BUROS
Center for Testing

Report on the
2006 and 2007 Florida Comprehensive Assessment Test Grade 3 Reading Scores:

Prepared for the

Florida Department of Education

and the

Florida FCAT External Review Committee

by:

Kurt F. Geisinger, Ph.D.

Brett P. Foley, M.S. (Doctoral Student)

Craig S. Wells, Ph.D.

Rebecca Norman, B.A. (Doctoral Student)

Andrew C. Dwyer, M.S. (Doctoral Student)

Carina M. McCormick, B.A. (Doctoral Student)

Anja Römhild, M. A. (Doctoral Student)

November, 2007

Questions concerning this report can be addressed to:
Kurt F. Geisinger, Ph.D.
Buros Center for Testing
21 Teachers College Hall
University of Nebraska – Lincoln
Lincoln, NE, 68588-0353
kgeisinger2@unl.edu

Executive Summary

In 2006, 8% more third-graders achieved proficiency status on the FCAT Reading test than in 2005. This sudden increase happened in the context of increases averaging 1-2% over several years. The following year, however, saw a 6% decrease in scores. These fluctuations raised concerns by the Florida Department of Education and its constituents The Department of Education and its FCAT External Review Committee therefore asked the Buros Institute for Assessment Consultation and Outreach to investigate this matter. In our review over the past months, we have determined the following:

1. The Florida Comprehensive Assessment Test (FCAT) is a professionally developed test monitored by a high quality staff, and is produced by a nationally recognized testing firm.
2. Until 2006, there was a pattern of steady increases in test scores on both Reading and Mathematics over the grades assessed on the FCAT. In general, we believe that these test score gains are valid reflections of increased student learning.
3. The 2006 testing year clearly appears to be aberrant. By inspection alone, it is clear that the shape of its distribution differs from that of any of the other years. The scores earned that year do not fit the pattern seen in previous years and in 2007.
4. The jump in 2006 3rd grade reading test scores is in our opinion, a combination of real student growth and other factors. Changes in student demographics and their educational backgrounds are examples of factors that can lead to such changes.

We did not find obvious evidence for these factors in grade 3 over the two years in question that explain the magnitude of the changes in student proficiency.

5. In 2006, the anchor items in the test were not placed on the test where they had been in previous years. These items ultimately became easier than was expected. Placement of items can have a dramatic impact and it appears that this has almost assuredly occurred in this instance. We are firmly convinced that this explanation is by far the most likely explanation. While it is possible that some other yet unknown cause is involved, we do not know if any other changes between 2005 and 2006 occurred and we have no evidence of such instances. Moreover, we believe that this report should end the discussion of this issue.
6. Random sampling errors could also explain some variation in assessment results. Our evidence suggests that the size of the calibration sample needs to be increased to help reduce year-to-year fluctuations. The recommendation was also included in our first report, and the Florida Department of Education has accepted this recommendation.
7. While we believe the placement of anchor items is the without a doubt the most likely factor involved in the unusual change in proficiency rates for 3rd graders in 2006, other unknown factors cannot be ruled out completely[1]. These include our expectation that some real student growth in learning occurred and confounded the combination of student learning and the effect of changing the placement of the anchor items. It is perhaps possible that there were changes in the 3rd grade population not reflected in the rather complete data we reviewed, and a random

[1] Of course, one can never eliminate unknown factors, but what we are saying is that there is no known factor other than the placement of the anchor items that we believe can explain the difference in score in 2006.

equating error (a factor that is always present to greater or lesser extent and can only be contained by increasing sample size and making the calibration sample more representative). Moreover, given both that there is no appropriate formula to adjust the performance on the anchors for the change in their positions on the test and that the Florida Department of Education and its contractors searched repeatedly and unsuccessfully for other ways to equate the test to the previous years' forms, we strongly recommend against any rescoring of the 2006 data.

1. The Purpose of this Report: The FCAT Third Grade Reading Test Situation

The Buros Institute of Assessment Consultation and Outreach was contracted by the Florida Department of Education (FDOE) to conduct an audit of certain matters pertaining to the Florida Comprehensive Assessment Test (FCAT). Of primary concern was a finding related to the third grade reading FCAT, which is used in at least some settings as a promotional test and is therefore a high stakes test. In short, over early years in this decade, the number of individuals achieving proficiency status on the FCAT (that is, earning an achievement-level classification of 3, 4, or 5 on the 5-point scale), increased a percentage point or two each year. Then in 2006, this percentage increased dramatically by 8%. Unfortunately, the same percentage then decreased by 6% in 2007. This factor appears to have been the primary justification for hiring Buros to look into the FCAT. We believe that certain aspects of this situation are artifactual and others demand explanation.

We have worked hard to keep our discussions that follow as non-technical as possible so that they can be understood by the widest possible audience. The State of Florida has adopted demanding academic standards—the Sunshine State Standards—and has authorized a respected and well known contractor, Harcourt Assessment, to build the FCAT to assess whether those standards in reading and mathematics are being met instructionally in terms of student learning statewide. We have been impressed thus far with both the knowledge about educational testing and the level of commitment both to the Sunshine State Standards and to quality education by the members of the Florida

Department of Education with whom we have interacted as well as the members of their various technical advisory groups.

The specific goal of this report is to explicate the apparent rise and drop in the rates of achieving proficiency status on the FCAT third-grade reading scores. Section 2 provides a brief explanation of test form equating and the use of anchor items. This section continues by describing the important findings regarding equating and the use of anchor items in 2006, the factor that we believe appears most likely to have caused the large score changes. The next section describes year-to-year fluctuations in scores on the FCAT and other similar measures. We then summarize a simulation of year-to-year changes in passing rates in the context of overall test score changes. A more complete technical description of the study may be found in Appendix A. A fifth section very briefly addresses security concerns and recommends further investigation into test security. Finally, we provide a summary of the major conclusions we have reached and recommendations that we offer in regard to future administrations of the FCAT. These sections will be developed more fully in our third report.

2. Equating and Anchor Items

In many assessment situations, especially in settings where results are of high stakes and where tests are administered repeatedly, there is often the need to generate multiple test forms. New forms are needed because the high-stakes nature of the testing increases threats to test security and therefore items and test forms cannot generally be used more than once. In order for states such as Florida to gauge the annual success of their students, and hence, their educational programs, however, the scores from separate forms of the test (i.e., separate years) need to be placed on the same scale in order to make score comparisons meaningful.

Test forms are often built following the same design for test construction (e.g., test content, item format), and the difficulty level of the forms is intended to be the same. These forms rarely achieve perfect equality, however, and as a result, adjustments need to be made to make the scores from one "easier" form directly comparable to another "harder" form. This process of making adjustments to scores from one form is known as test equating.

Those of us who carry an excess pound or two know that scales differ; some are much friendlier than others and these rarely can be found in doctors' offices. Inexpensive bathroom scales often diverge by as much as 5-10 pounds on individual measurements, and such differences are neither acceptable for physicians nor psychometricians, the individuals who study psychological and educational tests. We are all familiar with the adjustments that one often needs to make to inexpensive bathroom scales to bring them into alignment with higher quality scales. Such adjustments are actually a rudimentary

form of scale equating. Requirements for proper test equating may be found as Appendix B and various types of equating are briefly described in Appendix C.

