



ELSEVIER

Journal of School Psychology
xx (2007) xxx–xxx

Journal of
School
Psychology

Accuracy of the DIBELS Oral Reading Fluency Measure for Predicting Third Grade Reading Comprehension Outcomes

Alysia D. Roehrig*, Yaacov Petscher, Stephen M. Nettles,
Roxanne F. Hudson¹, Joseph K. Torgesen

Florida State University and The Florida Center for Reading Research, 227 N. Bronough St.,
Suite 7250, Tallahassee, FL, 32301 USA

Received 3 November 2006; received in revised form 8 May 2007; accepted 14 June 2007

Abstract

We evaluated the validity of DIBELS (*Dynamic Indicators of Basic Early Literacy Skills*) ORF (*Oral Reading Fluency*) for predicting performance on the *Florida Comprehensive Assessment Test* (FCAT-SSS) and *Stanford Achievement Test* (SAT-10) reading comprehension measures. The usefulness of previously established ORF risk-level cutoffs [Good, R.H., Simmons, D.C., and Kame'enui, E.J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, 5, 257–288.] for third grade students were evaluated on calibration ($n_{S1} = 16,539$) and cross-validation ($n_{S2} = 16,908$) samples representative of Florida's *Reading First* population. The strongest correlations were the third (February/March) administration of ORF with both FCAT-SSS and SAT-10 ($r_S = .70-.71$), when the three tests were administered concurrently. Recalibrated ORF risk-level cut scores derived from ROC (receiver-operating characteristic) curve analyses produced more accurate identification of true positives than previously established benchmarks. The recalibrated risk-level cut scores predict performance on the FCAT-SSS equally well for students from different socio-economic, language, and race/ethnicity categories.

© 2007 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.

Keywords: Oral reading; Reading comprehension; Validity; At risk populations

* Corresponding author. Tel.: +1 850 644 9080; fax: +1 850 644 9085.

E-mail address: aroehrig@fcrr.org (A.D. Roehrig).

¹ Current address: University of Washington, USA.

0022-4405/\$ - see front matter © 2007 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.

doi:[10.1016/j.jsp.2007.06.006](https://doi.org/10.1016/j.jsp.2007.06.006)

Please cite this article as: Roehrig, A. D. et al. Accuracy of the DIBELS Oral Reading Fluency Measure for Predicting Third Grade Reading.... *Journal of School Psychology* (2007), doi:[10.1016/j.jsp.2007.06.006](https://doi.org/10.1016/j.jsp.2007.06.006)

Accuracy of the DIBELS Oral Reading Fluency Measure for Predicting Third Grade Reading Comprehension Outcomes

Across the United States, 1.5 million of some of our students most at risk for not reading on grade level are being served by the No Child Left Behind Act's *Reading First* program (U.S. Department of Education [USDOE], 2006a). The *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS; Good & Kaminski, 1996) include a measure of *Oral Reading Fluency* (ORF) that is widely used in the *Reading First* assessment plans. While there is debate about the efficacy of *Reading First*, especially in relation to its assessment recommendations (e.g., allegations of impropriety that surround the DIBELS measures and debates about their validity; Goodman, 2006), DIBELS ORF is one of few empirically validated standardized reading fluency assessments available for widespread progress monitoring (Kame'enui et al., 2006).

The construct of reading fluency itself has been under a spotlight (e.g., Rasinski, 2006; Rasinski, Blachowitz, & Lems, 2006; Samuels & Farstrup, 2006) since being highlighted as one of the five components of reading instruction included in the National Reading Panel's (NRP, 2000) report (i.e., summary of research findings on reading instruction that guides *Reading First*). This brought the instruction and assessment of reading fluency to the forefront of debate by teachers, policy makers, and researchers. At the heart of much of the debate about DIBELS ORF are questions about its validity as a progress monitoring measure (e.g., Goodman, 2006; Pressley, Hilden, & Shankland, in press).

The goal of the research presented here was to answer the call for more empirical investigation of DIBELS ORF (Pressley et al., in press; Samuels, 2006a,b). First, to estimate the predictive and concurrent validity of using ORF to identify students who are at risk for below grade-level reading achievement, we examined the relationship between DIBELS ORF and two reading comprehension measures. Second, we investigated the appropriateness of ORF risk-level cut scores, which were developed by the DIBELS authors and adjusted for more frequent administration (University of Oregon Center on Teaching and Learning, 2006), for predicting outcomes on a reading comprehension assessment other than the *Oregon Reading Assessment* and for a different population. While Hintze, Ryan, and Stoner (2003) examined the predictive power of several DIBELS measures for accurately classifying students as at risk for not reading on grade level, no published evaluations of the diagnostic accuracy of DIBELS ORF could be identified. Adjusting cut scores, such that they are as accurate as possible with a preference for over- rather than under-classification of at risk students, would decrease the likelihood of not identifying students who need additional interventions. Third, given that the purpose of *Reading First* is to serve schools where students tend to be at risk for reading difficulties, particularly those with large proportions of socio-economically disadvantaged students, ORF also was evaluated for predictive bias across demographic subgroups. The following brief overview of theories and research on reading fluency and its assessment in the curriculum based measurement tradition provides a context in which to situate our evaluation of one measure of oral reading fluency, the DIBELS ORF.

