Appendix A: Statistical Indicators Used in Test Data Analysis

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

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

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

Table 16: Statistical Analyses for Test Data and Indicators

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

Indicator Definitions

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

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

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

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

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

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

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

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

Reliability Measures

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

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