2011 SREE Conference Abstract Template

Abstract Title Page

Not included in page count.

Title: Methodological Differentiation in Assessing the Value-added of Florida's Interim Reading Assessment System to Predicting FCAT's Mean Proficiency

Author(s): Barbara R. Foorman, Ph.D., and Yaacov Petscher, Ph.D.

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

Dissatisfaction with traditional "mean proficiency" approaches to accountability that judge students' achievement relative to a benchmark has led to an interest is basing accountability on individual academic growth. Recent alternative approaches are:

1. TN uses Saunders' (2000) EVAAS system that measures individual students' growth in terms of the degree of deviation from the mean level of growth. Students who fall more than 1 SD below the mean growth rate are identified as making less than a year's growth and those above 1 SD, are identified as making more than one year's growth (McCaffrey et al. 2004).
2. Ohio includes both a measure of student academic achievement using standardized test scores and a value-added score of students' academic progress over time. Based on this, individual schools are categories as (a) above expected growth, (b) met expected growth, and (c) below expected growth. In the case of Met Expected Growth, the school level gains in each respective grade and the growth in achievement scores from the previous year and the current year indicate that students are performing at a level reflective of the average student.
3. Colorado is using a new model (the Student Growth Percentile Model) based on conditional percentile ranks and quantile regression (Betebenner, 2008). Student's growth percentile scores characterize student growth by locating a student's current score within the distribution of students who had identical prior achievement. Contextualizing growth within the context of prior achievement helps address the concern that the traditional accountability approach put high-poverty schools at a disadvantage because aggregated test scores are used to create a mean proficiency score.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

In Florida, mean proficiency scores are reported on the Florida Comprehensive Achievement Test (FCAT) as well as recommended learning gains from the developmental scale score. Florida now has another within-year measure of growth in reading comprehension from the Florida Assessments for Instruction in Reading (FAIR). The FAIR reading comprehension screen is administered three times a year in a computer-adaptive system to students in grades 3-10. Our objective in this presentation is to answer the following research questions:

1. What are the correlations between FCAT and the FAIR reading comprehension scores (FCAT success probability and reading comprehension ability)?
2. What is the value added of the FAIR reading comprehension screen to prior FCAT?
3. Does the FAIR reading comprehension screen significantly reduce identification errors above and beyond prior FCAT?
4. What is the value added to predicting FCAT of using Bayesian mean estimates of growth in FAIR reading comprehension compared to difference scores?

5. What is the value added to predicting FCAT of using Bayesian mean estimates of growth and prior FCAT scores?

Setting:
Description of the research location.

The State of Florida's public elementary, middle, and high schools from 2009-2010.

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features or characteristics.

Participants were 951,893 students in grades 3-12 in Florida's public schools. The numbers of students per grade were: 156,265 in grade 3; 130,119 in grade 4; 129856 in grade 5; 107,737 in grade 6; 108,394 in grade 7; 105,071 in grade 8; 112,686 in grade 9; and 101,765 in grade 10.

Intervention / Program / Practice:
Description of the intervention, program or practice, including details of administration and duration.

> N/A.

Research Design:
Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

A secondary analysis of 2009-2010 archival FCAT and FAIR student-level data in grades 3-10 available in Florida Department of Education's Progress Monitoring and Reporting Network (PMRN) was used in this design.

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

The Florida Comprehension Assessment Test (FCAT) is a component of Florida's testing effort to assess student achievement in Reading, Writing, Mathematics, and Science represented in Florida's *Sunshine State Standards* (SSS) (Florida Department of Education [FDOE], 2001). The SSS Reading portion of the FCAT is a group-administered, criterion-referenced test consisting of 6 to 8 informational and literary reading passages (FDOE, 2005). Students in grades 3-10 respond to between 6 and 11 multiple choice items for each passage and are assessed across four content clusters: reading comprehension in the areas of words and phrases in context, main idea, comparison/cause and effect, and reference and research. Reliability for the FCAT-SSS has been shown to be high at .90; moreover, test score content and concurrent validity have been established through a series of expert panel reviews and data analysis (FDOE, 2001). The construct validity of the FCAT-SSS as a comprehensive assessment of reading outcomes recently received strong support in an empirical analysis of its relationships with a variety of other reading comprehension, language, and basic reading measures (Schatschneider et al., 2004).

The Florida Assessments for Instruction in Reading (FAIR; Florida Department of Education, 2009-2011; Foorman, Torgesen, Crawford, & Petscher, 2009)) are formative reading

assessments given three times a year in kindergarten through grade 10. In grades 3-10 students take a computer-adaptive reading comprehension screen consisting of up to three passages with multiple choice questions written to FCAT test specifications. Scores derived from the screen are (a) the FCAT success probability (FSP) score based on logistic regression, and (b) a reading comprehension ability score (RCAS; from which standard scores, percentiles and Lexile® reader measures are also derived). For students predicted not to pass FCAT, two diagnostic measures are available—a short passage reading efficiency measure (based on the maze format) and a word analysis measure. Scores used for these analyses are the FAIR reading comprehension screen's FSP, RCAS, and standard scores. In addition, we use the theta score associated with the ability score in some analyses, which we will call RCA to distinguish it from RCAS.

