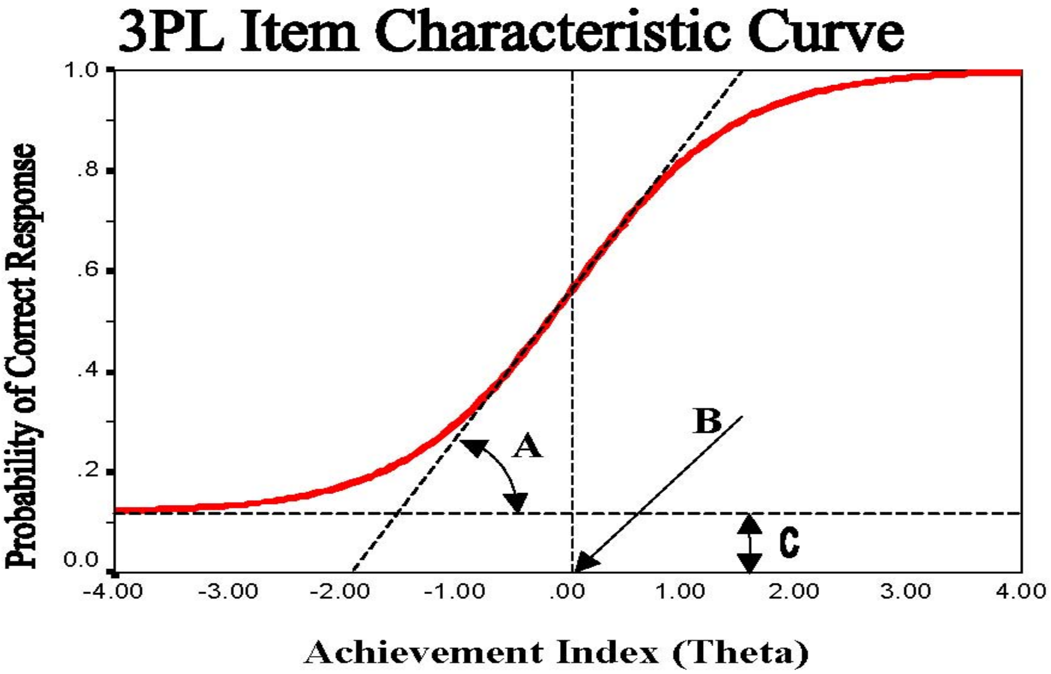

**Figure 1.** Item Characteristic Curve based on the three-parameter logistic trace line.

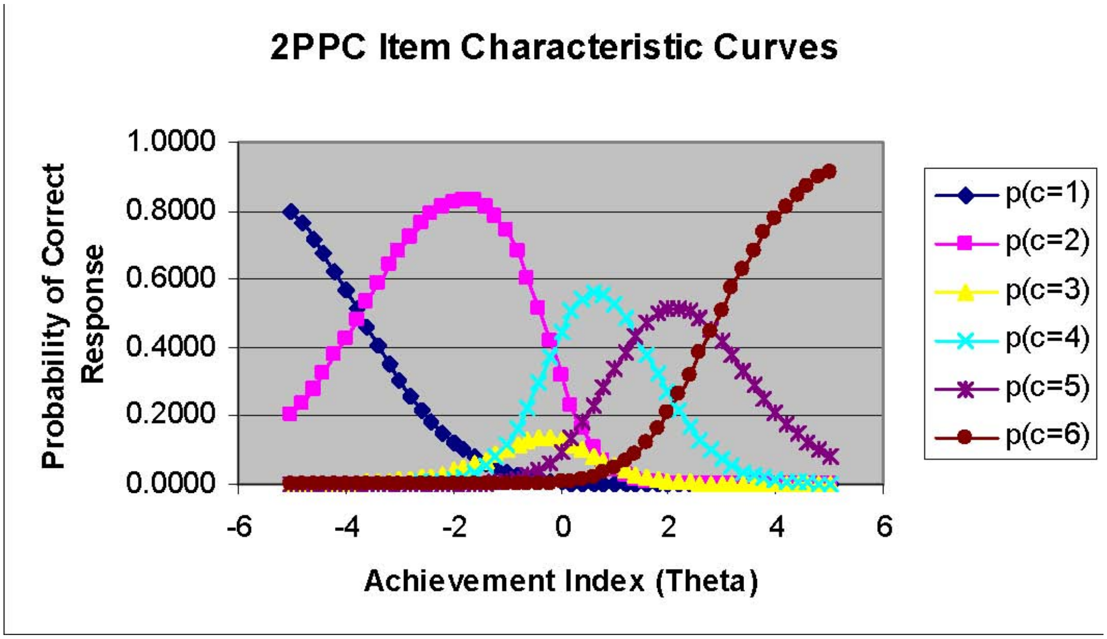

**Figure 2.** Probability of receiving a correct answer for a writing prompt item.

IRT item parameters provide the means for assigning achievement scores to individual students. Because the item parameters represent response probabilities, each student's achievement score is assigned as the level of achievement most likely to have created that student's observed responses.[5] Multiple-choice items and prompt ratings were analyzed separately, yielding an estimate of student writing achievement from MC items and an estimate of student writing achievement from the writing performance task. These two estimates were averaged to provide the final scale scores for Writing+. One issue arising from this process was how IRT estimates ability when the writing prompt on one test form appears to be more difficult than the prompt on another test form in the same grade.

In each grade, the mean scores for the Writing+ prompt pairs may not be equal, giving rise to a concern about the fairness of the test. Since no one student answers both prompts, there is no way to determine whether the difference in means for the prompts is a result of differences in students' abilities for the two groups, differences in the difficulties of the prompts, or some combination of the two. Nevertheless, all students in both groups took the same multiple-choice Writing+ portion. IRT uses student responses on the MC items relative to each prompt to make a determination about differences in prompt difficulty. Therefore, if students have equal MC ability estimates but perform differently on the two prompts, then the two prompts are not of equal difficulty and are assigned item parameters that capture that difference. The students are scored using the item parameters for the prompts and MC items. The student scale scores account for differences in prompt difficulty. In other words, a student's ability would be estimated in relation to the difficulty of the prompt such that the same prompt ratings would not imply the same level of ability, but would imply a level of ability captured by item parameters. For example, receiving a total rating of 8 on a "hard" prompt would imply more ability than receiving an "8" on an easy prompt, where hard and easy are defined by the overall prompt scores in relation to the same set of MC scores.

### *IRT Results*

Distributions of the three 3PL item parameters are presented in Table 14 for MC items. The parameters are in the IRT traditional metric,[6] and the achievement scale can be interpreted as a standard scale with a true score mean of 0 and standard deviation of 1. The "A" parameter indicates the slope of the curve. The steeper the slope (the larger the "A"), the more the item contributes to the estimation of achievement scores. "A" is similar to item-total correlation. For reference, the "A" for the sample curve in Figure 1 is 1.1. As long as there are enough items, items with lower slopes are useful. Table 14 shows that the median "A" parameter across all grades is approximately 0.70.

[5] Scores are calculated using maximum likelihood estimation. Interested readers should see Baker & Kim (2004).