Equating Design. The equating design used for the FCAT is called the "common-item, nonequivalent groups design for equating" (Kolen, 2007). The procedure is probably the most commonly used technique in testing today because it does not require the administration of two complete forms of the examination to a single sample of test takers. Using this approach, Harcourt, together with input from FDOE assessment specialists prepares approximately 30 forms of the FCAT for FDOE each year, four of which serve as *anchor forms*. Within each of these anchor (reading) forms, there is one reading passage with roughly seven associated items that has been administered during the previous years (but not formally released). These seven items on each of the anchor forms are known as *anchor items*. Given that there are four so-called anchor forms, there is a total of 28 potential anchor items. These anchor items provide a link between forms that can be used to adjust the current year's FCAT scores so that these scores are directly comparable to scores from previous years.

Anchor Items in Test Equating. Anchor items obviously play a critical role in test equating; they provide the link that permits scores to hold the same meaning from year to year by providing a consistent metric across years. That is, scores on the anchor items, sometimes called common items as they are common across forms, are used to adjust for any differences in difficulty between the two test forms. For this reason, "the set of common items should be proportionally representative of the total test forms in content and statistical characteristics" (Kolen & Brennan, 2004, p. 19). It is also important that such items behave similarly on the different forms of the test. To help ensure this

similarity, the wording must be identical on both forms and they should be placed similarly on the test. "To ensure that the common items behave the same way on the two forms, each of the common items is identical on the two forms and is in a similar position in the test booklet" (Kolen, 2007, p. 46). It is crucial that all conditions of measurement for these common items are equivalent for both test forms.

Anchor items are always items that have been previously administered and calibrated. Anchor items can either be internal or external. Internal anchor items contribute to the score on a test as well as providing an anchor for equating; until about 2004 Florida used such internal anchor items. External anchors are seemingly part of the test, but do not contribute directly to a student's score; rather, they are solely used to help determine whether one form is easier or more difficult than another form, and to permit testing professionals to adjust and set scores so that they are equivalent across forms. Beginning in the 2004 FCAT administration, Florida has used external anchors. An advantage of external anchors, of course, is that they need not be released when the operational forms are issued publicly. Anchor items can be thought of as a stratified sample of test items generally chosen (1) to represent content of that form even more appropriately than a random sampling of questions would provide, (2) to embody a range of difficulty around the average difficulty of items from the previous form, and (3) not to be subject to specific causes of change, such as context effects where their item difficulty is affected by the items placed around them.

Examination of the 2006 and 2007 FCAT anchor items. The anchor items on the 2006 and 2007 FCAT 3rd grade reading assessments were not always in the same location as when these items were originally administered on earlier assessments. As noted

previously, changing the position of an anchor item on a test may change its level of difficulty (i.e., the percentage of test takers who answer it correctly) appreciably. Figure 3 shows that changes in anchor item positions are associated with changes in item difficulty in terms of the 2006 and 2007 FCAT reading tests. More specifically, the closer an item is moved toward the beginning of a test form, the larger the proportion of students who tend to answer the item correctly. The Pearson correlation between the change in an item's difficulty and its change in location is both meaningful and statistically significant ($r = -.576$, $p < .001$). One goal in designing anchor forms is to keep anchor items as close as possible to the positions the items occupied on the earlier assessment (the assessment to which the current assessment is being equated). Figure 3 also shows that the typical change in anchor item location in 2006 (19.4 positions earlier on average) was larger than that in 2007 (4.6 positions earlier on average). The difference in the average anchor item location change between 2006 and 2007 is also statistically significant ($t(47) = -4.845$, $p < .001$). Consequently the large improvement in the percentage of students scoring at the Proficiency level (achievement level 3 and above) is likely to be at least partially due to a decrease in the difficulty of the anchor items (as a consequence of the items being moved toward the beginning of the assessment), rather than differences in student ability. We believe that this finding, noted previously by FDOE, is perhaps one of the most significant findings in this report.

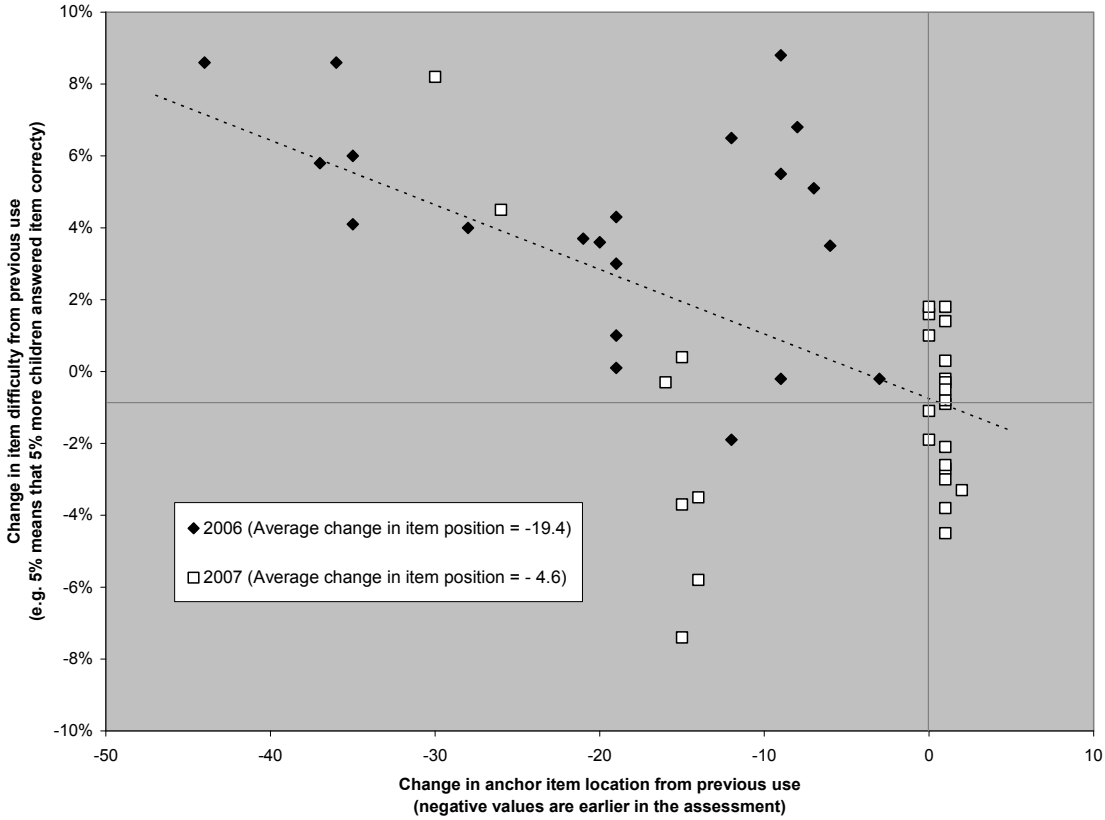

Figure 3. Change in anchor item location vs. change in item difficulty, FCAT 3rd grade reading 2006-2007

Equating and Validity. A discussion between Dr. Geisinger and several of the Florida state senators indicated a certain amount of confusion between two concepts: equating and validity. A description of each of these terms can be found in Appendix D.

3. The Context: Third Grade FCAT Reading Yearly Score Fluctuations and Cohort Attributes in 2006 and 2007

In 2006 the percentage of Florida third graders scoring at proficient level and above on the FCAT increased by about 8% to 75%. In 2007 that same percentage decreased by 6% to 69% (see Table 1). While it appears that some have assumed that such large changes could not be the result of chance, it should be pointed out that such a change is not necessarily impossible, nor is extremely unusual.

