BUROS
Center for Testing

Initial Report to the Florida Department of Education
Recommendations on FCAT for 2007-2008

Report prepared for the

Florida Department of Education

by:

Kurt F. Geisinger, Ph.D.

Craig S. Wells, Ph.D.

Brett P. Foley, M.S.

September, 2007

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

Initial Report to the Florida Department of Education
Recommendations on FCAT for 2007-2008

Buros Institute for Assessment Consultation and Outreach
Buros Center for Testing
University of Nebraska, Lincoln
September, 2007

This brief report is the first report from the Buros Institute for Assessment Consultation and Outreach (BIACO), a division of the Buros Center for Testing at the University of Nebraska, Lincoln to the Florida Department of Education (FDE). It provides some initial thoughts regarding the Florida Comprehensive Assessment Test (FCAT) program in Florida used as part of the accountability system for Florida schools.

The focus of this report is on the 2007-08 FCAT program. We realize that some of these comments may come too late to change the program for this year; however, we believe they will improve the FCAT program this year and, even more so, in the future. These suggestions relate to the strategies for: a) calibration sampling and equating, b) use of content clusters, and c) the involvement of oversight bodies, as well as a few minor issues.

Our second report will deal with what some might consider equating or scoring strategies that have been employed in recent years on the FCAT, especially in regard to the third grade Reading test, which went up 8 points from 2005-2006 and dropped 6 points from 2006-2007[1]. We note that changes of this size should not be expected and make it difficult to gauge the performance of students across the state in a precise and meaningful manner. In this report, we believe we can provide several procedural suggestions that would help deal with these matters. Our suggestions are based on published measurement literature and can also be observed as matters of practice in others states. Below, first we deal with the calibration sample. Then, we address the content of the anchor tests, the use of content clusters, and the role of oversight bodies.

Size and Nature of the Calibration Sample

The key to year-to-year stability of the FCAT test scores is the equating or calibration sample. Sampling errors in equating can be either random in effect or systematic. Harcourt (and its subcontractor, HumRRO) have been using a stratified proportional sampling strategy. In so doing, they have attempted to maximize the similarity of this sample to the students in Florida public education with some other restrictions. For example, the comparability is to students in "standard" educational programs and schools taking the National Assessment of Educational Progress (NAEP) examinations during a given year and students in the juvenile detention schools are excluded. In utilizing this method of equating, the testing contractor is attempting to reduce systematic errors in sampling. We note that gender is not one of the stratifying

[1] We note that this is not a change in the average scores, but a change in the percentage of students categorized as performing at or above the acceptable level (at or above the third level).

variables. This is important to note because female students do better on average on the Reading tests than male students and there have been a somewhat larger proportion of female students in the calibration samples than in the student population as a whole. While our next report will deal with this issue in more depth, we note at this time that it may also prove useful to stratify on gender.

The numerical goal for the equating sampling is that the calibration sample sizes are approximately 8,000 students. We estimate that the average class statewide is approximately 190,000 students (range across the years is approximately 180,000-200,000). Thus, the sampling is approximately 4%-4.4% of the grade-level population. We believe that this sample is smaller than preferable. Moreover, there are four distinct anchor forms, each composed of one reading passage (for the reading tests) and seven multiple-choice test items.

The goal of the current sampling design is to have 2,000 responses for each anchor form (composed of one reading passage and seven questions) so that there are at least 1,500 standard education students in that sample. Again, we question the size of this sampling plan from a stability perspective, especially given the reasonably large number of anchor items that have been eliminated due to instability of parameters in recent years. We also note that it may not be possible to change this number in the short term and that this change may demand some additional research. Nevertheless, random sampling errors would be reduced by increasing the equating sample size. To be sure, the literature is not certain in terms of how large calibration samples should be.

Stocking (1990) employed samples of 2500 for item calibrations and then stated, “Calibration samples, particularly for more complex models, typically consist of several thousand examinees” (p. 474). Florida, we note, uses a three-parameter IRT model for their multiple choice items. On the other hand, Tsutakawa and Johnson (1990) recommended a minimum sample size of 1,000 for item calibrations, but also recommend lengthening tests, probably beyond the seven items used as anchors on the FCAT reading test form anchors.