[6] A, B, and C are reported, where $P(\theta) = C + (1-C)/(1+ \exp(-1.7A(\theta-B))$ (Lord & Novick, 1968). See FDOE 2004 for a more detailed explanation of IRT metrics.

**Table 14.** Multiple-Choice Item Parameter Summary—Traditional Metric— Writing+ Core Items

| Grade<br>(No. of MC Items) | Parameter | Minimum | 25th Percentile | Median | 75th Percentile | Maximum |
|---|---|---|---|---|---|---|
| 4 | A | 0.29 | 0.55 | 0.67 | 0.85 | 1.06 |
| (44) | B | -3.48 | -1.02 | -0.36 | 0.21 | 2.78 |
| | C | 0.03 | 0.10 | 0.22 | 0.30 | 0.59 |
| 8 | A | 0.28 | 0.57 | 0.69 | 0.85 | 1.10 |
| (44) | B | -3.20 | -1.18 | -0.18 | 0.46 | 1.51 |
| | C | 0.03 | 0.08 | 0.19 | 0.29 | 0.57 |
| 10 | A | 0.31 | 0.56 | 0.71 | 0.92 | 1.32 |
| (44) | B | -2.65 | -0.85 | -0.26 | 0.50 | 2.60 |
| | C | 0.03 | 0.09 | 0.20 | 0.28 | 0.58 |

The “B” parameter indicates the difficulty of the items by indicating where the item slope at the point of inflection is centered along the achievement scale. “B” is conceptually similar to an item’s *p*-value. For reference, the “B” in Figure 1 is set at 0, which indicates that the curve is centered at the population mean. “B” parameters should be spread across a wide range of achievement to accurately measure students at all levels of ability (i.e., because of the way the curve flattens on the ends, an item centered in the middle of the achievement scale functions well only for students in the center of the achievement distribution). Items with higher and lower “B” parameters help to measure achievement for students in the upper and lower ends of the achievement distribution. Most students score towards the center of the distribution (near the mean, 0), and Table 14 shows that the majority of items have “B” parameters that are within one standard deviation of the mean. Because item information is the highest at the point of the item “B” parameter, the test is most reliable where the majority of the students score. Reliability is not as strong toward the ends of the distributions, or for very high- or low-ability students.

The 3PL “C” parameter factors in the effects of examinees not knowing the answer and still getting the item correct. This is also called the “pseudo-guessing” parameter. Notice in Figure 1 that the curve asymptotes are at a lower value of about 0.2. For MC items with four possible responses, without knowing anything about the item content, the chances of responding correctly are about one in four. Typically, “C” values should be around 0.2. Well-designed items have distractors that are very attractive to those with limited skills who have no knowledge of the correct answer. For this reason, the “C” parameter is sometimes referred to as pseudo-chance, and this aspect of test design results in low “C” values for these items. Higher values may signal poorly functioning distracters or some unusual curriculum emphasis in certain portions of the state. Table 14 shows that median “C” parameters tend to fall in the expected range.

Test characteristic curves (TCCs) were plotted using item parameters from each grade/subject test. In other words, ICCs for all items were summarized into one curve, a TCC. The results for each grade are shown in Figure 3. Achievement (*x*-axis) was transformed to the 100-500 scale (see next section “Scale Conversion”).

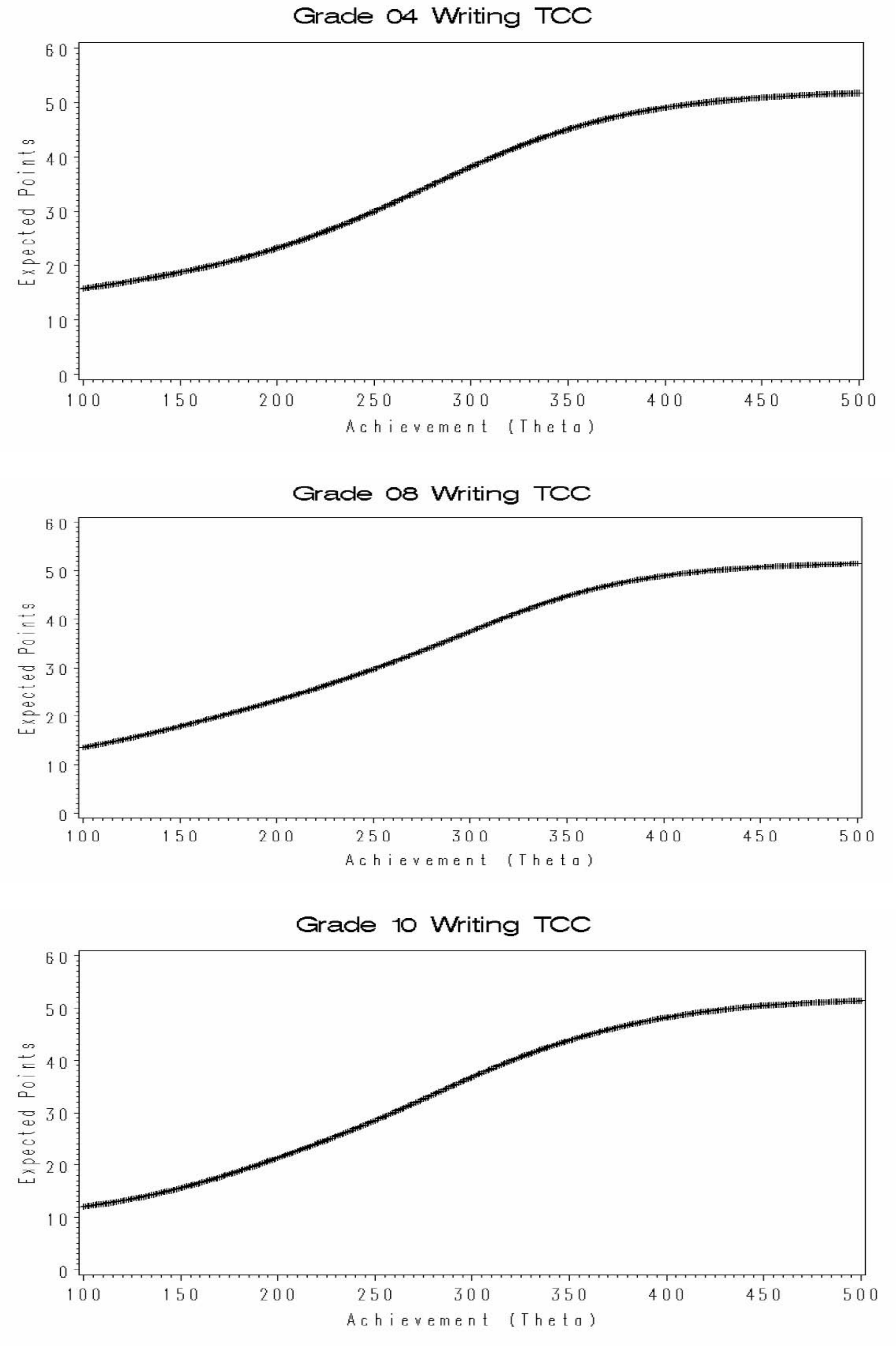

**Figure 3.** Test characteristic curves (TCCs) for FCAT Writing+ by grade.

The item parameters for the 2PPC model used to score PT items are conceptually more difficult to translate graphically. Therefore, Table 15 presents the average of the two ratings for each prompt, as noted above, for the discrimination (A) and category difficulty (category intersection) parameters (D1-D5). The “A” parameters for PT items tend to be higher than those for MC items. Because IRT processing is trying to fit the same achievement construct to all items, the magnitude of the “A” parameters is evidence of the convergence or similarity between the knowledge and skills required for the different item types. The category intersection parameters indicate where the category characteristic curves (see Figure 2) intersect along the achievement index (theta) continuum.