An 8% increase in proficiency rates is relatively large compared to other years and grade levels. However, it is not an unprecedented change. For example, increases that are as large or larger for reading were observed in grades 6 and 7 in 2006, and grade 4 in 2004. In 2004 there was also a 10% increase in grade 4 mathematics. In 2007 the percentage of third graders who scored at achievement level 3 or higher on the FCAT Reading assessment dropped by 6% to 69%. While this 6% drop was the largest percentage drop in reading or math from 2001 to 2007, drops of 5% and 4% have also been observed, and a drop of 6%, while extreme, does not seem to fall outside the typical distribution of change values. A histogram showing this distribution is presented in Figure 1.

Table 1. FCAT 3rd grade reading scores, 2001-2007

Year	% of students at or above proficient
2001	57%
2002	60%
2003	63%
2004	66%
2005	67%
2006	75%
2007	69%

It should also be noted that it is not necessarily unusual for assessment scores to increase one year, only to decrease the next. Tables 2 to 4 present three examples of this phenomenon including a different Florida assessment, as well as assessments in other states in the region. In each of these situations (including 3rd grade FCAT reading), there

is a generally increasing trend in the percentage of proficient students over time (Figure 1 shows that the percent of students achieving proficiency across grades and subjects has typically been increasing a percentage point or so each year). The year-to-year increase-decrease cycles may be due to a variety of factors, perhaps including sampling errors and regression to the mean[2]. That is, there may be an underlying rate of improvement combined with year-to-year fluctuations based on various factors other than changes in ability alone.

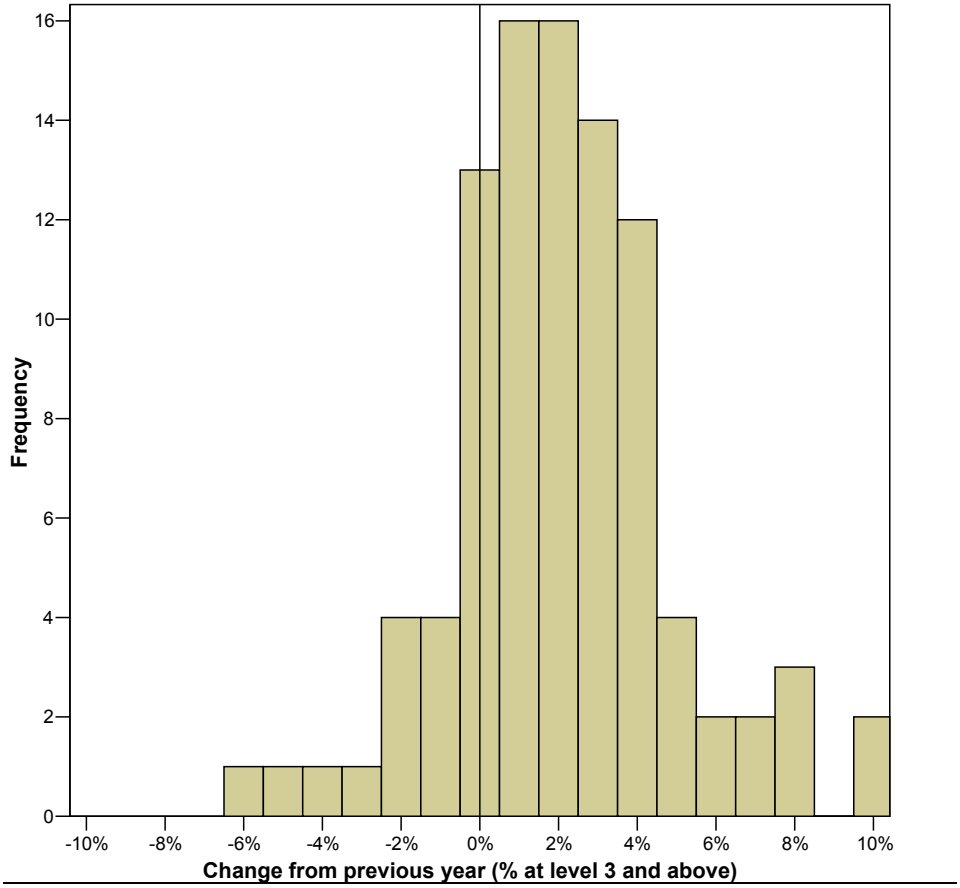

Figure 1. Histogram of year-to-year changes in % of students at or above achievement level 3 for the FCAT mathematics and reading assessments, 2001-2007. Source: Florida Department of Education, May 2007

[2] Regression to the mean is a complicated issue that affects many research designs. In general, what is meant by regression to the mean is that when one encounters an extreme score on a test, the individual who earned that score is likely to score high or low on a second, similar testing as the case may be, but in a less extreme manner. In the same way, a state that scores extremely well (highly) on a given testing would likely fall just a little on a second testing.

Table 2. NAEP Reading results for 4th grade students in Florida public schools

Year	% of students at or above proficient
2002	27%
2003	32%
2004	-
2005	30%
2006	-
2007	34%

Note: Observed differences are not necessarily statistically significant

Table 3. CRCT Reading/Language Arts assessment results for students in Georgia, 2002-2007

Year	% of students at or above proficient
2002	-
2003	78%
2004	84%
2005	87%
2006	85%
2007	87%

Table 4. Mississippi Curriculum Test (Reading) results for 3rd grade student

Year	% of students at or above proficient
2002	79%
2003	81%
2004	84%
2005	84%
2006	87%
2007	84%

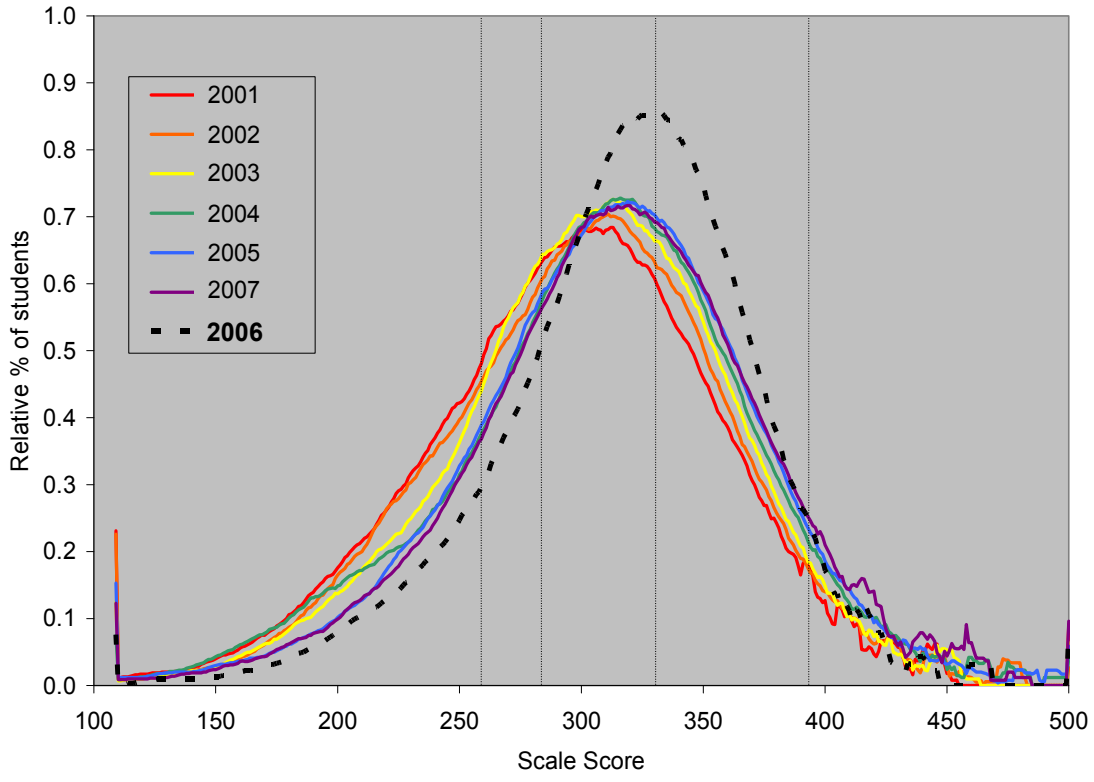

Figure 2. Smoothed relative frequency distributions of 3rd grade FCAT reading scores, 2001-2007