Analyses examined: 1) correlations between FAIR's FSP and RCAS; 2) multiple regression estimates of variance explained in current FCAT by prior FCAT, FCAT plus FAIR's RCA, and FAIR's RCA by itself; 3) comparisons of negative predictive power in predicting current FCAT with either prior FCAT or prior FCAT plus FAIR's RCA; 4) comparisons of HLM estimates of variance explained in current FCAT by Bayesian mean estimates of growth in RCAS vs. difference scores; and 5) comparison of HLM estimates of variance explained in current FCAT by Bayesian mean estimates of growth with and without prior FCAT.

Findings / Results:
Description of the main findings with specific details.

Correlations between FCAT and FAIR's FSP and standard scores are provided in Table 1.

(Please Insert Table 1 here)

The correlations across grades show strong relations between FCAT and FAIR's RC screen's standard score and between FSP and FCAT. As expected, the FSP correlation with FCAT is stronger than the relation between the RC screen standard score and FCAT because prior FCAT was included in the calculation of FSP. These correlations indicate that both score types strongly predict to end of year FCAT performance.

Results of multiple regression in Table 2 provide information on the value added of FAIR's RCA (theta) to prior FCAT in predicting current FCAT. Prior FCAT accounts for a majority of the variance in predicting current FCAT (minimum=49.5% in grade 4, maximum=64.7% in grade 8). When RCA is added to prior FCAT, the resulting FSP accounts for significant unique variance in predicting current FCAT (minimum=1.7% in grade 9 to 7.3% in grade 6).

(Please Insert Table 2 here)

Although this amount of unique variance may initially seem small, the implications are more readily observed when examining the data in the context of predicting student risk on the FCAT (i.e., FCAT Level <3). The data from Table 3 report the percentage of students who are identified by the variable as "not at-risk" for failing the FCAT using either Prior FCAT or the combination of Prior FCAT and FAIR's RCA.

(Please Insert Table 3 here)

The results from Table 3 demonstrate that using the RCA in conjunction with Prior FCAT produces significantly different results in predicting whether a student will be successful on the current FCAT. For example, in grade 4, if we were to use a student's prior year FCAT score to determine whether or not they will pass current FCAT, we would only be able to correctly say that 86% of the students of the students who were level 3or above last year would be level 3 or above on the current FCAT. As such, 14% of students (i.e., 100%-86%) would mistakenly be identified as being not at-risk for failing the FCAT when they in fact did fail and did not receive appropriate interventions. These values range from 61% in grade 10 to 86% in grades 3-5, meaning that between 14% and 39% of students are mistakenly identified as being on grade level.

Conversely, when using both the prior FCAT and RCA together, this range of incorrect identification reduces from 14% in grade 4 when only prior FCAT is used, to 2% when both are used. Similarly, the incorrect identification rate decreases by 8% in grade 5, 7% in grade 6, 1% in grade 7, 14% in grade 8, and 20% in grade 9, with an increase in error in grade 10 of 7%. With the exception of grade 10, where only 13 of 1,602 had a FSP >=0.85, a significant reduction in identification errors occurs when using both the prior FCAT and RCA.

In Table 4, the base R^2 for predicting FCAT from the winter assessment period is compared to the R^2 when the Bayesian slope from fall to winter is added to the HLM. The Simple Difference Score uses the gain score as the predictor. The value-added can be extrapolated from the Bayesian Slope-Base or the Simple Difference-Base. It is practically important to see how much unique variance is added to the prediction of FCAT with Simple Differences in the FAIR reading comprehension screen and no autoregressor. Additional analyses with comparisons of the fall-spring and winter-spring assessment points are underway.

(Please insert Table 4 here)

The final research question asked what the value added to current FCAT prediction was of using Bayesian mean estimates of growth and prior FCAT scores. To address this we compared the R^2 between the models in Table 4 using procedures identified by Alf and Graff (1999) to statistically compare the estimates of Bayesian slope with and without the autoregressor of prior FCAT. Grade 3 is excluded from the table because of lack of prior FCAT for that grade.