**Table 15.** Writing Prompt Parameters in Traditional Metric

| Grade | Prompt | A | D1 | D2 | D3 | D4 | D5 |
|---|---|---|---|---|---|---|---|
| 4 | 1 | 0.594 | -5.798 | -2.099 | -0.753 | 1.961 | 4.382 |
| | 2 | 0.661 | -4.461 | -3.841 | -2.026 | 0.781 | 3.381 |
| 8 | 1 | 1.108 | -4.092 | -3.041 | -1.255 | 1.069 | 2.308 |
| | 2 | 1.112 | -3.767 | -3.130 | -1.174 | 1.325 | 2.720 |
| 10 | 1 | 1.029 | -3.038 | -2.571 | -1.071 | 1.089 | 2.295 |
| | 2 | 1.124 | -2.673 | -2.243 | -0.907 | 0.816 | 1.748 |

## Scale Conversion

IRT scaling produces item parameters for an achievement scale targeted to a true score mean of 0 and true score standard deviation of 1. The FCAT, however, reports scores on a 100–500 scale. Therefore, a transformation is needed for the IRT item parameters in order for them to produce the appropriate scores. Table 16 contains the multiplicative and additive constants used to transform the traditional IRT metric to the FCAT reporting scale.

**Table 16.** Scale Conversion Multiplicative and Additive Constants

| Grade | M1<br>Multiplier | M2<br>Additive Constant |
|---|---|---|
| 4 | 50 | 300 |
| 8 | 50 | 300 |
| 10 | 50 | 300 |

## IRT Fit Statistics

Again, IRT scaling algorithms attempt to find item parameters (numerical characteristics) that create a match between observed patterns of item responses and theoretical response patterns defined by the selected IRT models. The Q1 statistic (Yen, 1981) may be used as an index for how well theoretical item curves are found that match observed item responses. Q1 is computed by (1) conducting IRT item-parameter estimation, (2) estimating students' achievement using the estimated item parameters, and (3) using student achievement scores in combination with estimated item parameters to compute expected performance on each item. Differences between expected item performance and observed item performance are then compared at selected intervals across the range of student achievement. Q1 is computed as a ratio involving expected and observed item performance and is therefore interpretable as a chi-square ($\chi 2$) statistic, which is a statistical test that determines whether the data (observed item performance) fit the hypothesis (the expected item performance).

Q1 for each item type has varying degrees of freedom because the different types of items have different numbers of IRT parameters. This means that Q1 is not directly comparable across item types. An adjustment or linear transformation (translation to a *z*-score, $Z_{Q1}$) is made for different numbers of item parameters and sample sizes to create a more comparable statistic. The FCAT has set a criterion for a minimum $Z_{Q1}$value standard for an item to have acceptable fit (FDOE, 1998).[7] Table 17 presents the distributions of $Z_{Q1}$ for Writing+. Table 18 presents the number of poorly fitting items by item type. Nearly all of the items in Grade 4 and a large number of items in Grades 8 and 10 exhibit poor fit.

**Table 17.** $Z_{Q1}$ Statistic, Summary Data—All Writing+ Items

| Grade | Minimum | 25th Percentile | Median | 75th Percentile | Maximum |
|---|---|---|---|---|---|
| 4 | 29.78 | 154.80 | 267.99 | 389.10 | 703.26 |
| 8 | 11.92 | 36.63 | 57.53 | 89.22 | 203.95 |
| 10 | 17.06 | 48.68 | 62.51 | 97.30 | 217.03 |

**Table 18.** Number of Poorly Fitting Items According to Q1 Statistics—All Items

| | Writing+ | |
|---|---|---|
| Grade | MC | Prompt |
| 4 | 42/44 | 2/2 |
| 8 | 19/44 | 2/2 |
| 10 | 26/44 | 2/2 |

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

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

## Achievement Scale Unidimensionality

By fitting all items simultaneously to the same achievement scale, IRT is operating under the assumption that there is a strong, single construct that underlies the performance of all items. Under this assumption, performance on the items should be related to achievement (as depicted by Figure 1). 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 dependence" assumption of unidimensional IRT which suggests a relatively straightforward test for unidimensionality, called the Q3 statistic (Yen, 1984).

Computation of the Q3 statistic begins the same as the Q1 statistic: expected student performance on each item is calculated using item parameters and estimated achievement scores. Then, for each student and each item, the difference between expected and observed item performance is calculated. The difference can be thought of as the "leftover" variance in performance after accounting for underlying achievement. If performance on an item is driven by a single achievement construct, then not only will this residual be small (as tested by the Q1 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 Q3 statistic.

With *n* items, there are *n*(*n* – 1)/2 Q3 statistics. For example, Grade 4 Writing+ has 45 items and 990 Q3 values. The Q3 values should all be small. Q3 data are summarized in Table 19 by minimum, 5th percentile, median, 95th percentile, and maximum values for each grade. To add perspective to the meaning of the Q3 distributions, the average zero-order correlation (item intercorrelation) among item responses is also shown. If the achievement construct is "accounting for" the relationships among the items, Q3 values should be much smaller than the zero-order correlations. Table 19 indicates that the median Q3 for all grades is within acceptable range, but there are a number of Q3 values that suggest the writing assessment is not strictly unidimensional.

**Table 19.** Q3 Statistic, Summary Data—All Writing+ Items

| | | **Q3 Distribution** | | | | |
|---|---|---|---|---|---|---|
| **Grade** | **Average Zero-order Correlation** | **Minimum** | **5th Percentile** | **Median** | **95th Percentile** | **Maximum** |
| 4 | 0.112 | -0.348 | -0.256 | 0.060 | 0.142 | 0.343 |
| 8 | 0.112 | -0.241 | -0.173 | 0.022 | 0.064 | 0.419 |
| 10 | 0.120 | -0.257 | -0.192 | 0.020 | 0.063 | 0.240 |

## Item Bias Analyses

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

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

Table 20 presents the distribution of SMD summary ratings by grade. Ratings for the vast majority of items fall in the lowest two categories. Nevertheless, one item in Grade 4 and four items in Grade 10 obtained a rating of “4,” indicating a non-trivial performance difference.

**Table 20.** Item DIF Rating Summary—Writing+

| | **Standardized Mean Difference (SMD) Rating** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Low** | | | | **High** | | |
| **Grade** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| 4 | 32 | 10 | 3 | 1 | 0 | 0 | 0 |
| 8 | 36 | 8 | 2 | 0 | 0 | 0 | 0 |
| 10 | 35 | 5 | 2 | 4 | 0 | 0 | 0 |

## Test Reliability, Standard Error of Measurement, and Information

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

Test reliability concerns the concept that a test score results from some true level of achievement plus measurement error. For a population of students, reliability is a ratio of variation in true achievement compared with variation in observed test scores. The less that measurement error contaminates test scores, the closer the ratio is to 1. Under classical test theory, measurement error is assumed to be the same at all levels of achievement, and one reliability coefficient can be estimated to acknowledge that error.