However, like members of the Florida Department of Education, the FCAT External Review Committee, and others, we do believe that the scores on the 2006 3rd Grade FCAT reading assessment deserve extra scrutiny. One can see in Figure 2 (which shows the percentage of students at each FCAT score point, averaged over 10-point score ranges) that the scores for 2006 are generally higher and less spread out than in other years. The curve differs from that of all other years and, we believe, is indicative of a sort of aberrancy that demands further explication.

What might account for these differences? First, one should consider the possibility that this finding may be the result at least in part of real changes in the reading ability of students. Changes in student demographics and educational backgrounds are also examples of factors that can lead to such changes, since each group of 3rd graders differs from that of the previous year. However, student race, gender, and educational background variables were examined and we did not find obvious evidence for these factors over the two years in question that explain the magnitude of the changes in grade 3 student proficiency.

In this section, we have shown how year-to-year fluctuations in proficiency rates are not necessarily unheard of, nor are they unique to the FCAT. However, we have also noted that grade 3 FCAT scores in 2006 appear to deviate from the pattern both before and after this year. We've noted that this deviation may be due to real changes in student performance, or due to differences in the 3rd grade population over time (though we did not find obvious evidence for this). In other sections of this report, we examine how characteristics of the assessment and of the equating procedures could possibly have contributed to these fluctuations as well.

4. The Effect of Random Equating Error on Measuring Improvement

There are two types of errors that must be minimized for the equating procedure to produce reasonably comparable scores between groups: Random error and systematic error. To examine the effect of random and systematic equating errors on the proportion of students scoring at or above proficiency for the Grade FCAT, we performed two separate Monte Carlo[3] simulation studies. The first Monte Carlo simulation study examined the variability in estimated improvement assuming that the percentage of students classified as proficient increased 2% from the base year to the following year. In other words, given a true improvement rate (e.g., 2%), how often we would *observe* improvement other than the true improvement rate due to random error?

The results indicated that there was a surprising amount of variability in observed improvement rates over the simulated samples from a population where the true improvement is 2%. The average amount of improvement seen was 2%, implying that, on average, the equating procedure will correctly calculate the percentage of students at or above proficient. The standard deviation of percent improvement (i.e., average "spread"), however, was 0.043, indicating that even though the expected amount of improvement represents the true improvement of 2%, it is not uncommon to observe improvement rates much different than 2%. To further illustrate this point, Table 1 reports the proportion of samples exhibiting specific amounts of improvement.

---

[3] [3] Monte Carlo simulation studies are computer simulations of actual situations. An advantage of such studies is that one can run multiple simulations based on randomly equivalent data and determine the range, nature and types of results that are likely to result. In this case, 100 samples of simulated data were randomly assembled.

Table 1. *Percentage of samples exhibiting an increase or decrease in improvement (Note: The true improvement was*

Improvement Rates	Proportion of Samples
< -8%	0.00
-8% to -6%	0.03
-6% to -4%	0.09
-4% to -2%	0.06
-2% to 0%	0.10
0% to 2%	0.20
2% to 4%	0.17
4% to 6%	0.18
6% to 8%	0.06
> 8%	0.11

It is apparent from Table 1 that while a large number of samples exhibited reasonable improvement rates compared to the true improvement rate of 2% (e.g., 37% of the samples exhibited improvement within 0% to 4%), there were a considerable number of samples exhibiting improvement well beyond the true improvement rate of 2%. For example, 17% of the samples exhibited improvement greater than 6%. Interestingly, a non-ignorable amount of the samples exhibited declines in performance (e.g., 28% of the samples exhibited a decrease in the number of students scoring at or above proficient compared to the previous administration). This result illustrates that the amount of random error in the equating may have a large impact on the proportion of students being classified into the performance categories, making it difficult to judge improvement over years.

While random error represents noise, systematic error can lead to a consistent over (or under) prediction of improvement. Systematic changes in anchor item parameter values over time (i.e., item parameter drift) or changes in the representativeness of the calibration sample can both be sources of systematic error. In the second simulation study we examined the effect of having an unrepresentative calibration sample on accuracy of the equating results. Specifically, the calibration sample was assumed to improve by 4% while the rest of the population was assumed to grow at a rate of 2%. The results from this study were similar to the first study with respect to random error, but the entire

distribution was systematically shifted up, indicating that the student's scores (and the percentage of students at or above proficiency) were over-predicted due to systematic equating error. These finding suggest that the representativeness of the calibration sample plays a major role in providing accurate estimates of year to year improvement.

The details of both simulation studies (methodology, computer software, etc.) can be found in Appendix A.

5. Test Security Concerns

One additional factor must be considered, the security of the anchor items (as well as the operational items themselves). Operational items on the FCAT have been pretested typically during the previous year's testing, whereas anchor items have either been used as anchors during the previous year(s) or have been pretested previously. While very few students would be likely to have seen any of these items on the FCAT, it is possible that educators have seen them and have shared them with students in preparing those students for the test. Please note that in stating this point, we are not suggesting anything nefarious. Rather, teachers are under significant pressure to help their students to do well. If they are able to do so, they may look over the test so that they are familiarized with the typical test content so that they can plan their instruction accordingly. In so doing, they may provide instruction on specific items that they have seen to help make this instruction relevant to students and to provide them with realistic expectations in terms of what the test includes.

To assure the State of Florida that such actions have not occurred and will not occur in the future, we recommend that the State consider working with a company that specializes in test security services. We understand that the FDOE is open to this

suggestion. We believe that being assured that the test items are secure would be helpful to the State of Florida in understanding the reasons for test score changes, if any.

6. Conclusions

Identifying what happened in a situation after the fact is almost always extremely difficult. Nevertheless, in this case we have a few conclusions.

1. The Florida Comprehensive Assessment Test (FCAT) is a professionally developed test. The individuals involved at the Florida Department of Education have excellent reputations, and we have been impressed with them in our interactions. The companies working on these measures are well known in the industry.
2. Until 2007, there is a pattern of steady increases in test scores on both Reading and Mathematics over the grades assessed on the FCAT. In general, we believe that these test score gains are valid reflections of increased student learning.
3. The 2006 testing year clearly appears to be aberrant. By inspection alone, it is apparent that the shape of its distribution differs from that of any of the other years. The scores earned that year do not fit the pattern seen in previous years and in 2007.
4. The anchor items in the test were not placed on the test where they had been in previous years and these items ultimately became easier than was expected. We believe they were easier because of the location of their placement on the test form. Placement of items can have a dramatic impact and it appears that this may have occurred in this instance. Because we were not present at and throughout

the times of the test administration, we do not know if any other changes between 2005 and 2006 occurred, but we have no evidence of such instances.