Defining reading fluency

If one is automatic in the ability to decode text, “the text can be decoded with ease, speed, and accuracy” (Samuels, 2006a, p. 9; see also Meyer & Felton, 1999). Accuracy and

automaticity are considered central components of reading fluency because they are consistently associated with the end goal of reading, comprehension (e.g., Breznitz, 1987, 1991; Chard, Vaughn, & Tyler, 2002; Deno, Marston, Shinn, & Tindal, 1983; Dowhower, 1987; Fuchs, Fuchs, & Maxwell, 1988; Perfetti & Hogaboam, 1975; Rasinski, 1989, 1990; Tenenbaum & Wolking, 1989). Thus, reading fluency is defined by many as rate and accuracy of oral reading in connected text (e.g., Fuchs & Fuchs, 1992; Hasbrouk & Tindal, 2006; Shinn, Good, Knutson, Tilly, & Collins, 1992; Torgesen, Rashotte, & Alexander, 2001).

Along with many others, Hudson, Lane and Pullen (2005) define reading fluency as comprising three primary elements. These include not only accuracy and rate, but also prosody. Prosody describes the expressiveness of oral text reading as it is related to intonation, stress patterns, and phrasing (for a review see Hudson et al., 2005; see also Allington, 1983; Dowhower, 1991; NRP, 2000; Schreiber, 1980, 1991). Others, such as Allington et al. (see Mathson, Allington, & Solic, 2005), suggest that accuracy, automaticity, and prosody are all equally important components of fluency. Still other researchers have noted that “while rate and accuracy of oral reading are relatively straightforward characteristics both to observe and measure, it has proven more difficult to capture and measure the fluency, or ‘ease,’ with which children read texts” (Daane, Campbell, Grigg, Goodman, & Oranje, 2005, p. 27).

Empirically, it remains unclear whether prosody is an aid to or a result of comprehension (Mathson et al., 2005; Schwanenflugel, Hamilton, Kuhn, Wisenbaker, & Stahl, 2004). If prosody is enabled by comprehension, then its usefulness in the assessment of fluency would be primarily as an index of the extent to which students comprehend what they are reading. However, one recent study that examined this question did not find evidence that direct assessment of prosody provided evidence of reading comprehension beyond that provided by assessments of accuracy and rate (Schwanenflugel et al.). For these reasons, for the purposes of this study, we define oral reading fluency as accuracy and rate in connected text, or correct words per minute.

Assessing reading fluency

The DIBELS ORF measure is an example of curriculum based measurement (CBM), which was developed to incorporate data-based decision-making into instructional planning (Deno, 1989; Shinn, 1989). CBM focuses on the direct and continuous measurement of student progress toward specific instructional objectives in order to determine appropriate instructional methods (Marston, 1989; Tindal & Marston, 1990). It answers two questions: Is a particular student or group of students performing at an expected level given particular instructional conditions, and is that instruction strong enough for the student or group of students to make sufficient progress to achieve an expected goal at the end of the instructional period? There is substantial evidence of the validity and reliability of curriculum based measures in assessing oral reading fluency (e.g., Deno, Mirkin, & Chiang, 1982; Deno et al., 1983; Fuchs & Fuchs, 1992; Fuchs et al., 1988; Fuchs, Fuchs, Hosp, & Jenkins, 2001).

In addition to the substantial empirical base supporting CBM use in general, there also is a growing body of evidence from technical reports supporting DIBELS ORF as a progress monitoring assessment. Correlations of .65–.80 were found between DIBELS ORF and several state assessments of reading (Barger, 2003; Buck & Torgesen, 2003; Good, Simmons,

& Kame'enui, 2001; Shaw & Shaw, 2002; Vander Meer, Lentz, & Stollar, 2005; Wilson, 2005). Reliability and validity also have been studied for second grade ORF passages (Good, Kaminski, Smith, & Bratten, 2001), with median alternate form reliability of .95 and concurrent validity with the *Test of Oral Reading Fluency* (Children's Educational Services, 1987) of .92–.96. Although these correlations with state-developed outcome measures provide substantial evidence of the validity of using DIBELS ORF to predict end of year performance, there is little research on how well ORF predicts performance on more comprehensive measures of reading comprehension. Pressley et al., in press, however, found that DIBELS ORF predicted reading comprehension as measured by the *TerraNova* reading comprehension assessment (CTB/McGraw-Hill, 2004) better than teachers' informal assessment of reading comprehension.

Context of assessment

Florida *Reading First* serves a large and diverse population, including hundreds of schools and thousands of students (e.g., 59,728 third graders in 2006). Of third grade students educated in Florida *Reading First* schools, 75% are eligible for free or reduced-price lunch, 20% have at least one identified disability, 12% of students speak a first language other than English, and 36% of students are White, 36% African American, and 23% Latino. The Florida *Reading First* population of students, however, is not unique. Other large states also serve diverse populations through *Reading First*. For example, in California across 3 cohort years, 86 to 88% of students at *Reading First* schools were economically disadvantaged, 7–9% of students had at least one identified disability, 56–59% were English Language Learners, 4–10% were White, 6–15% African American, and 73–78% Latino (Educational Data Systems, 2005). Across the nation, the Title I program also serves schools and students with demographics very similar to those at *Reading First* schools (USDOE, 2006b).