Within the IRT framework, however, measurement error is not assumed to be constant across the range of ability. Rather, standard error of measurement (SEM) is a function of how well a student's pattern of item responses matches the expected response pattern uncovered by the IRT modeling processes. In other words, with IRT modeling, score assignment is more accurate for a student who correctly answers the easy items and misses the difficult items than for a student who gets as many easy items correct as difficult items. Furthermore, score assignment tends to be more accurate for students toward the center of the distribution than for students with more extreme scores. Another way to determine the amount of precision in estimating achievement is to look at *information*. In IRT, a test's information is inversely related to SEM ($1/\sigma_2$). Therefore, if the amount of information on the ability scale is large, then ability can be estimated with precision for students whose true ability is at that level (Baker, 2001).

Conditional standard error curves, depicted in Figure 4, are used to depict test reliability. The curves plot the average SEM extracted from student score records as a function of achievement level. SEM is like a standard deviation, so that approximately two-thirds of the students with a given level of achievement will have observed test scores within one SEM of the given true score. For example, the Grade 4 SEM plot in Figure 5 shows that a student whose true achievement level is 200 will have an SEM of approximately 25. That means that approximately two-thirds of those students will have test scores between 175 and 225. The remaining one-third of the students with a true achievement level of 200 will have test scores more than 25 points away from 200. As expected, SEM is larger at the tails of the achievement level distribution and smaller in the center. Most students, however, are in the center of the distribution.

Test information functions (TIFs), seen in Figure 5, show the amount of information as plotted on the 100-500 achievement scale. The TIFs generally peak around an achievement value of 300. The peaks can be interpreted to mean that these tests estimate achievement more precisely around 300 and with less precision at other levels of achievement. A flatter curve means a test estimates achievement with more equal precision across that range of achievement.

It is possible to synthesize an overall reliability system from the standard error curves by using the average SEM for all students to compute a "marginal" reliability. These values, which can be interpreted like traditional reliability statistics such as Cronbach's alpha, are presented in Table 22.

While marginal reliability estimates were computed using only the calibration sample, it is important to note that the SEM curves and reliability estimates were computed using all students who received scores, including the non-standard curriculum students. This was done in order to make reliability data consistent across grades and subjects and not confounded by any differences in calibration samples. In addition, these estimates are consistent with the application of the FCAT; they characterize test results for all students who receive scores.