5. If 2006 data are eliminated, and 2005 and 2007 seen side-by-side, when evaluating the results of the 2007 administration, then the 2007 year continues to represent improvement of student performance.
6. The generally accepted score for achieving proficiency on the 3rd grade Reading FCAT is 284. This value is rather centrally located in the distribution of test takers, with approximately 72% of the students found as proficient in recent years. Any changes in testing practices can impact that passing score to a great degree, as was seen in 2006. It is a distinct strength of the FCAT program that proficiency scores have been consistent year after year. Otherwise, comparisons such as those presented in this report would be far more difficult both to perform and to report.
7. The jump in 2006 3rd grade reading test scores is, in our opinion, a combination of real student growth and other factors. The most likely factors that we would include among these are:
    a. The change in location of the anchor items. This change leads us to be substantially concerned. As noted by the brief literature review from highly respected sources, position effects can certainly have a major impact, as they appear to have on the 3rd grade FCAT. Nevertheless, we cannot say with 100% certainty that this was the cause in this instance. It is our most likely suspect, however, and the position of anchors needs to be consistent.

b. Random error effects. As noted both by historical data and our simulation study, it is clear that random error effects alone could explain some or all of the changes in year-to-year passing rates, even those that appear as highly unusual fluctuations.

c. The security of the items. As also noted below, any knowledge of items on the FCAT forms—and most certainly of the anchor items—would make them appear much easier during the second or later year (2006) of their administration. We believe that Florida should take proactive steps to ensure that the items on the test continue to be as secure as possible, even after they have been administered. We note, however, that we have no direct evidence of inappropriate actions leading to security concerns.

7. Recommendations

1. The performance of the educational system should not be judged by proficiency rates alone. Average test scores are probably a better index, and should at least be used in conjunction with proficiency rates, as they reflect the performance of all students to a greater extent. Multiple indices are almost always best. While we believe that the FCAT should be employed as a critical "dashboard indicator" it should not be the only one used to evaluate the schools in Florida.
2. Because of year-to-year fluctuations in test scores due simply to equating and other random factors, it may be prudent to consider multiple years of data (e.g., average proficiency rates over the three most recent years) when making decisions regarding accountability sanctions or rewards (e.g., teacher or administrator merit pay).

3. It appears extremely likely that the aberrant scores in 2006 have resulted from the placement of anchors on that test. Therefore, Florida needs to adopt and strictly follow rules about the placement of such items. The policy of another state with which we are familiar requires that the position of items not change by more than 10 places from one year (or form) to the next. Even such policies are concerning because it means that the position of an item could conceivably change by 30 places over a 3-year period. A suggestion that anchors should always be in a given location, however, is itself not without problem. If test anchors are always the questions associated with the second reading on a test, then students and teachers would learn this fact quickly. Procedural rules for the placement of anchor items are needed and they must be followed.
4. Anchor items need to be retired regularly. Although there is no generally agreed upon standard in the profession in this regard, we suggest that retiring all anchor items after they have been used three times probably makes sense to avoid heightened item exposure.
5. The State of Florida should consider making the calibration sample representative of all students, not just the standard curriculum students. We look forward to addressing this issue as our work continues.
6. Although we dislike recommending a specific company, we do recommend that the Florida Department of Education contract with a service provider to investigate any allegations of inappropriate testing behaviors. While such behaviors might include those generally considered to be cheating, we believe that more typical behaviors are those where teachers share their knowledge of specific

items from previous years' examinations toward the goal of making their instruction both more relevant and effective. The FDOE has issued, we understand, strong recommendations against such practices, but a test security, could be able to provide an evaluation of whether these potential risks have actually occurred or not.

8. Future Directions

1. Scores must be released in a timely manner. This pressure is clearly one of the stressors for test developers and states alike. We believe that Florida should scrutinize test scores carefully before they are released to ascertain that no errors or potential errors in equating have been made, prior to scores being released. Once released, perceptions are hard to change. Procedures probably need to be developed or enhanced to increase the review of scaling issues prior to score release. We would like to work with the FCAT External Advisory Committee address the nature of indicators that could be used to postpone the reporting of scores in future reports.
2. Not being privy to decisions that Harcourt must make and their discussions in regard to such questions, we question whether 26 of the 30 forms are needed to pre-test possible new items. We believe that the number should be re-evaluated with the possibility that more students would take the equating-anchor forms of the test. While the State of Florida is presently using 15-28 equating items for 45 operational-core questions on the 3rd grade reading test, only approximately 5% of the students take the equating items. Increasing the number of students taking the equating forms, while keeping the representativeness at least equal to its present

status, could reduce equating concerns. Another option is to pretest reading sections with a larger number of test items associated with each one. Then if one or two prospective test items are dropped due to poor performance, it is possible that the entire section would not need to be dropped.

3. Discussions among FDOE and Florida legislators indicate that the State would very much like to use tests that are diagnostic in intent. We believe that the tests used to monitor academic attainment are generally not well suited to such purposes and probably cannot be adapted to be such measures. Nevertheless, the State, committed to helping all learners achieve, should consider whether it wishes to add a more diagnostic measure. As noted in our prior report, the content clusters are not well suited for individual student diagnosis, although they might be able to be used to track the progress of the state or of individual school districts. We also believe that they may have import in making curricular decisions.

References and Sources

Dorans, N. J., Pommerick, M. & Holland, P. W. (2007). *Linking and aligning scores and scales*. New York: Springer.

Holland, P. W. (2007). A framework and history for score linking. In Dorans, N. J., Pommerick, M. & Holland, P. W. (2007). *Linking and aligning scores and scales* (pp. 5-30). New York: Springer.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Westport, CT: Praeger.

Kolen, M. J. (2007). Data collection designs and linking procedures. In Dorans, N. J., Pommerick, M. & Holland, P. W. (2007). *Linking and aligning scores and scales* (pp. 31-55). New York: Springer.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices (2nd ed.).* New York: Springer-Verlag.

Michaelides, M. & Haertel, E. (2004). Sampling of common items: An unrecognized source of error in test equating. CSE Report 636. Los Angeles: CRESST, University of California at Los Angeles

Nunnally, J. C. (1978). Psychometric theory (2nd. Ed.) New York: McGraw-Hill.

Peterson, N. S. (2007). Equating: Best practices and challenges to best practices. In Dorans, N. J., Pommerick, M. & Holland, P. W. (2007). *Linking and aligning scores and scales* (pp. 59-72). New York: Springer.

Peterson, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*, 137-156.

Puhan, G. (In press). Scale drift in test equating on a test that employs cut scores. *Applied Measurement in Education*.

Skaggs, G. & Lissitz, R. W. (1986). An exploration of the robustness of four test equating models. *Applied Psychological Measurement, 10*(3), 303-317.