Hence, it is of particular concern to learn more about how well widely-used assessments, such as DIBELS ORF, work in such diverse populations. Since many of the schools using the ORF measure serve a large proportion of students who are traditionally considered at risk for academic achievement, it is important to discern whether there is any bias in ORF's predictive utility for different demographic groups (Reynolds, 2000). While there is evidence that CBM measures of oral reading fluency other than DIBELS ORF are not biased in the prediction of comprehension outcomes for African American and White students (Hintze, Callahan, Matthews, Williams, & Tobin, 2002; Hixson & McGlinchey, 2004) and that oral reading probes are similarly valid and reliable for use with English-only and bilingual Spanish-speaking students (Baker & Good, 1995), apprehension about the use of oral reading fluency probes with students of varying language backgrounds still exists (Klein & Jimerson, 2005). In this study, we examine the extent of predictive bias of the DIBELS ORF for FCAT-SSS reading comprehension outcomes across several demographic characteristics.

Answering the call for research

The purpose of this study was to examine the concurrent and predictive validity of DIBELS ORF not only with the *Florida Comprehensive Assessment Test* (FCAT-SSS; a

third grade reading comprehension assessment used in Florida for accountability decisions), but also with the 10th edition of the *Stanford Achievement Test* (SAT-10; a nationally normed standardized reading comprehension assessment). Three research questions were addressed: 1) Does the DIBELS *Oral Reading Fluency* measure predict performance on a widely-used standardized measure of reading comprehension as well as it predicts performance on a state-developed reading accountability measure (i.e., FCAT-SSS)? 2) Can the risk levels established by the developers of DIBELS ORF (Good, Simmons et al., 2001) be improved for third grade students in Florida *Reading First* schools who take FCAT-SSS? 3) Do the optimal risk-level cutoffs for the DIBELS ORF measure predict performance on FCAT-SSS equally well for students from different categories of socio-economic status, language status, or race/ethnicity?

Method

Participants

The participants were 35,207 third grade students enrolled in Florida *Reading First* schools during the 2004–2005 school year (i.e., all third grade students attending Florida *Reading First* schools who were assessed according to the *Reading First* plan in 2004–2005). According to school records, this cohort of third graders reflect the diversity often found in *Reading First* schools: 49% were female, 36% were identified as White, 36% as African American, 23% as Latino, 3% as Multiracial, 1.5% as Asian, and less than 1% as Native American. Across the sample, 75% were eligible for free or reduced-price lunch, and 17% were served on an Individual Education Plan for a disability (9% Specific Learning Disability; 4% Speech Impaired, 3% Language Impaired, 4% Various Other Impairments). Programs for limited English proficiency served 12% of students, and 3% of students were identified as gifted.

Because we intended to evaluate the usefulness of DIBELS ORF as a predictor of future reading comprehension achievement, the participants were split into two samples: calibration (S1) and cross-validation (S2). We did so using a stratification procedure to ensure that the two samples correctly reflected the demographics of the larger Florida *Reading First* population. This resulted in 17,409 students in the calibration sample and 17,798 students in the cross-validation sample. The samples were equitable on all demographic variables mentioned above (Table 1).

Measures

The data used in this study were obtained from the Progress Monitoring and Reporting Network (PMRN) maintained by the Florida Center for Reading Research (FCRR) as part of its role in providing support for statewide *Reading First* programs. The PMRN is a centralized data collection and reporting system through which *Reading First* schools in Florida both report reading data and receive reports of the data for instructional decision-making. State, district and school staff in *Reading First* schools use this system as the primary source of data regarding student performance. The progress monitoring measures, which were administered four times a year in 2004–2005 as mandated by the Florida

Table 1
Demographic frequency percentage comparisons

Category	PMRN	S1	S2	Missing FCAT data
Gender				
Male	51.0	51.2	51.4	53.0
Female	49.0	48.8	48.6	47.0
Ethnicity				
White, non-Latino	36.0	35.8	36.2	38.1
African American, non-Latino	36.0	36.3	36.1	34.1
Latino	23.0	22.5	22.4	21.5
Multiracial	4.0	3.5	3.5	4.0
Asian	2.0	1.5	1.5	1.8
American Indian	<1.0	0.4	0.3	0.5
FRL				
Eligible	75.0	75.2	74.7	75.4
Not eligible	25.0	24.8	25.3	24.6
ESE				
None	77.0	76.9	76.9	74.7
Specific Learning Disability	9.0	8.9	8.3	10.2
Speech Impaired	4.0	4.5	4.8	3.0
Language Impaired	3.0	3.0	3.3	3.1
Other Impairments	4.0	3.0	3.0	7.0
Gifted	3.0	3.0	3.0	2.0
English proficiency				
Limited English proficient	12.0	20.0	20.0	23.0
English proficient	88.0	80.0	80.0	77.0

Note. In Florida, gifted students are included in the ESE category.

Reading First assessment plan, vary by grade level. In kindergarten, the progress monitoring measures focus primarily on phonological awareness, letter knowledge, and decoding while in third grade only the ORF is given. In addition to progress monitoring measures, the following outcome measures are given at the end of each grade: the *Peabody Picture Vocabulary Test* (PPVT), the SAT-10 (first through third grades), the reading vocabulary subtest of the *Gates-MacGinitie Reading Test* (second and third grade), and the *Florida Comprehensive Assessment Test* (FCAT-SSS; third grade only).

DIBELS Oral Reading Fluency

ORF (5th Edition; Good et al., 2001) is a measure that assesses oral reading fluency in grade-level connected text. This standardized, individually administered test of accuracy and reading rate in connected text was designed to identify children who may need additional instructional support and to monitor progress toward instructional goals (Good & Kaminski, 1996). Students read three passages aloud for 1 min each, with the cue to “be sure to do your best reading” (Good, Kaminski, Smith, Laimon, & Dill, 2001, p. 30). Words omitted, substituted, and hesitations of more than 3 s are scored as errors. Words self-corrected within 3 s are scored as accurate (Good, Wallin, Simmons, Kame’enui, & Kaminski, 2002). The assessor notes errors, and the score is the number of correct words read per minute; the median score from the three passages is the data point for decision-making.