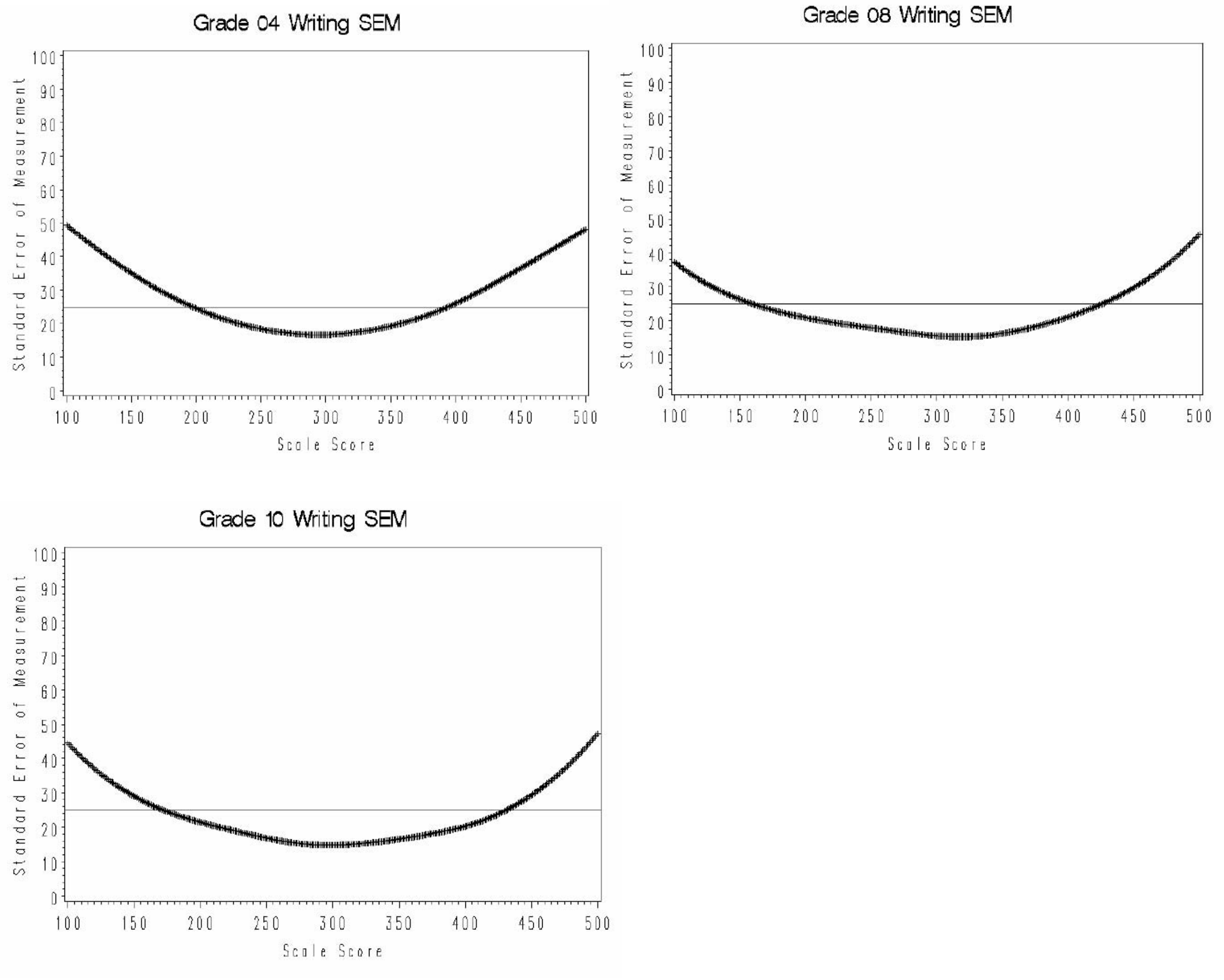

**Figure 4.** Standard error of measurement plots for 2006 FCAT Writing+ by grade.

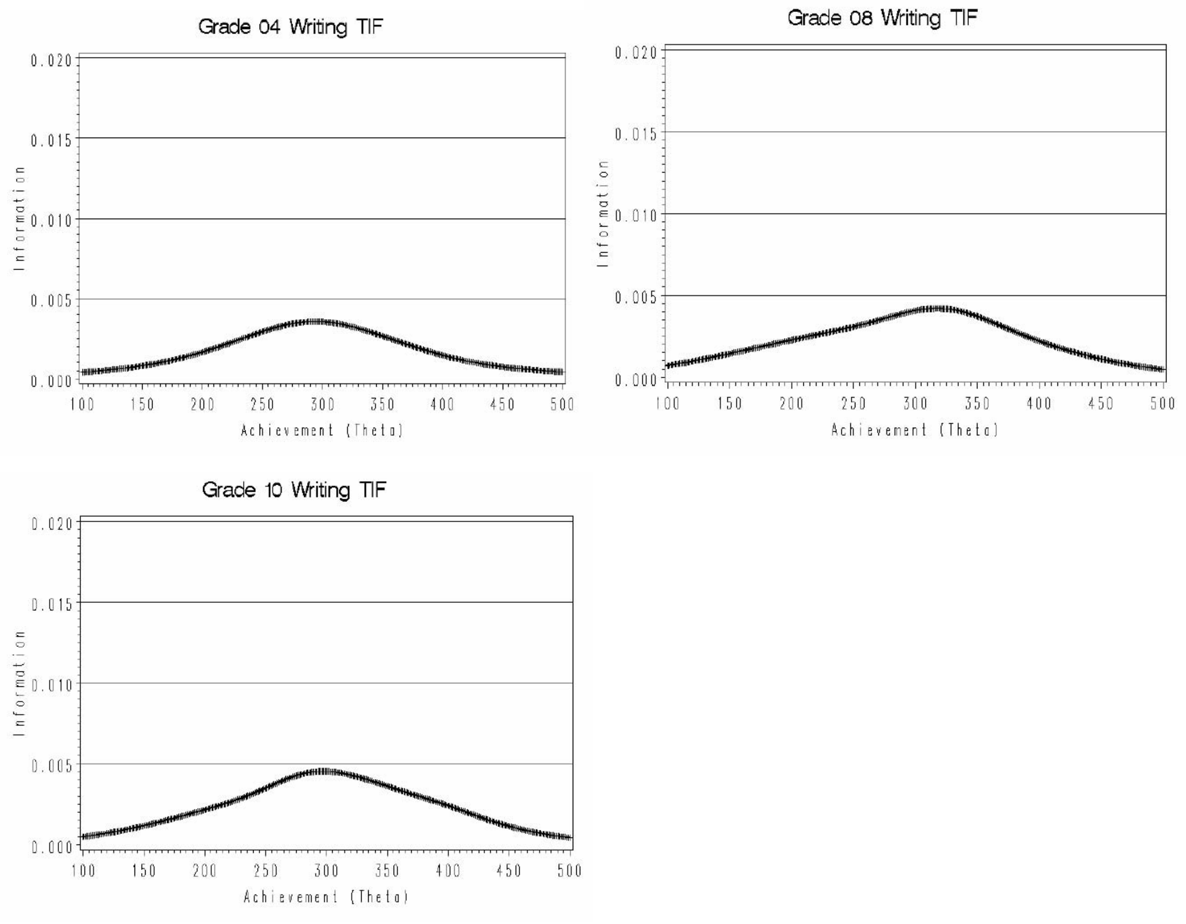

**Figure 5.** Test information functions (TIFs) for 2006 FCAT Writing+ by grade.

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

Table 21 also shows traditional Cronbach’s alpha reliability statistics. These estimates are based on raw scores only and have been calculated for the total set of items and for the items comprising each of the separate reporting categories. The numbers of items are in parentheses.

**Table 21.** IRT Marginal Reliabilities and Cronbach’s Alpha.

| Writing+ Grade | IRT Marginal $r_{ii}$ | Cronbach’s Alpha | | | | |
|---|---|---|---|---|---|---|
| | | **Total** | **Focus** | **Organization** | **Support** | **Conventions** |
| 4 | 0.875 | 0.848 | 0.582 (10) | 0.670 (10) | 0.568 (8) | 0.611 (16) |
| 8 | 0.896 | 0.847 | 0.629 (13) | 0.546 (6) | 0.650 (10) | 0.607 (15) |
| 10 | 0.904 | 0.857 | 0.667 (11) | 0.587 (8) | 0.601 (9) | 0.622 (16) |

## Intercorrelations among Reporting Categories and Scale Scores

Intercorrelations among IRT-derived scale scores, total raw scores, and the FCAT reporting categories are presented in Tables 22–27. Correlations between total raw scores and IRT overall scale scores are high (0.83 to 0.91), but not as high as observed in other FCAT assessments (i.e. Reading, Mathematics, or Science). This is due to the weighting of the prompt and MC components of the Writing+ assessment. That is, although the MC items comprise nearly ninety percent of the raw score points, the MC and prompt components are simply averaged (i.e., each receive a weight of 0.5) in determining the overall scale score.

Similarly, correlations between writing prompt raw scores and IRT overall scale scores are also high (0.90–0.91). Correlations between prompt scores and MC scores exhibit moderate relationships (0.39–0.58), indicating that the MC and prompt components are neither entirely redundant nor unrelated. Such correlations suggest the constructs assessed by MC and prompt components of the assessment are measuring constructs that are related, yet still distinct.

Comparisons of the correlations among reporting category scales are affected by differences in scale reliabilities (see Table 21) that result from differences in numbers of items in the categories. For example, in Grade 8 (Table 24), observed correlations in the Organization category would be expected to be lower than the other categories because it is measured with fewer items than the other categories. This means that all of the correlations among the reporting categories are underestimated due to lower reliabilities of corresponding subscores.

**Table 22.** Grade 4 Writing+ Reporting Category and Scale Score Intercorrelations

| | Total Raw Score (45) | Focus (10) | Organization (10) | Support (8) | Conventions (16) |
|---|---|---|---|---|---|
| **Scale Score** | 0.828 | 0.581 | 0.599 | 0.533 | 0.559 |
| **Total Raw Score** | 1 | 0.797 | 0.824 | 0.735 | 0.790 |
| **Focus** | -- | 1 | 0.616 | 0.549 | 0.502 |
| **Organizations** | -- | -- | 1 | 0.589 | 0.527 |
| **Support** | -- | -- | -- | 1 | 0.436 |

Note: Number of items in parentheses; N = 19,362

**Table 23.** Grade 4 Writing+ Reporting Category and Scale Score Intercorrelations

| | Prompt 1 Raw Score | Prompt 2 Raw Score | Prompt Scale Score | MC Raw Score | MC Scale Score |
|---|---|---|---|---|---|
| **Scale Score** | 0.919 | 0.904 | 0.917 | 0.704 | 0.725 |
| **Prompt 1 Raw Score** | 1 | -- | 0.995 | 0.378 | 0.389 |
| **Prompt 2 Raw Score** | -- | 1 | 0.986 | 0.395 | 0.401 |
| **Prompt Scale Score** | -- | -- | 1 | 0.383 | 0.391 |
| **MC Raw Score** | -- | -- | -- | 1 | 0.966 |

N = 19,362

**Table 24.** Grade 8 Writing+ Reporting Category and Scale Score Intercorrelations

| | Total Raw Score (45) | Focus (13) | Organization (6) | Support (10) | Conventions (15) |
|---|---|---|---|---|---|
| **Scale Score** | 0.900 | 0.700 | 0.602 | 0.695 | 0.661 |
| **Total Raw Score** | 1 | 0.839 | 0.705 | 0.826 | 0.810 |
| **Focus** | -- | 1 | 0.561 | 0.622 | 0.543 |
| **Organizations** | -- | -- | 1 | 0.532 | 0.450 |
| **Support** | -- | -- | -- | 1 | 0.545 |