BUROS
Center for Testing

Report on the
2006 and 2007 Florida Comprehensive Assessment Test Grade 3 Reading Scores:

Prepared for the

Florida Department of Education

and the

Florida FCAT External Review Committee

Appendices

Appendix A: Simulation Study on the Effect of Equating Error on Measuring Improvement

Equating is a crucial aspect of maintaining a stable score scale over time and measuring improvement. In fact, improvement (e.g., measured by comparing the proportion of examinees scoring at or above proficient over consecutive years) is primarily determined by the performance on the anchor items while the scoring items are used to increase precision of the proficiency estimates. Any error in the equating may have a detrimental effect on the measure of improvement. There are two types of errors that must be minimized for the equating to produce reasonably comparable scores between groups or over time: Random error and systematic error. We examined the effect of random and systematic equating error on measuring improvement for the Grade 3, Florida Comprehensive Assessment Test, as determined by the proportion of students scoring at or above proficiency.

Effect of Random Equating Error. Random error represents noise in the data that may make it difficult to observe the effect of interest (e.g., improvement). The effect of random error may be simply conceptualized as the variability in the estimates (e.g., item parameter estimates, test scores, proportion of students at or above proficient, etc.) over multiple samples drawn from the same population. When measuring any educational or psychological construct, it is crucial to minimize the amount of random (and systematic) error in order to draw valid inferences from the results. The important question in this context is, "How much does random error influence the measure of improvement?" More specifically, "Can random error help explain, in part, the 8% increase in students scoring at or above proficient from 2005 to 2006 on the FCAT, Grade 3 Reading assessment?"

Examining random equating error in the FCAT is particularly relevant considering that the parameter estimates for the anchor items are based on roughly 2,000 to 2,700 examinees. Therefore, the amount of error in the item parameter estimates may influence the equating significantly.

To explore the potential effect of random equating error, we performed a Monte Carlo simulation study to examine the variability in the proportion of students scoring at or above proficiency given an actual 2% increase from the previous administration. In other words, given a true rate of improvement (e.g., 2%), the simulation study examined how often we would an *observe* rate of improvement other than the true improvement rate. For example, how often would we observe improvement of 8% or higher given the current test specifications, equating design, other operational specifications, and a true improvement rate of 2%?

The simulation study was intended to replicate the essential aspects of the operational procedure used in scaling the FCAT, Grade 3, Reading assessment given in 2006. A brief description of the simulation will be provided below, followed by the results.

Simulation Conditions: Base Year. Dichotomous item responses for 21 items, which correspond to the anchor items used in the 2006 administration of the Grade 3 Reading assessment, were simulated from the three-parameter logistic model (3PLM) to represent a base year. The base year was used to define the score scale and provide item parameters estimates for the anchor items used in 2006 administration. The generating item parameter values were based on the anchor item parameter estimates for the 2006 Administration reported on page 117 in Appendix A of the Reading and Mathematics

Technical Report for 2006 FCAT Test Administration. Ability parameter values were sampled for 5,000 examinees from the standard normal distribution and a cutscore of -0.41 was chosen because it would produce roughly 65% at or above proficient. The item parameters were estimated using the software package PARSCALE.

Simulation Conditions: Calibration Sample. Dichotomous data were generated to represent the 2006 FCAT, Grade 3, Reading administration. The item parameter estimates from the 3PLM provided in Appendix A (pp. 116-117) of the Reading and Mathematics Technical Report for 2006 FCAT Test Administration were used as generating item parameter values to simulate the dichotomous item responses. To represent the Calibration Sample, item responses for 8,100 examinees were simulated to represent the three test forms used in equating (i.e., F27, F28, and F29) containing 8, 5, and 8 anchor items, respectively. All 8,100 examinees were used to generate data for the 45 scoring items while the sample was split into three equal sizes to generate responses for the anchor items (i.e., N=2,700 per form). Examinees were sampled from a normal ability distribution with $\mu = 0.058$ producing a roughly expected 67% at or above proficient. The item parameters were calibrated concurrently using PARSCALE. The anchor items were used to obtain the linking coefficients between the current administration and the base year. The linking coefficients were used to place the item parameter estimates for the scoring items onto the base year scale.

Simulation Conditions: State Sample. To represent the remaining students who take the FCAT, item responses for 180,000 examinees were generated for the scoring items only. Examinees were sampled from an ability distribution with $\mu = 0.058$ producing a roughly expected 67% at or above proficient. The transformed item

parameter estimates from the Calibration Sample were used to estimate proficiency scores for each examinee using PARSCALE. Lastly, the proportion of examinees at or above proficient was determined by comparing the proficiency estimates to the respective cutscore of -0.41.

The procedure was repeated 100 times (i.e., 100 samples) in order to examine the sampling distribution of the proportion of examinees being classified as proficient or above.

Simulation Results. The results from the simulation study indicated that there was a meaningful amount of variability regarding observed improvement over samples from a population where the true improvement is 2%. Figure 1 provides the histogram representing the proportion of students scoring at or above proficient across the 100 simulated samples.

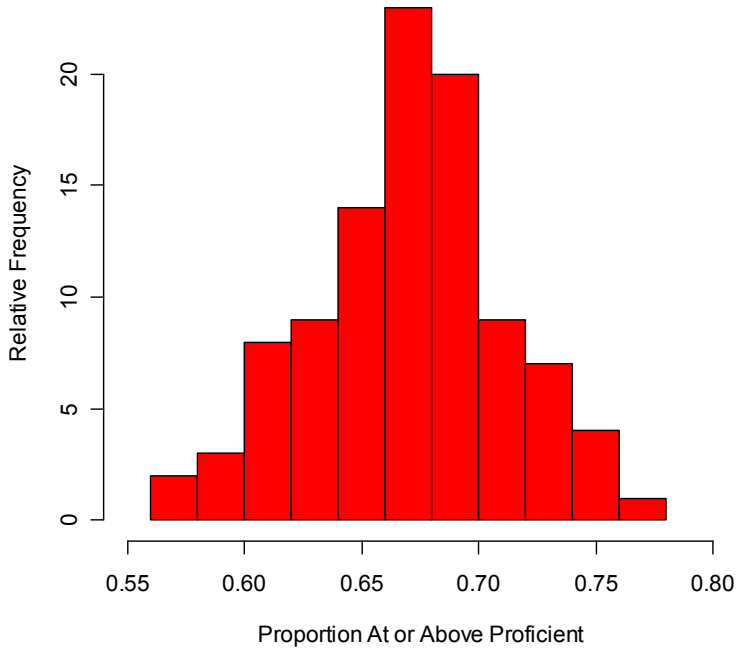

Figure 1. Histogram representing the proportion of students scoring at or above proficient (Note: The previous administration observed 65% at or above proficient).

The average proportion of students across the 100 samples who scored at or above proficient was 0.67 (representing 2% improvement above the base year where 0.65 of the students were at or above proficient). However, the amount of variability across samples, as measured by the standard deviation, was 0.043, indicating that even though the expected amount of improvement represents the true improvement of 2%, it is not uncommon to observe improvement rates much different than 2%. To further illustrate this point, Table 1 reports the proportion of samples exhibiting specific amounts of improvement.

Table 1. Percentage of samples exhibiting an increase or decrease in improvement (Note: The true improvement was +2%).

Improvement Rates	Proportion of Samples
< -8%	0.00
-8% to -6%	0.03
-6% to -4%	0.09
-4% to -2%	0.06
-2% to 0%	0.10
0% to 2%	0.20
2% to 4%	0.17
4% to 6%	0.18
6% to 8%	0.06
> 8%	0.11

It is apparent from Table 1 that while a large number of samples exhibited reasonable improvement rates compared to the true improvement rate of 2% (e.g., 37% of the samples exhibited improvement within 0% to 4%), there were a considerable number of samples exhibiting improvement well beyond the true improvement rate of 2%. For example, 17% of the samples exhibited improvement greater than 6%. Interestingly, a non-ignorable amount of the samples exhibited reduced proportions (e.g., 28% of the samples exhibited a decrease in the amount of students scoring at or above proficient compared to the previous administration). This result illustrates that the amount of random error in the equating may have a large impact on the proportion of students being

classified into the performance categories making it difficult to judge improvement over years.