Although information about the development and setting of ORF benchmarks for third grade is available for a three assessment administration (Good et al., 2002), no data is currently available on a four assessment period. Florida administrations of ORF typically occur during the months of September (Fall Assessment; A1), December (Winter 1 Assessment; A2), February/March (Winter 2 Assessment; A3), and April/May (Spring Assessment; A4). The fall and spring assessments' indicators of risk are based on the DIBLES goals for assessing three times per year. Winter 1 and 2 performance targets in Florida were based on interpolation between the Fall, Winter, and Spring DIBLES benchmarks assuming linear growth in oral reading rate between measurement points.

Florida Comprehensive Assessment Test-Sunshine State Standards (FCAT-SSS)

The FCAT-SSS 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 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).

Students completing the reading portion of the FCAT-SSS are placed in one of five performance levels based on a scale score ranging from 100–500. Levels 1 and 2 reflect below grade-level performance in reading, with Level 1 being the lowest indication of reading performance. Levels 3 and above represent proficiency in reading comprehension at or above grade-level standards. For this study, Levels 1 and 2 were collapsed into one grouping variable representing below grade-level reading performance (coded as 1 for analyses), and Levels 3, 4 and 5 were combined into one grouping variable representing at or above grade-level performance following previously used methodology (coded as 0; Buck & Torgesen, 2003).

Stanford Achievement Test

SAT-10 (10th edition; Harcourt Brace, 2003a) is a widely-used standardized measure of reading comprehension. Classroom teachers administer this untimed test in a group format in all Florida *Reading First* schools, and it is scored by the test publisher. Students answer a total 54 multiple choice items (assessing initial understanding, interpretation, critical analysis, and awareness and usage of reading strategies) in response to reading literary, informational, and functional text passages. The reliability coefficient for SAT-10 on a nation-wide representative sample of students was .88. Validity was established with other standardized assessments of reading comprehension,

providing evidence of content, criterion-related, and construct validity (Harcourt Brace, 2003b).

Procedures

Assessment administration

Students in both samples were administered DIBELS ORF four times per year by well-trained and reliable school- or district-based assessment teams that contained no classroom teachers. Administrations of ORF occurred during two week assessment windows in the months of September, December, February/March, and April/May. Students also completed the FCAT-SSS and SAT-10 during the February/March assessment window of each year, as required by Florida statute. All assessments were administered in the classroom setting following appropriate assessment protocols.

Diagnostic accuracy analysis

Diagnostic efficiency of each ORF assessment was tested by generating a receiver-operating characteristic (ROC) curve and examining the sensitivity (i.e., proportion of students correctly classified as at risk on both FCAT-SSS and ORF) and specificity (i.e., proportion of students correctly classified as *not* at risk on both measures) of cut score values used to screen for readers at risk for not reaching grade-level reading achievement as measured by FCAT-SSS. Several methods exist for evaluating the appropriateness of existing cut scores, including: discriminant analysis, equipercetile equating, logistic regression, and ROC curve analysis. Of these approaches, ROC curve analysis has been shown to provide greater flexibility with regard to estimated diagnostic accuracy and predictive power (Silbergliitt & Hintze, 2005), as well as in determining the balance between Type I and II errors. Since ORF's Spring Assessment occurs after FCAT-SSS is given, the recalibration of current indicators of risk was only conducted on the first three assessments. Practically, it was inappropriate to make a prediction about FCAT-SSS risk from an assessment given *after* FCAT.

A selected number of indices were calculated to provide multiple perspectives on the effectiveness of the individual ORF scores: sensitivity (SE), specificity (SP), area under the curve (AUC), likelihood ratios (LR), predictive power, and overall correct classification (OCC). AUC is a probability index ranging from 0.5–1.0, and provides the probability of the independent variable correctly classifying a pair of individuals, where one student is at risk and the other is not. Values closer to 0.5 indicate that the independent variables do not classify well, while an AUC of 1.0 reflects a test that perfectly classifies; the AUC also is considered to be sufficient as an effect size indicator (Swets, 1988). Positive and negative likelihood ratios (LR) are the odds that positive test results (e.g., ORF High Risk or Low Risk) come from a student who is either truly at risk (LR+) or truly not at risk (LR–). Values close to 1.0 provide evidence that the independent variable is not diagnostically useful and does not contribute to classification accuracy (Streiner, 2003). Positive Predictive Power (PPP) is the probability that a student who is identified as being at risk according to ORF is truly at risk. Similarly, Negative Predictive Power (NPP) is the chances that a student who is identified as being not at risk according to ORF is truly not at risk. Positive and negative predictive power are limited in their generalizability as they are

affected by the base rate of the presented condition in the sample: if there are more people at risk in a sample, then the predictive power estimates could be artificially inflated. Overall correct classification is the proportion of students who are correctly classified as truly at risk or truly *not* at risk divided by the total number of students as a function of the criterion and predictor variables. These seven indices were selected for reporting as they represent the most commonly reported efficiency statistics (Streiner) and all but positive and negative predictive power are not affected by base rates, which is an important property when examining diagnostic accuracy in a high or low prevalence population (Meehl & Rosen, 1955; Streiner).