Note: Number of items in parentheses; N = 22,238

**Table 25.** Grade 8 Writing+ Prompt Score and Scale Score Intercorrelations

| | Prompt 1 Raw Score | Prompt 2 Raw Score | Prompt Scale Score | MC Raw Score | MC Scale Score |
|---|---|---|---|---|---|
| **Scale Score** | 0.910 | 0.905 | 0.909 | 0.822 | 0.841 |
| **Prompt 1 Raw Score** | 1 | -- | 0.997 | 0.539 | 0.546 |
| **Prompt 2 Raw Score** | -- | 1 | 0.996 | 0.526 | 0.535 |
| **Prompt Scale Score** | -- | -- | 1 | 0.532 | 0.539 |
| **MC Raw Score** | -- | -- | -- | 1 | 0.971 |

N = 22,238

**Table 26.** Grade 10 Writing+ Prompt Score and Scale Score Intercorrelations

| | Total Raw Score (45) | Focus (11) | Organization (8) | Support (9) | Conventions (16) |
|---|---|---|---|---|---|
| **Scale Score** | 0.911 | 0.719 | 0.664 | 0.677 | 0.636 |
| **Total Raw Score** | 1 | 0.836 | 0.779 | 0.800 | 0.791 |
| **Focus** | -- | 1 | 0.620 | 0.622 | 0.525 |
| **Organizations** | -- | -- | 1 | 0.582 | 0.492 |
| **Support** | -- | -- | -- | 1 | 0.512 |

Note: Number of items in parentheses; N = 21,395

**Table 27.** Grade 10 Writing+ Prompt Score and Scale Score Intercorrelations

| | Prompt 1 Raw Score | Prompt 2 Raw Score | Prompt Scale Score | MC Raw Score | MC Scale Score |
|---|---|---|---|---|---|
| **Scale Score** | 0.908 | 0.912 | 0.915 | 0.823 | 0.845 |
| **Prompt 1 Raw Score** | 1 | -- | 0.995 | 0.535 | 0.544 |
| **Prompt 2 Raw Score** | -- | 1 | 0.986 | 0.568 | 0.577 |
| **Prompt Scale Score** | -- | -- | 1 | 0.547 | 0.556 |
| **MC Raw Score** | -- | -- | -- | 1 | 0.969 |

N = 21,395

## Student Classification Accuracy and Consistency

Based on their FCAT scale scores, students are classified into one of five performance levels. While it is important to know the reliability of student scores in any examination, of even greater importance is assessing the reliability of the classification decisions based on these scores. Evaluation of the reliability of classification decisions is performed through estimation of the probabilities of correct and consistent classification of students. Procedures were used from Livingston and Lewis (1995) and Lee, Hanson, and Brennan (2000) to derive measures of the accuracy and consistency of the classifications. A brief description of the procedures used and the results derived from them are presented in this section. As a reference, Table 28 provides the cutpoints for classification into the FCAT Writing+ performance levels.

**Table 28.** Cutpoints for FCAT Writing+ Performance Level Classifications

| Grade | Performance Level | | | |
|---|---|---|---|---|
| | LEVEL 2 | LEVEL 3 | LEVEL 4 | LEVEL 5 |
| **4** | 240 | 290 | 365 | 427 |
| **8** | 250 | 299 | 356 | 416 |
| **10** | 250 | 300 | 342 | 403 |

### *Accuracy of Classification*

According to Livingston and Lewis, the accuracy of a classification is “. . . the extent to which the actual classifications of the test takers . . . agree with those that would be made on the basis of their true score, 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 somehow be obtained (Young and Yoon, 1998). Since these true scores are not available, Livingston and Lewis provide a method to estimate the true score distribution of a test and create the cross-tabulation of the true score and observed score classifications. An example of the 5×5 cross-tabulation of the true score versus observed score classifications for FCAT Grade 4 Writing+ is given in Table 29. It shows the proportions of students who were classified into each performance category by the actual observed scores and by estimated true scores. The detailed procedure for calculating accuracy of classification is presented in Appendix B.

**Table 29.** 2006 FCAT Grade 4 Writing+ True Scores vs. Observed Scores Cross-Tabulation (Accuracy Table)

| True Score | Observed Score | | | | | |
|---|---|---|---|---|---|---|
| | LEVEL1 | LEVEL2 | LEVEL3 | LEVEL4 | LEVEL5 | Total |
| LEVEL1 | 0.126 | 0.030 | 0.000 | 0.000 | 0.000 | 0.156 |
| LEVEL2 | 0.043 | 0.200 | 0.053 | 0.000 | 0.000 | 0.296 |
| LEVEL3 | 0.001 | 0.066 | 0.315 | 0.048 | 0.001 | 0.430 |
| LEVEL4 | 0.000 | 0.000 | 0.028 | 0.069 | 0.015 | 0.113 |
| LEVEL5 | 0.000 | 0.000 | 0.000 | 0.002 | 0.003 | 0.005 |
| Total | 0.169 | 0.296 | 0.397 | 0.119 | 0.019 | 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 and Lewis, 1995, p. 179). 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. An example of 5×5 cross-tabulation between a form taken and an alternate form for FCAT Grade 4 Writing+ is provided in Table 30. The table shows the proportions of students who were classified into each performance category by the actual test and by another (hypothetical) parallel test form.

Note that the consistency table is symmetrical, but the accuracy table is non-symmetrical because it compares classifications based on two different types of scores. Also note that agreement rates are lower in the consistency table because both classifications contain measurement error, whereas in the accuracy table, true score classification is assumed to be errorless.  The detailed procedure for calculating consistency of classification is presented in Appendix B.

**Table 30.** 2006 FCAT Grade 4 Writing+ True Scores vs. Observed Scores Cross-Tabulation (Consistency Table)

| Form Taken | Alternate Form | | | | | |
|---|---|---|---|---|---|---|
| | LEVEL 1 | LEVEL 2 | LEVEL 3 | LEVEL 4 | LEVEL 5 | Total |
| LEVEL 1 | 0.117 | 0.048 | 0.003 | 0.000 | 0.000 | 0.169 |
| LEVEL 2 | 0.048 | 0.168 | 0.079 | 0.001 | 0.000 | 0.296 |
| LEVEL 3 | 0.003 | 0.079 | 0.261 | 0.050 | 0.003 | 0.397 |
| LEVEL 4 | 0.000 | 0.001 | 0.050 | 0.056 | 0.011 | 0.119 |
| LEVEL 5 | 0.000 | 0.000 | 0.003 | 0.011 | 0.005 | 0.019 |
| Total | 0.169 | 0.296 | 0.397 | 0.119 | 0.019 | 1.000 |

Note: Columns and row totals are computed from non-rounded values. Shaded cells are used for computing consistency index conditional on level.