Finally, Hambleton and Jones (1994) reported that sample size is one of the most important factors in estimating item parameters and found overestimation of parameters even with sample sizes of 2,000. They stated, “Use large samples in item calibration to gain precision in the item parameter estimates. An increase in the precision of item parameter estimates will reduce the size of the overestimation in the test information function” (p. 184). We believe that increasing the size of the calibration sample will reduce year-to-year fluctuations that are based on random error. If, on the other hand, those changes are due to other factors—improvements in the educational process, changes in the overall population of the state, and so on, then increasing the sample size will not reduce year-to-year changes. They will, however, make one more certain of the results.

Given the stratified sampling strategy used by the test contractor, we expect that the scorings are unbiased. Nevertheless, we believe that by increasing the size of this equating sample, the year-to-year fluctuations will be decreased as the standard error of the equating is reduced by increasing the number of students tested in the equating sample. In our phone conversation with representatives of Florida on September 26, 2007, we learned that the State of Florida is now planning both to pre-equate using samples as has been the practice in recent years, and also to sample about 5,000 students randomly across the State to check these calibrations. We are encouraged by this step which we see as supportive of our recommendation.

Alignment of Anchors and Actual Assessments

A long held belief in equating is that the anchor items on a test should parallel the actual test to the extent possible (e.g., Kolen & Brennan, 2004). That is, the items composing the anchor should parallel the operational test in terms of content coverage, difficulty, cognitive complexity, and cognitive processes employed by the test takers. However, while recent research (Sinharay & Holland, 2007) has raised questions in regard to the necessity of this assumption, one wonders about the alignment between the questions found on the anchor test and the test as a whole. The FDOE and its contractors are obviously going to great lengths *a priori* to locate items for which the psychometric characteristics match those on the core test as a whole.

Sinharay and Holland (2007) reported that content similarity between the full test and the anchor is probably more important than the spread of item difficulties being critical to an appropriate anchor. We believe that systematic reviews by content experts of those items on the anchor would prove highly beneficial. These reviews would have the goal of determining that (1) they measure the skills required on the core test to the extent possible, (2) are in the same proportions as are on the core test, and (3) to the extent possible at the same level of cognitive complexity. This analysis should be performed on the reading, mathematics, science, and writing tests. That is, we recommend that formal alignment studies be performed[2]. First, it should be ascertained that the tests do appear to expert judges to be measuring the types of learning called for in the Florida State Standards. Second, it should be ascertained that the anchors parallel the operational test to the maximum extent possible.

Use of Content Clusters

We have some concerns about the use of the “content clusters” of test questions. Commissioner Pfeiffer has informed us that in the case of some school districts, these “scale” values are shared with teachers who may adapt instruction for individual students on the basis of their scores. While such uses are not high stakes uses to be certain, a number of these clusters are nevertheless too short (that is, have too few items) to

[2] During our conference call on September 26, 2007, we learned that FDOE employs a group of highly respected and experienced former classroom teachers to aid in the test design, development, and review. These individuals are involved in the selection of anchor items and check that the content benchmarks are appropriately balanced. Nevertheless, we believe that making this process a formal part of test on one hand, and describing it more completely in the test documentation on the other, would prove beneficial.

generate scale scores that have adequate reliability. With regard to the Reading tests, this point pertains primarily to the Reference and Research cluster and, to a lesser extent, to the Words and Phrases in Context cluster. We believe that the best use of such clusters is at school-level or district-level reporting whereby teachers and other educational leaders may choose to enhance certain aspects of their curricula because students are not learning at the level that is desired. These scales lack both the reliability and validity (in this case, accuracy) to be used in making educational decisions about individual students, even those decisions that do not appear likely to be detrimental to individual students, such as suggesting remediation. During the phone conversation on September 26, 2007, we were informed that FDOE has sent out instructions to school districts warning them against such uses. We support such instructions and believe that the development of in-service training for teachers in this regard might prove useful.