Effect of Systematic Equating Error. While random error represents noise, systematic error represents biased estimates of the true signal (e.g., an estimate that consistently over (or under) predicts true improvement). There are a few factors that may lead to systematic equating error. For example, anchor item parameter values that have changed over time (i.e., item parameter drift), especially in a particular direction, may lead to a consistent overestimate of improvement due to the effect on the linking function. In the FCAT, the potential effect of several sources of systematic error is reduced by performing a thorough examination of the appropriateness of the item response theory (IRT) model (e.g., differential item functioning, item parameter drift, model misfit). However, the equating design used in the FCAT may be susceptible to systematic equating error.

In order to increase the turn-around time for reporting student test scores, a Calibration Sample consisting of roughly 8,000 examinees is used to estimate the item parameters for the scoring and anchor items. The anchor items are distributed onto four forms, resulting in about 2,000 examinees answering a particular item (however, note that the FCAT, Grade 3, Reading assessment uses three forms; therefore, approximately 2,700 examinees respond to each anchor item). The estimates for the anchor items are then used to determine the linking coefficients between the current administration and the base year. The linking coefficients are used to transform the item parameter estimates for the scoring items onto the base year scale. The transformed scoring item parameter estimates are then used to score the examinees for the entire state.

Theoretically, this procedure will not produce systematic error as long as the Calibration Sample is representative of the entire population (i.e., the linking relationship will be appropriate). However, if the Calibration Sample does not represent the entire population, the equating relationship between the current administration and the base year may systematically under or over predict improvement. While FCAT expends a great deal of effort to implement a sophisticated methodology to select schools for the Calibration Sample, inevitably, the representativeness of the Calibration Sample will be questionable in some years and some assessments (it is important to note that although the Calibration Sample will never represent the entire population *exactly*, it will be close-enough to accomplish its goal of determining an appropriate equating relationship between the current administration and the base year).

Since improvement is primarily captured in the anchor items, the Calibration Sample is essential in determining improvement. To illustrate this point, we performed a Monte Carlo simulation study to explore the effect of systematic equating error on examinee proficiency classification. The same procedure for generating the data was followed as described in the section examining the effect of random error except that examinees from the Calibration Sample were drawn from an ability distribution in which 69% were at or above proficient ($\mu = 0.114$), representing a 4% improvement rate compared to the base year. However, the ability parameter values for the entire state population were drawn from a distribution in which 67% were at or above proficient. Therefore, the Calibration Sample is not representative of the entire state. It is interesting to note that the mean for the Calibration Sample ($\mu = 0.114$) need only differ slightly from that of the entire state ($\mu = 0.058$) to produce an improvement rate of 4% versus 2%.

Figure 2 illustrates the histogram representing the proportion of students scoring at or above proficient across the 100 replications (samples). The average proportion of students across the 100 samples who scored at or above proficient was 0.69 representing a 4% improvement rate even though the true improvement rate was 2%. There are a few interesting observations from this result. First, improvement is primarily determined by

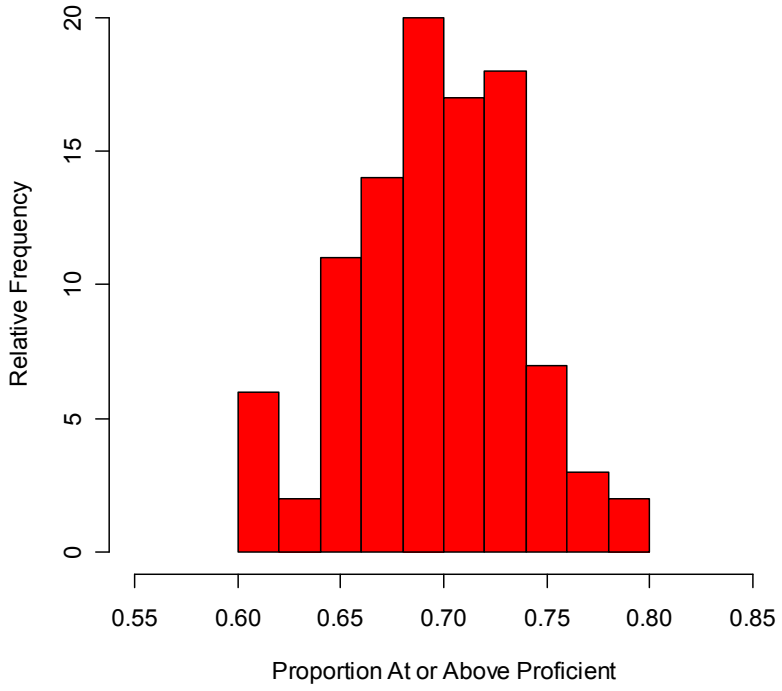

Figure 2. Histogram representing the proportion of students scoring at or above proficient (Note: The previous administration observed 65% at or above proficient).

the Calibration Sample even though the Calibration Sample only comprises about 4% of the student population. Second, if the proficiency distribution of the Calibration Sample is slightly different from the state population, then the improvement rate will likely be over or under predicted. Third, random error still plays an important role when the Calibration Sample is not representative of the student population (the standard deviation of the proportions across the 100 simulated samples was 0.040). To illustrate this point

Table 2. Percentage of samples exhibiting an increase or decrease in improvement (Note: The true improvement was +2%).

Improvement Rates	Proportion of Samples
< -8%	0.00
-8% to -6%	0.00
-6% to -4%	0.02
-4% to -2%	0.04
-2% to 0%	0.07
0% to 2%	0.13
2% to 4%	0.11
4% to 6%	0.25
6% to 8%	0.21
> 8%	0.17

further, Table 2 reports the proportion of samples exhibiting specific amounts of improvement. Given the systematic over prediction of improvement due to the non-representative Calibration Sample and random error, it is not uncommon to observe large (false) improvement rates. For example, 38% of the 100 simulated samples produced observed improvement rates of 6% or higher (17% exhibited 8% improvement or higher).

Factors Influencing Random and Systematic Equating Error. There are several factors that may introduce systematic or random error into the equating. In general, random error will be greater when the information (i.e., item parameter estimates) used in the equating contains a relatively large amount of error. In this case, the item parameter estimates from the three-parameter logistic model (3PLM) are used in the equating and the amount of error in the estimates may be measured by their respective standard errors. Given the equating design in which only 2,000 to 2,700 examinees respond to each anchor item, the standard errors for the item parameter estimates may be larger than desired. Therefore, it is advisable to have at least 5,000 examinees respond to each anchor item (this is a common sample size used in statewide assessments). While sample size per anchor item is an important factor influencing error in the item parameter estimates, test design also affects item parameter estimation. Tests that contain item parameter values that are matched to the proficiency distribution produce less error in the

estimates. Therefore, ensuring that the test characteristics match the ability distribution for each assessment will reduce the effect of random error on the equating. In addition to the test characteristics matching the proficiency distribution, it is prudent to use at least moderately discriminating items (a test with primarily low discriminating items will have more equating error). A fourth crucial consideration is the number of anchor items used in the equating. As the number of anchor items increases the random error in the equating tends to decrease. Considering the previous factors will minimize random error in the equating.