### *Accuracy and Consistency Indices*

There are three types of accuracy and consistency indices that can be generated from these tables: *overall, conditional on level,* and *cutpoint*. In order to facilitate interpretations of these indices, a brief outline of computational procedures used to derive accuracy indices will be presented using the example of the FCAT Grade 4 Writing+ 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 areas in Table 29. This is a proportion (or percentage) of correct classifications across all the levels. In this particular example, the overall accuracy index for the FCAT Grade 4 Writing+ test equals 0.713 (71.3 percent). This means that 71.3 percent of students are classified in the same performance categories based on their observed scores as would be classified based on their true scores, if the true scores 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 30, it can be determined that the overall consistency index for the FCAT Grade 4 Writing+ test equals 0.607 (60.7 percent). In other words, 60.7 percent of Grade 4 students would have been classified in the same performance levels based on the alternate test form if they had taken that test instead. Another way to express *overall consistency* is to use Cohen's *kappa (κ)* coefficient (Cohen, 1960). The overall coefficient kappa when applying all cutoff scores together is

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

where *P* is the probability of consistent classification, and $P_c$ is the probability of consistent classification by chance (Lee, 2000). Kappa is a measure of ". . . how much agreement exists beyond chance alone . . ." (Fleiss, 1973), which means that it assesses the proportion of consistent classifications between two test forms after removing the proportion of consistent classifications expected by chance alone. The data from Table 32 indicates that Cohen's κ for FCAT Grade 4 Writing+ equals 0.449. Compared to the previously described overall consistency estimate, Cohen's κ has lower value because it has been 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 students classified into that level (marginal entry). In Table 30, the row LEVEL 4 is outlined, and corresponding cells are shaded. The ratio between 0.056 (proportion of correct classifications) and 0.119 (total proportion of students classified into the LEVEL 4) yields 0.471, which represents the index of consistency of classification for FCAT Grade 4 Writing+ that is conditional on LEVEL 4. It indicates that 47.1 percent 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 29 the proportion of agreement between true score status and observed score status at LEVEL 1 is 0.126, whereas the total proportion of students with true score status at this level is 0.156. The accuracy conditional on level is equal to the ratio between those two proportions, which yields 0.808. This indicates that 80.8 percent of the students estimated to have true score status on LEVEL 1 are correctly classified into that category by their observed scores on the FCAT Grade 4 Writing+ test.

Perhaps the most important indices for accountability systems are those for the accuracy and consistency of classification decisions made at specific cutpoints. To evaluate decisions at specific cutpoints, the joint distribution of all the performance levels has been collapsed into a dichotomized distribution around that specific cutpoint. For example, the dichotomization at the cutpoint that separates LEVEL 1 through LEVEL 3 (combined) from LEVEL 4 and LEVEL 5 (combined) for FCAT Grade 4 Writing+ is depicted in Table 31. The proportion of correct classifications below that particular cutpoint is equal to the sum of the cells in the upper left shaded area (0.834), and the proportion of correct classifications above the particular cutpoint is equal to sum of the cells in the lower right shaded area (0.089).

**Table 31.** 2006 FCAT Writing+ Grade 4 True Scores vs. Observed Scores Cross-Tabulation (Accuracy Table)

| **True Score** | **Observed Score** | | | | | |
|---|---|---|---|---|---|---|
| | **LEVEL1** | **LEVEL2** | **LEVEL3** | **LEVEL4** | **LEVEL5** | **Total** |
| **LEVEL1** | 0.126 | 0.030 | 0.000 | 0.000 | 0.000 | 0.156 |
| **LEVEL2** | 0.043 | 0.200 | 0.053 | 0.000 | 0.000 | 0.296 |
| **LEVEL3** | 0.001 | 0.066 | 0.315 | 0.048 | 0.001 | 0.430 |
| **LEVEL4** | 0.000 | 0.000 | 0.028 | 0.069 | 0.015 | 0.113 |
| **LEVEL5** | 0.000 | 0.000 | 0.000 | 0.002 | 0.003 | 0.005 |
| **Total** | 0.169 | 0.296 | 0.397 | 0.119 | 0.019 | 1.000 |

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

The *accuracy index at cutpoint* is computed as the sum of the proportions of correct classifications around a selected cutpoint.[8] In the example from Table 31, the sum of both shaded areas (upper left shaded areas added to lower right shaded areas) equals 0.923, which means that 92.3 percent of students were correctly classified either above or below the particular cutpoint. The sum of the proportions in the upper right non-shaded area (0.049) indicates false positives (i.e., 4.9 percent of students are classified above the cutpoint by their observed score, but fall below the cutpoint by their true score). The sum of the lower left non-shaded area (0.028) is the

[8] All cuts are from the Writing+ Standard Setting Committee, prior to State Board of Education approval. They are subject to change.

proportion of false negatives (i.e., 2.8 percent of students are observed below the cutpoint level whose true level is actually above the cutpoint).

The *consistency at cupoint* is obtained in an analogous way. For example, if data are taken from Table 30 and the distribution is dichotomized at the cutpoint between LEVEL 1 and all other levels combined, it can be determined that the proportion of correct classifications around that cutpoint equals 0.896. This means that 89.6 percent of students would have been classified in the same two categories (LEVEL 1, or LEVEL 2 through LEVEL 5 combined) as their actual test form taken on an alternate test form (if they had taken it).

### *Accuracy and Consistency Results for 2006 FCAT*

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

The overall indices of accuracy and consistency of classification for 2006 FCAT Writing+ tests are presented in Table 32.

**Table 32.** Estimates of Accuracy and Consistency of Performance-Level Classification for Writing+ by Grade

| Grade | Accuracy | Consistency | Kappa (κ) |
|---|---|---|---|
| **4** | 0.714 | 0.608 | 0.449 |
| **8** | 0.693 | 0.591 | 0.443 |
| **10** | 0.729 | 0.630 | 0.516 |

Table 32 shows that overall accuracy indices are in the range between 0.693 and 0.729, overall consistency indices range between 0.591 and 0.630, and κ coefficients fall in the range between 0.443 and 0.516.

In addition to overall ratings of decision accuracy, the levels of agreement at each performance level are also of interest. Table 33 displays the probability of students being classified as being in a particular performance level, given that their “true status” was the same category. In most tests, the accuracy indices at the lowest performance level (LEVEL 1) are substantially higher than at other levels. Similarly, the accuracy at the highest performance level is also elevated, but not so evidently as at the lowest level. This effect is due to the fact that extreme performance levels usually cover a wider range of the measured construct than the intermediate levels, and misclassification can occur in only one direction. It should be noted that the percentage of students whose observed scores are classified in the highest performance level is relatively low [below 10 percent in most of the tests (see Appendix A)] which makes indices conditional at that level less reliable. In one instance (Grade 8 Writing+), the percentage of students whose estimated true scores fall in the LEVEL 5 equals zero, which makes it impossible to estimate the accuracy at that level. It is possible, however, to estimate accuracy of decisions at the cutpoint between LEVEL 4 and LEVEL 5. Moreover, this estimate can be high (see Table 34).

**Table 33.** Accuracy of Classification at each Proficiency Level for Writing+ by Grade

| Grade | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 |
|---|---|---|---|---|---|
| **4** | 0.809 | 0.677 | 0.732 | 0.614 | 0.597 |
| **8** | 0.861 | 0.695 | 0.573 | 0.670 | * |
| **10** | 0.860 | 0.708 | 0.627 | 0.750 | 0.713 |

*No accuracy estimates were calculated at LEVEL 5 for Grade 8 Writing+ because the number of estimated true scores in this cell is zero.