Optimal cut scores for the samples were determined by an examination of values represented at the shoulder of the ROC curve, and tested in a 2×2 contingency table. The specific classification indices were calculated for several ranges of cut scores for each ORF administration in order to observe the changes in sensitivity and specificity that occur when scores are adjusted along the horizontal and vertical axes on the ROC curve. Analyses for within-group AUC differences were run to evaluate the dependency of the results on the sample (Hanley & McNeil, 1983). Establishment of the new benchmarks also was considered in conjunction with methodology employed by Good et al. (2002). Part of their guiding principles in developing the original cut scores was to retain intervals for *low risk* levels that resulted in at least 80% of students meeting the end of year goal. Additionally, they wanted to set an interval for *high risk* whereby 20% or fewer of students met the third grade goal. Good et al. also outlined that the *some risk* students should have a 50% probability of meeting the end of year goal.

Predictive bias analysis

Testing if the optimal risk-level cutoffs for ORF predicted performance on FCAT-SSS equally well for students with different socio-economic status, language status, or race/ethnicity involved a series of logistic regression analyses. The calibration and cross-validation samples were combined to utilize the power of the full sample, with FCAT-SSS performance modeled as the outcome (0=not at risk, 1=at risk) and selected independent variables (Free or Reduced-Priced Lunch Status, English Language Learner Status, Race/Ethnicity) entered as predictors in separate models with ORF. Each demographic characteristic was dummy coded, with the referent individual reflecting the traditionally low risk group: Free or Reduced-Priced Lunch (0=not eligible, 1=eligible), English Language Learner Status (0=English proficient, 1=limited English), and Race/Ethnicity (0=White, 1=African American or Latino).

Main effects for the demographic variables and ORF (0=not at risk, 1=at risk) were simultaneously entered as covariates, along with the interaction between the two variables. Odds ratios were calculated from the beta estimates provided to enhance interpretation and provide a communicable effect size. Using a temporal based rationale, it was hypothesized that the concurrent administrations of reading assessments (i.e., ORF Winter 2 Assessment and FCAT-SSS) would maximize any subgroup differentiation in predictive dependency. Thus, analyses were conducted only using the Winter 2 Assessment ORF data. It was expected that ORF would be a significant predictor of FCAT-SSS risk, as would main effects for the demographic characteristics. Conversely, it was believed that the interaction

between ORF and demographics would not be significant in predicting FCAT-SSS risk, indicating a lack of bias.

Results

Missing data analysis

Since the recalibration of cut scores was contingent on students, true risk identification (i.e., FCAT level), participants without an FCAT score were removed, reducing the number of participants in the calibration ($n_{S1} = 16,539$) and cross-validation ($n_{S2} = 16,908$) samples. The demographic makeup of excluded students was examined to determine if data were missing at random, or if a student characteristic bias was introducing systematic error into data removal. Frequency distributions indicated that data were not missing in a consistent pattern, and the participant characteristics were comparable to the makeup of the full sample. Similarly, the reduced participant pool matched the original sample in terms of demographic frequencies (Table 1; analyses available upon request).

Reading fluency and comprehension performance

Descriptive statistics and correlations for students' third grade (2004–05) ORF, FCAT-SSS, and SAT-10 scores are reported in Table 2 for both samples. Moderate to strong correlations were observed for the relationship among students' ORF scores in third grade. The relationship between ORF and FCAT-SSS in both groups increases in magnitude over time, peaking at the ORF Winter 2 Assessment ($r_{S1} = .71$, $r_{S2} = .70$), corresponding with the time both FCAT-SSS and ORF were administered. Similarly, the correlations of ORF scores with SAT-10 peak at the Winter 2 Assessment, when they were concurrently administered, with comparable magnitudes to the relationship between ORF and FCAT-SSS ($r_{S1} = .71$, $r_{S2} = .70$). In regards to the first research question, oral reading fluency predicts reading comprehension performance on FCAT-SSS equally as well as on SAT-10.

Table 2
Sample correlations and descriptive statistics for oral reading fluency scores

	I	II	III	FCAT-SSS	SAT-10
Calibration (S1)					
I. Fall Assessment	1.00			.66	.68
II. Winter 1 Assessment	.90	1.00		.68	.68
III Winter 2 Assessment	.88	.92	1.00	.71	.71
Cross-validation (S2)					
I. Fall Assessment	1.00			.67	.69
II. Winter 1 Assessment	.91	1.00		.68	.68
III Winter 2 Assessment	.89	.92	1.00	.70	.70
Mean	70.15 (69.20)	93.90 (93.06)	101.12 (100.39)	292.39 (290.91)	611.23 (610.61)
SD	32.73 (32.78)	35.20 (35.19)	33.42 (33.36)	60.98 (60.92)	36.55 (36.37)

Note. Sample sizes for 2004–05 correlations in both samples range ($n = 15,804$ – $16,864$). All correlations significant at $p < .01$ level. Cross-validation sample means and SDs are in parentheses.

Risk-level recalibration

First, an analysis of variance (ANOVA) was run to test the differences between the two samples on the ORF, FCAT-SSS, and SAT-10 measures to ensure that the samples did not have significant mean differences. Bonferonni correction was applied, increasing the significance threshold to $p < .01$ (.05/5). Results indicated that groups did not statistically differ on the selected outcomes (A1ORF [$F(1,33384)=4.32, p=.043$]; A2ORF [$F(1,31974)=4.51, p=.037$]; A3ORF [$F(1,33354)=3.93, p=.049$]; FCAT-SSS [$F(1,33414)=4.89, p=.027$]; SAT-10 [$F(1,33475)=2.40, p=.122$]).