(Please insert Table 5 here)

From Table 5 we see that the addition of the autoregressor of prior FCAT adds, on average, 2% unique variance to the prediction of current FCAT.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

Are there ways to improve on the traditional "mean proficiency" approach to accountability? Using Florida's new formative assessment system (FAIR) with nearly 1 million students in grades 3-10 from 2009-2010, we showed that the answer is Yes and that the degree of improvement depends on the methodology. The FAIR reading comprehension screen correlated highly with FCAT reading scores. Incorporating the autoregressor of prior FCAT and FAIR

performance at the winter assessment into the prediction of current FCAT accounted for up to 7% unique variance. The Simple Difference approach to measuring FAIR growth accounted for up to 2-3% unique variance above and beyond the model which included the FCAT autoregressor and winter performance on the FAIR. Such improvements in FCAT prediction lead to (a) reduction in false positives in classifying student risk and (b) better placement for reading intervention.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Alf, E.F., Jr., & Graf, R.G. (1999). Asymptotic confidence limits for the difference between two squared multiple correlations: A simplified approach. *Psychological Methods*, *4*, 70-75.

Betebenner, D.W. (2008). *A primer on student growth percentiles*. Dover, NH: National Center for the Improvement of Educational Assessment.

Florida Department of Education (2001). *FCAT handbook—A resource for educators*. Tallahassee, FL: Author.

Florida Department of Education (2005). *FCAT briefing book*. Tallahassee, FL: Author.

Florida Department of Education (2009). *Florida Assessments for Instruction in Reading (FAIR)*. Tallahassee, FL: Author.

Florida Department of Education (2009). FAIR 3-12 Technical Manual. Tallahassee, FL: author. Retrieved from http://www.fcrr.org/FAIR/3-12_Technical_Manual_FINAL.pdf

Foorman, B., Torgesen, J., Crawford, E., & Petscher, Y. (2009). Assessments to guide reading instruction in K-12: Decisions supported by the new Florida system. *Perspectives on Language and Literacy*, *35(5)*, 13-19.

McCaffrey, D., Lockwood, J., Koretz, D., Louis, T., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*, 67-101.

Sanders, W. (2000). Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education*, *14*, 329-339.

Schatschneider, C., Buck, J., Torgesen, J.K., Wagner, R.K., Hassler, L., & Hecht, S., et al., (2004). A multivariate study of factors that contribute to individual differences in performance on the Florida comprehensive Reading Assessment Test (Technical Report No. 5), Florida Center for Reading Research, Tallahassee, FL. Available at: http://www.fcrr.org/TechnicalReports/Multi_variate_study_december2004.pdf.

Appendix B. Tables and Figures

Not included in page count.

Table 1. Correlations between the FCAT and both RC Screen and FSP

Grade	Fall RC Screen Standard Score	FSP	Winter RC Screen Standard Score	FSP	Spring RC Screen Standard Score	FSP
3	0.72	0.71	0.74	0.72	0.75	0.73
4	0.69	0.74	0.73	0.74	0.74	0.74
5	0.70	0.75	0.73	0.76	0.73	0.76
6	0.73	0.74	0.72	0.74	0.72	0.74
7	0.71	0.72	0.69	0.72	0.69	0.72
8	0.71	0.76	0.71	0.76	0.71	0.76
9	0.69	0.73	0.67	0.73	0.67	0.73
10	0.69	0.75	0.67	0.74	0.66	0.74

Table 2. Estimates of Variance Explained by Prior FCAT, FCAT + RCA, and FSP by Itself.

	Grade						
Variables	4	5	6	7	8	9	10
Prior FCAT	49.5%	58.6%	53.0%	60.0%	64.7%	57.5%	59.0%
Prior FCAT + RCA	53.3%	62.5%	60.3%	63.0%	68.4%	59.2%	61.3%
FSP	3.8%	3.9%	7.3%	3.0%	3.7%	1.7%	2.3%

Table 3. Comparing Negative Predictive Power in Predicting Current FCAT with either Prior FCAT or Prior FCAT + FAIR's RCA

	Grade						
Variables	4	5	6	7	8	9	10
Prior FCAT	86%	86%	86%	84%	78%	70%	61%
Prior FCAT + RCA	98%	98%	93%	85%	92%	90%	54%

Table 4. HLM Estimates of FCAT Comparing R^2 in Growth vs. Difference Score Models

Grade	Base	Bayesian Slope	Simple Difference
3	0.55	0.55	0.62
4	0.53	0.53	0.59
5	0.53	0.53	0.60
6	0.51	0.51	0.61
7	0.47	0.47	0.57
8	0.50	0.50	0.59
9	0.45	0.45	0.55
10	0.45	0.45	0.55

Table 5. HLM Estimates of FCAT Comparing R^2 in Autoregressive, Growth, and Difference Score Models

Grade	FCAT	FCAT + RCA	Bayesian Slope	Simple Difference
4	0.58	0.65	0.65	0.68
5	0.6	0.66	0.66	0.68
6	0.62	0.67	0.68	0.69
7	0.56	0.61	0.62	0.63
8	0.51	0.57	0.57	0.59
9	0.5	0.53	0.54	0.55
10	0.52	0.57	0.57	0.59