The most important decisions about student scores often involve dichotomous choices. For example, the stakes are usually highest regarding decisions made at the pass-fail cutpoint, which makes it desirable to know the accuracy and consistency of dichotomous decisions made around that specific cutpoint. For example, if a college gives credits to advanced and proficient students who achieved LEVEL 4 and LEVEL 5, but not to those in LEVEL 1 through LEVEL 3, the focus of interest would be on the accuracy and consistency of dichotomous decisions below (versus those at or above) the LEVEL 4 threshold. Reporting in a "percent at-or-above cut" (PAC) metric requires a judgment about whether the student score is below or at-or-above a particular cutpoint. Table 34 presents the accuracy and consistency information for these dichotomous categorizations.

**Table 34.** Accuracy and Consistency of Dichotomous Categorizations for Writing+ by Grade (PAC metric)

| Grade | Accuracy | | | | Consistency | | | |
|---|---|---|---|---|---|---|---|---|
| | **1<br>/<br>2+3+4+5** | **1+2<br>/<br>3+4+5** | **1+2+3<br>/<br>4+5** | **1+2+3+4<br>/<br>5** | **1<br>/<br>2+3+4+5** | **1+2<br>/<br>3+4+5** | **1+2+3<br>/<br>4+5** | **1+2+3+4<br>/<br>5** |
| **4** | 0.927 | 0.880 | 0.922 | 0.982 | 0.897 | 0.833 | 0.891 | 0.972 |
| **8** | 0.914 | 0.877 | 0.907 | 0.992 | 0.879 | 0.829 | 0.870 | 0.985 |
| **10** | 0.933 | 0.905 | 0.916 | 0.973 | 0.906 | 0.867 | 0.882 | 0.962 |
| **10<br>P / F** | 0.906 | | | | 0.868 | | | |

The data in Table 34 reveal that the level of agreement, in terms of both accuracy and consistency for these dichotomous categorizations, is very high. Although the rates of agreement for decision consistency are slightly lower, the rate of agreement does not fall below 82.9

percent. This means high rates of accuracy and consistency are available to support decisions about PACs.

The conclusion about high accuracy of PAC decisions is also supported by data on the percentages of false positives and false negatives derived from the dichotomized “true status” versus “observed status” categorizations (see Table 35). On average, only 3.71 percent of students were classified in a lower or higher level than their “true” level across all grades and subjects. The range of false positives and false negatives is from 0.000 to 0.071, indicating that not more than 7.1 percent of students were classified differently from a level meeting the standard.

**Table 35.** Accuracy of Dichotomous Categorizations: False Positives and False Negatives Rates (PAC Metric)

| **Grade** | **False Positives** | | | | **False Negatives** | | | |
|---|---|---|---|---|---|---|---|---|
| | **1<br>/<br>2+3+4+5** | **1+2<br>/<br>3+4+5** | **1+2+3<br>/<br>4+5** | **1+2+3+4<br>/<br>5** | **1<br>/<br>2+3+4+5** | **1+2<br>/<br>3+4+5** | **1+2+3<br>/<br>4+5** | **1+2+3+4<br>/<br>5** |
| **4** | 0.030 | 0.053 | 0.049 | 0.016 | 0.043 | 0.066 | 0.029 | 0.002 |
| **8** | 0.029 | 0.052 | 0.050 | 0.008 | 0.057 | 0.071 | 0.043 | 0.000 |
| **10** | 0.029 | 0.050 | 0.049 | 0.020 | 0.037 | 0.045 | 0.036 | 0.006 |
| **10<br>P / F** | 0.050 | | | | 0.044 | | | |

* False negatives could not be estimated at 1+2+3+4 vs. 5 cutpoint for Grade 8 Writing+ because the number of estimated true scores in the LEVEL 5 cell is zero.

The issue of dichotomous classifications has particular relevance in the case of high-stakes situations, such as that exemplified by the high school graduation standard proposed for the Grade 10 Writing+ test. Under these circumstances, a student hoping to receive a regular diploma will be required, among other things, to achieve a score of 295 or better on the FCAT Writing+ test.[9] In principle, it is possible for three situations to be found.

1. A student’s observed performance is accurately reflected in terms of the standard and in terms of his or her true level of ability (i.e., a student whose ability is at or above the minimum acceptable standard achieves a test score at or above that standard, and a student whose true ability is below the standard achieves a score below the standard). Students scoring below the standard will be required to take the test again.

2. A student whose true ability is below the standard receives a score that is, in fact, above

[9] At the time the data for this report was generated, a score of 295 on the Grade 10 FCAT Writing+ assessment was the proposed cutpoint for passing. This cutpoint was later changed to 300 by the State Board of Education, but legislation ultimately suspended the writing graduation requirement

the standard (false positives).

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

False-positive and false-negative rates for all dichotomous classifications for FCAT tests are presented in Table 35. An examination of the FCAT results for the Grade 10 Writing+ test in terms of the high school standards reveals that because the threshold score for fail-pass decisions in Grade 10 Writing+ falls within performance LEVEL 2, a separate analysis to estimate the accuracy of fail-pass decisions for this test needed to be performed. The analysis showed that 90.6 percent of students were classified correctly into either a pass or fail category (situation 1) based on their observed performance in Grade 10 Writing+.

# REFERENCES

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

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

Baker, F. B. and Kim, S. *Item Response Theory: Parameter Estimation Techniques*. New York: Marcel Dekker.

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

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

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

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

Florida Department of Education (1996). *Sunshine State Standards.* Retrieved September 20, 2002, from the Florida Department of Education website: http://www.firn.edu/doe/curric/prek12/frame2.htm

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

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

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

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

Florida Department of Education (January 2002). *Florida Comprehensive Assessment Test Technical Report Field Test Supplement for Test Administration in Spring 2001.*

Unpublished. Tallahassee, Fla.: Author.

Florida Department of Education (November 2003) *Plan for Selecting the Calibration Sample for the 2006 FCAT Administration.* Unpublished. Tallahassee, Fla.: Author.

Florida Department of Education (2004). *The FCAT 2005 Test Construction Specifications.* Unpublished. Tallahassee, Fla.: Author.

Lee, W., Hanson, B. A., & Brennan, R. L. (2000, October). *Procedures for computing classification consistency and accuracy indices with multiple categories*. (ACT Research Report Series 2000-10). Iowa City, Iowa: ACT, Inc.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32(2), pp. 179-197.

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

Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *Journal of American Statistical Association*, 58, pp. 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, pp. 719-748.

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

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

Rogosa, D. (2000). Statistical topics in educational assessment: individual scores, group summaries, and accountability systems. Presented to the March 14, 2000 CCSSO Technical Issues in Large Scale Assessment Workshop, San Diego, Calif.

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

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, pp. 2, 245-262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement,* 2, pp. 125-145.

Young, M. J. & Yoon, B. (1998, April). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment.* (CSE Technical Report 475). Center for the Study Evaluation, National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles, Calif.: University of California, Los Angeles.

Zwick, R., Donoghue, J. R. & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*. 30(3), pp. 233-251.