In order to examine the appropriateness of the cut scores for *Reading First* schools in Florida in making risk-status decisions (research question 2), student classification across each ORF assessment was examined using the current benchmarks used in the state of Florida. Using these criteria, the frequency percentage of students at each risk level (i.e., low risk, some risk, at risk) as well as the percentage of students who met the end of year benchmark goal (i.e., FCAT-SSS level ≥ 3) were reported and compared to results found by Good et al. (2002). The recalibrated cut scores were compared to Good et al.'s cut scores for the Fall Assessment; however, comparisons were not possible for the Winter 1 and Winter 2 Assessments since Good et al.'s results were reported on ORF administration three times per year rather than four.

Fall assessment

The University of Oregon Center on Teaching and Learning (2006) reported that students with ORF scores at or above 77 on the Fall Assessment are considered to be at low risk for developing below-grade-level reading achievement; students with scores between 53 and 77 are at some risk; and scores less than 53 are at high risk for below-grade-level reading achievement at the end of third grade. Furthermore, in their sample they reported 90% of low risk, 34% of some risk and 3% of at risk students achieved the grade-level goal at the end of third grade. In our combined sample, a narrower range was observed across risk levels (i.e., 86%, 62%, 20%). The recalibrated cut scores were evaluated in terms of their diagnostic efficiency and were compared to other plausible values to achieve the optimal sensitivity and specificity estimates. Results indicated that adjusting the at risk threshold to <45 and low risk to >76 improved efficiency of classification, evidenced by the higher values in the sensitivity, specificity, and overall correct classification (Table 3) for both samples as compared to the benchmarks currently used. Additionally, the proportion of students meeting the end of year benchmark goal using the recalibrated scores was comparable to those found using the current criteria in our sample (Table 4). Still, even with improved classification, the distribution in the percentage of students achieving the end of year benchmark were not as close to the one's observed by Good et al. (2002); however, they were generally within the guidelines of at least 80% or greater of low risk and 20% or fewer of high risk meeting the end of year goal.

Subsequent frequency distributions of students in all risk levels were examined since diagnostic efficiency analyses and cut score adjustment do not consider students who are classified as some risk. Generally, students falling within the some risk range of scores on ORF as currently defined are equally likely to have an adequate or inadequate performance on FCAT-SSS (Buck & Torgesen, 2003). Contrasting the equal likelihood of FCAT-SSS

Table 3
Diagnostic efficiency results

	Sensitivity	Specificity	LR+	LR-	PPP	NPP	OCC
Fall Assessment (A1)							
Current cut scores (S1 & S2)	.74	.86	5.29	3.31	.81	.80	.81
Recalibrated cut scores (S1)	.83	.87	6.38	5.12	.80	.89	.85
Recalibrated cut scores (S2)	.84	.87	6.46	5.44	.81	.89	.86
Winter 1 Assessment (A2)							
Current cut scores (S1 & S2)	.88	.80	4.40	6.67	.56	.96	.82
Recalibrated cut scores (S1)	.93	.75	3.72	10.71	.49	.99	.78
Recalibrated Cut Scores (S2)	.92	.78	4.18	9.75	.49	.98	.81
Winter 2 Assessment (A3)							
Current cut scores (S1 & S2)	.91	.81	4.79	9.00	.56	.97	.83
Recalibrated cut scores (S1)	.94	.82	5.22	13.67	.53	.99	.84
Recalibrated Cut Scores (S2)	.95	.81	5.00	16.20	.53	.99	.83

Note. S1=Calibration sample, S2=Cross-validation sample, LR+=Positive Likelihood Ratio, LR-=Negative Likelihood Ratio, PPP=Positive Predictive Power, NPP=Negative Predictive Power, OCC=Overall Correct Classification.

risk, students identified at some risk on ORF at the beginning of the year using the recalibrated cut scores in both samples were more likely to be identified not at risk (S1 = 59%; S2 = 58%) on FCAT-SSS (Table 5). Although the overall frequency distribution

Table 4
Cut score descriptive statistics for ORF

	Cut scores	Risk status	% met BM	Frequency (%)
Fall Assessment (A1)				
Current	<53	At risk	20.2	32.5
	53-76	Some risk	61.9	28.8
	77+	Low risk	85.9	38.7
Recalibrated	<45	At risk	17.4	22.9
	45-79	Some risk	59.2	41.4
	80+	Low risk	86.8	35.7
Winter 1 Assessment (A2)				
Current	<62	At risk	11.6	16.8
	62-87	Some risk	44.8	25.8
	88+	Low risk	79.7	57.4
Recalibrated	<55	At risk	7.8	13.1
	55-84	Some risk	39.8	25.8
	85+	Low risk	73.5	61.1
Winter 2 Assessment (A3)				
Current	<70	At risk	8.5	15.2
	70-97	Some risk	40.6	27.8
	98+	Low risk	80.9	57.0
Recalibrated	<65	At risk	7.8	12.5
	66-97	Some risk	41.5	31.8
	98+	Low risk	80.9	55.7

Note. % met BM is the percentage of students who met the 3rd grade end of year benchmark goal (FCAT-SSS level ≥ 3).