Oversight Bodies

Role of internal bodies. It is our understanding that FDOE has set up at least two bodies that review and discuss FCAT findings. One of these is composed entirely of school personnel, primarily at the school district level, and the other is a Technical Advisory Committee (TAC), comprised of nationally recognized testing experts. We believe that both of these bodies need to continue to meet with regularity and to be given clear and direct charges. We know some of the people on the latter body and respect them highly. We believe that advice that they can provide on an on-going basis could address some of the facing the State of Florida. The Technical Advisory Committee includes Drs. Allan Cohen, Mark Reckase, and Mark Shermis. This is an immensely well-qualified group, but it is smaller than the committee of professionals from which most states profit[3]. We recommend increasing the size of the TAC to diversify the expertise they can provide as advisors to the FDE.

Role of an external auditor. With increasing scrutiny of testing programs, a number of states have moved to a system where some of the work of the testing contractor is audited or replicated by an independent, third party organization. Although the scope of work varies from state-to-state, the work serves as an additional quality control mechanism. Among the states that have moved in this direction are Massachusetts, Michigan, Minnesota, Ohio, New Hampshire, Rhode Island, and Vermont. The involvement of such a contractor would seek to reduce the possibility of psychometric mistakes, and provide an additional set of psychometric eyes on the judgments involved in high-stakes educational testing. In fact, some of the previously mentioned states will not release the test scores until the external auditor satisfactorily replicates the results. We believe that in an era of high-stakes testing, this model is justified.

After writing the above paragraph, Buros received a couple of documents from Harcourt that we had not previously received related to equating and the calibration samples used to equate the tests. We learned that both HumRRO and the California Test

[3] We were told in the phone conversation on September 26, 2007 that additional members are being added to the Technical Advisory Committee, a change that we heartily support.

Bureau (CTB)/McGraw-Hill along with Harcourt have responsibilities in the equating process. Until learning that other agencies were involved in the equating, our understanding was that Harcourt alone performed this work. The joint responsibilities are still not well understood by Buros[4]. We understand that HumRRO may have some special expertise in this work, but we are somewhat surprised that Harcourt does not equate the tests and has HumRRO confirm these analyses. We believe that CTB/McGraw-Hill then confirms the results using somewhat different analytic procedures (i.e., software). We believe that these independent analytic procedures are likely to suggest that computational mistakes are extremely unlikely. We would like to understand the relationship better and plan to discuss the level of independence in these relationships with our contact person at Harcourt.

More Minor Matters

It was noted that the classical test theory item analysis procedures employed by Harcourt include point-biserial correlations between individual items and overall test scores to demonstrate what has historically been called item discrimination, a strong and important quality in items. This index is a commonly employed procedure and simply represents the Pearson correlation coefficient between the dichotomously scored test items and the overall test score (or content cluster scores). However, we also believe that there are adjustments to this index that are preferable. Correlations between items and test scores (or content cluster scores) are spuriously high due to the fact that the item itself contributes to the variance of the test scores. Essentially, the spuriousness needs to be removed. Computationally, it is relatively easy to have correlations computed between items and the scale scores without including the item in question. Henryson (1971) represented a correction that removes this spuriousness. We recommend that this correction be used. Moreover, we also note that the impact of this correlation is not likely to be critical except in the instance where item-cluster correlations are calculated for the smaller clusters (e.g., Reference and Research, Words and Phrases in Context).

Summary

There will be year-to-year variations in test scores, even in equated test scores. In our next report, we will discuss the nature and size of these disparities from year to year. Some of these differences relate to changes in the underlying population, some to educational differences—mostly improvements, and some to psychometric concerns. Our goal is to reduce these psychometric issues so that real score differences may be seen for exactly what they are. Toward this end, our strongest suggestion in this report is to recommend that the equating-calibration samples be increased in size. We also propose that gender be considered for use as an additional stratifying variable in the calibration sample.

[4] We were told in the phone conversation on September 26, 2007 that HumRRO and CTB are subcontractors to Harcourt. We believe that all contractors should ultimately be responsible to the State of Florida, who is, ultimately, the client.

Procedurally, we also recommend that a body be identified to independently review the analyses performed by Harcourt and HumRRO. If this is the role that CTB/McGraw-Hill is already assuming, then this recommendation may already be operational.

References

Hambleton, R. K. & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education, 7*, 171-185.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 130-159). Washington, DC: American Council on Education.

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

Sinharay, S. & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or some restrictions be relaxed. *Journal of Educational Measurement, 44*, 249-275.

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika, 55*, 461-475.

Tsutakawa, R. K. & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*, 371-390.