In the FCAT, systematic error will be a problem when the proficiency distribution for the Calibration Sample does not match the proficiency distribution for the entire state population. The ability distribution may vary with respect to location (e.g., mean, median), spread (e.g., standard deviation), and/or shape (e.g., normal, skewed) (in this report, we only examined differences in location; however, the other two factors could systematically alter the equating as well). All three factors could systematically influence the equating. One solution to this problem is not to use a subset of the population (i.e., Calibration Sample) to link the current year's scale to the base scale. However, considering FCAT's goal of reporting the test scores quickly, using all students may not be possible.

Systematic error may also be introduced when the item parameter values have changed from the base year to the current administration (i.e., item parameter drift). The FCAT operation procedure addresses this type of error by thoroughly examining the parameter estimates each anchor item to determine if its value has changed from the

previous administration. Any items that are deemed to have drifted from their original values are excluded from the equating.

Appendix B: The Requirements of Equating

There are five stringent requirements that should be met in order to apply any equating method properly (Holland, 2007; Holland & Dorans, 2006; von Davier, Holland, & Thayer, 2004):

1. Each test form should measure the same construct. (The term, *construct*, is the formal term used by psychologists to designate a theoretical entity, attribute or quality whose corresponding characteristics, a test is measuring.)
2. The forms should be highly and equally reliable. (Reliability is a formal psychological term that means that the measures consistently find the same or similar scores for similar levels of performance.)
3. The equating function should be symmetric so that equating the scores of test form **X** to form **Y** produces the same results as equating **Y** to **X**.
4. The equated score obtained by an individual should be equivalent regardless of whether they were administered form **Y** or **X**.
5. The equating function should be population invariant – this means if different samples of people are used to compute the equating function between the test scores on two forms, the use of these different samples should not make a difference.

Holland and Dorans (2006) expanded upon the first criterion above that each test should measure the same construct. Not only must the test items be similar in content, but also they should be of the same format (i.e., multiple-choice or open-ended questions, in the case of educational achievement measures). Holland and Dorans also argued that the consequences of the test should be equivalent – because the consequences—good and

bad—of a test affects the test takers and hence the construct measured by the test forms. In addition, tests should be administered under secure and standardized conditions, and examinees used for equating purposes should be representative of the population of test takers for which the equating will be applied. The issues of English language learners and of students who receive accommodations when taking a test has rarely if ever been considered in equating, but minimally, wisdom suggests that the two samples used in the equating study (that is, for forms X and Y) need to have comparable portions of all relevant groups.

Appendix C: A Brief Description of Some Equating Methods

Equating methods for dichotomous items include mean, linear, equipercentile, and item response theory (IRT) techniques. Mean, linear, and equipercentile methods fall under the category of *observed-score test equating.* While mean equating is presented simply to help clarify the concept of equating, linear and equipercentile represent traditional forms of test score equating using observed scores. Observed scores are simply the actual scores that result from a testing. In observed score equating, some characteristics of the score distributions are set to be equal for a specific population of examinees (Kolen & Brennan, 2004). More modern methods often use the item response theory approach, which generally involves *true score equating*, and which is the method that the State of Florida is using in equating the FCAT. An assumption of most test theories is that each individual's score is based upon his or her true score[4] and a certain amount of what we think of as random error, e.g., lucky guesses on multiple-choice test questions. Classical methods do not directly consider true scores and have been the most commonly used methods likely because they predate other equating methods (Kolen & Brennan, 2004). Item response theory (or IRT) methods employ techniques that permit us to estimate true scores.

Item response theory (IRT) methods have also been developed in which a mathematical relationship between scores on two tests is modeled. The relationship is based on item parameters from each test and results in the placement of the score

[4] A true score may be thought of as an average of many testings for an individual, a testing of an individual by all items in the domain, or some idealized score that truly represents a person's capability in a given domain of content. Ultimately, it is a score that has no error component at all and truly measures the ability. It is generally accepted that the measurement of true scores is not possible in educational and psychological measurement (Nunnally, 1978). An actual test that is highly reliable, however, provides scores where observed scores more closely approximate true scores.

estimates of the same metric (Skaggs & Lissitz, 1986). These methods require assumptions of unidimensionality and local independence to be met in order to be properly implemented. The unidimensionality assumption requires that the test measure only one latent construct, with only one underlying ability distribution (as opposed to multiple abilities impacting the distribution). Local independence means that the responses to items on an exam are statistically independent after taking into account examinee ability (Kolen & Brennan, 2004). For example, answering item 1 correctly should not increase the probability of correctly answering item 2.

Appendix D: Equating and Validity

Equating was described in a certain amount of non-technical detail in the beginning of this section in this report. In short, it involves making scores on different forms of the same general test to be as comparable as possible. It is a matter related purely to scoring and has many assumptions. When equating is effectively performed, one can use scores from different forms of the same test, such as the FCAT, interchangeably. For example, scores from the 2005 third-grade reading FCAT test should mean the same as scores from the 2006 FCAT.

Validity is generally considered to be another thing altogether. Validity is the most important consideration related to a test for testing professionals. Validity relates to whether a test is measuring what it intends to measure. While a discussion of validity is certainly beyond the scope of the current report, it is possible to describe test validity as it impacts the FCAT. For tests of educational achievement such as school learning, intended test score interpretations are in the main considered to be valid if the test represents the material that is intended to be covered. For example, the use of a classroom test developed by an individual teacher would be considered valid if it systematically covered the material taught in the instructional unit (this concept is known both as content validity and instructional validity). A test that did not systematically cover the unit might nevertheless assess knowledge learned during instruction but its usage might be seen as less than optimally valid. For example, imagine a course final examination that consists solely of a single essay question that covered material presented in only a single chapter of the book. Such an examination might have a certain amount of validity, but would not generally be accepted as a good and valid measure of learning

throughout the semester or year. Tests should represent the entire domain of learning in a representative manner. In order for tests that measure standards such as the Sunshine State Standards to be used validly, the tests should measure the standards identified by educational leaders in proportions representative of their importance to be valid. Alignment studies are performed by many states according to the judgments of the state's educators to demonstrate that the items composing the state tests are seen as actually measuring content associated with the state standards for those grades.

Relationship between FCAT Equating and Validity. The most important issue in terms of establishing the validity of the FCAT is that it measures content related to the reading and mathematics standards in the Sunshine State Standards. A related issue is that the standards actually influence both the curriculum of the State of Florida and what is taught in classrooms; these are the issues related to the quasi-legal terms curricular and instructional validity, respectively. We have neither been privy to results of alignment studies that may have been performed to assess these correspondences nor would review of such an analysis be within the scope of our current charge, although data indicating the general increase in test scores over time suggest that both of these types of validity are increasingly present. Nevertheless, one aspect of validity is clearly within our charge. This aspect relates to the meaningfulness of individual scores. One legitimate use of the scores from FCAT relates to use of the scores to assess the functioning of the state system of education. The FDOE and the State legislature alike desire such assessments as gauges of effectiveness. Indeed, in our highly operationalized life, the use of so-called dashboard indicators has become both commonplace and important. We believe, however, that the use of limited pieces of information, such as the percentage of students

passing a given test, is simply one incomplete picture of the test results and is less likely to be useful than some other, broader picture. For example, we suggest using average scores (or arithmetic means) in conjunction with the percentages of students passing the test, as shall be seen later in this report.