Table 5
Contingency table for FCAT-SSS risk prediction from ORF 1 scores

Oral Reading Fluency, Fall Assessment (A1)	FCAT performance			
	Not at risk	At risk	Total	
Low risk				
	Calibration (S1)	5233	654	5887 (35.0%)
	Cross-validation (S2)	5082	631	5713 (34.1%)
Some risk				
	Calibration (S1)	4138	2845	6983 (41.6%)
	Cross-validation (S2)	3985	2863	6848 (41.4%)
High risk				
	Calibration (S1)	798	3105	3903 (23.3%)
	Cross-validation (S2)	757	3234	3991 (24.1%)
Total				
	Calibration (S1)	10,169 (60.1%)	6604 (39.9%)	16,773
	Cross-validation (S2)	9824 (59.4%)	6728 (40.6%)	16,552

of students in risk categories for recalibrated scores was centered more towards the some risk students, the improved diagnostic accuracy in all estimated indices provided justification for retaining the suggested adjustments.

Winter 1 Assessment

Similar adjustments to the cut scores were made to the ORF Winter 1 Assessment. Results indicated that lower cut scores of <55 for the at risk criteria and >84 for low risk improved the classification of students when compared to the current benchmarks (<62 at risk, >87 low risk). The specificity of both S1 (.75) and S2 (.78) were smaller in magnitude than using the original scores (.80; Table 3), also directly affecting the overall correct classification; however, the improvement in sensitivity (Current_{S1&S2} = .88; Recalibrated_{S1} = .93; Recalibrated_{S2} = .92) and subsequent reduction in the rate of false negatives (Current_{S1&S2} = .12; Recalibrated_{S1} = .07; Recalibrated_{S2} = .08) provided a balance to the overall utility of the recalibrated adjustments.

The percentages of students meeting the end of year goal in the combined sample for low risk (73.5%), some risk (39.8%), and at risk (7.8%) using the recalibrated scores were lower than observed when the current recommended cutoffs were used in our sample (79.7%, 44.8%, 11.6%). The overall distribution of students across risk levels using the FCRR cut scores was comparable to results using the current cut scores (Table 4). Compared to the Fall Assessment findings, students who were classified as at some risk on ORF using the recalibrated cut scores after the Winter 1 Assessment were more likely to be categorized as at risk in both samples (S1 = 60%; S2 = 60%) on the FCAT-SSS (Table 6).

Winter 2 Assessment

In examining the distribution in our sample, the adjustment of the cut scores for the ORF Winter 2 Assessment indicated that the best fit was found when a more liberal estimate was applied to the at risk group (<65) and the low risk threshold was kept

Table 6
Contingency table for FCAT-SSS risk prediction from ORF 2 scores

Oral Reading Fluency, Winter 1 Assessment (A2)	FCAT performance		
	Not at risk	At risk	Total
Low risk			
Calibration (S1)	7733	163	7896 (49.0%)
Cross-validation (S2)	7441	164	7605 (48.1%)
Some risk			
Calibration (S1)	1656	2508	4164 (25.8%)
Cross-validation (S2)	1642	2508	4150 (26.3%)
High risk			
Calibration (S1)	2120	1934	4054 (25.2%)
Cross-validation (S2)	2063	1986	4049 (25.6%)
Total			
Calibration (S1)	11,509 (71.4%)	4605 (28.6%)	16,114
Cross-validation (S2)	11,146 (70.5%)	4658 (29.5%)	15,804

constant (>97). Even though the criteria were only adjusted by five points for the at risk category, the diagnostic efficiency improved across all indices in the calibration group and all but the specificity and overall correct classification in the cross-validation sample (Table 3). The percentages of students meeting the end of year goal in the combined sample for low risk (80.9%), some risk (41.5%), and at risk (7.8%) using the recalibrated cut scores were closer to the percentages observed implementing the current recommended benchmarks (80.9%, 40.6%, 8.5%). The distribution of students identified as at some risk for the ORF Winter 2 Assessment was similar to findings from the Fall Assessment (i.e., greater propensity toward end of year at risk classification) for both samples (S1 = 58%; S2 = 59%; Table 7).

Table 7
Contingency table for FCAT-SSS risk prediction from ORF 3 scores

Oral Reading Fluency, Winter 2 Assessment (A3)	FCAT performance		
	Not at risk	At risk	Total
Low risk			
Calibration (S1)	7747	123	7870 (46.8%)
Cross-validation (S2)	7512	111	7623 (46.2%)
Some risk			
Calibration (S1)	2128	2992	5120 (30.5%)
Cross-validation (S2)	2051	2992	5043 (30.6%)
High risk			
Calibration (S1)	1840	1977	3817 (22.7%)
Cross-validation (S2)	1801	2024	3825 (23.2%)
Total			
Calibration (S1)	11,715 (69.7%)	5092 (30.3%)	16,807
Cross-validation (S2)	11,364 (68.9%)	5127 (31.1%)	16,491

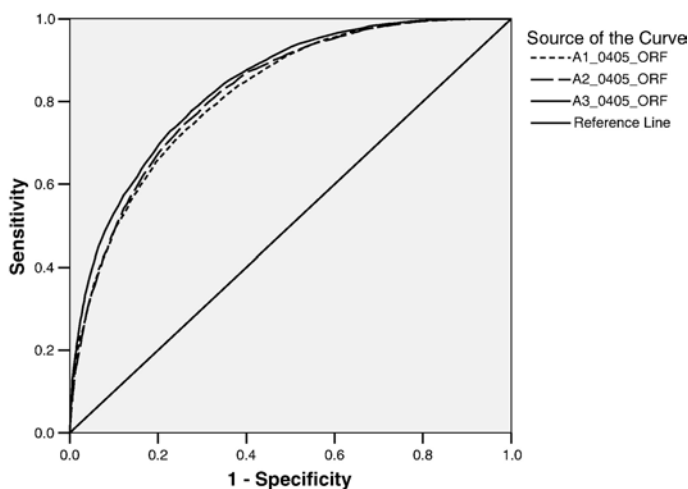


Fig. 1. ROC curve for calibration sample.

Post hoc analysis

A *post hoc* analysis was conducted to test which of the three ORF administrations was predictive of FCAT-SSS risk. Results indicated the ORF Winter 2 Assessment had the highest AUC index for both the calibration and cross-validation samples (.840, .841) and was significantly better than ORF at the Fall (S1 [.821, $z=9.77$]; S2 [.825, $z=7.00$]) and Winter 1 Assessments (S1 [.826, $z=7.04$]; S2 [.828, $z=6.53$]). The ROC curves for the three assessments in both samples are presented in Figs. 1 and 2.

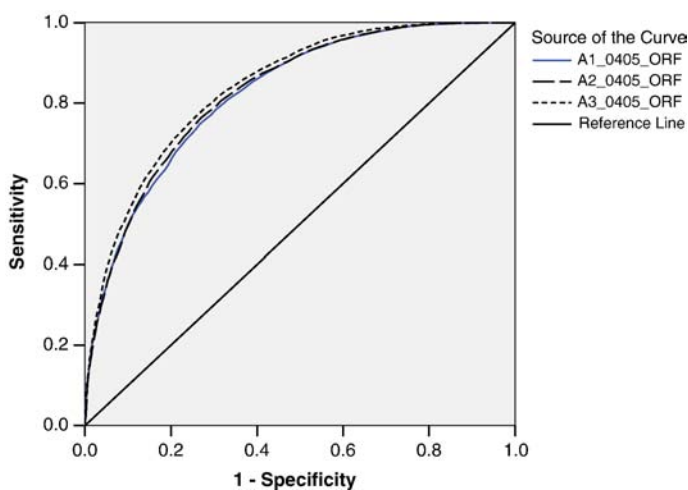


Fig. 2. ROC curve for cross-validation sample.

Table 8
Percent of students meeting end of year benchmark by ORF risk-level classification

Assessment	Subgroups	<i>n</i>	At risk	Some risk	Low risk
Fall Assessment	ELL				
	Non-ELL	26,672	17.7	60.6	87.7
	ELL	3658	12.3	42.7	71.4
	FRL				
	Non-FRL	8495	23.8	68.1	93.2
	FRL	24,803	15.5	55.9	83.6
	Ethnicity				
	White	12,175	23.3	69.7	92.6
	African American	12,107	12.4	49.4	79.3
Latino	7244	15.8	57.0	84.9	
Winter 1 Assessment	ELL				
	Non-ELL	25,462	7.9	41.4	79.8
	ELL	3596	5.5	29.1	59.4
	FRL				
	Non-FRL	8040	11.7	47.5	87.5
	FRL	23,853	6.9	37.8	74.5
	Ethnicity				
	White	11,532	11.1	51.2	87.0
	African American	11,696	5.1	31.4	69.1
Latino	6985	8.0	36.9	75.8	
Winter 2 Assessment	ELL				
	Non-ELL	26,621	6.3	43.0	82.4
	ELL	3687	2.5	29.1	61.2
	FRL				
	Non-FRL	8488	9.9	50.6	88.4
	FRL	24,783	4.7	38.8	77.3
	Ethnicity				
	White	12,157	8.9	51.8	88.6
	African American	12,076	4.1	32.9	75.3
Latino	7267	4.1	39.9	77.1	

Predictive bias

The third research question addressed differences among the subgroups with regard to the percentage of students who met the end of year benchmark according to ORF risk classification (see Table 8). Within each subgroup, the first reported group represented the referent in the logistic regression. Results from the descriptive classification table indicated that the referent in each subgroup (i.e., not eligible for free or reduced-price lunch, not an English Language Learner, White) had a greater percentage of its students meeting the end of year benchmark on FCAT-SSS by ORF risk-level classification than the subgroups traditionally considered at risk. It also was important, however, to test if the predictive accuracy of ORF was dependent on the demographic characteristics of poverty, language, and race/ethnicity, with a significant interaction evident of ORF bias in classification.

Results from the Free or Reduced-Priced Lunch Status (FRL) model with three predictors (FRL, ORF, FRL × ORF) revealed the presence of significant FRL ($z^2=5.80$, $p<.05$) and ORF effects ($z^2=382.87$, $p<.001$), with a large odds ratio for ORF (66.13)

Table 9
Logistic regression for classification bias

	β	SE	Wald	p -value	e^{β}
Model 1					
ELL	.83	.45	3.34	.065	2.30
Winter 2 ORF	3.99	.44	80.94	.000	53.95
ELL \times Winter 2 ORF	.13	.23	.30	.582	1.14
Model 2					
FRL	.76	.32	5.80	.016	2.14
Winter 2 ORF	4.19	.21	382.87	.000	66.13
FRL \times Winter 2 ORF	.02	.16	.02	.883	1.02
Model 3					
A.A.	.55	.32	2.97	.085	.58
Winter 2 ORF	4.38	.11	1596.38	.000	80.05
A.A. \times Winter 2 ORF	.27	.17	2.67	.102	.76
Model 4					
Lat.	.80	.37	4.80	.028	.45
Winter 2 ORF	4.38	.11	1596.38	.000	80.05
Lat. \times Winter 2 ORF	.02	.19	.01	.925	.98

Note. A.A. = African American, Lat. = Latino.

confirming the predictive nature of oral reading fluency. No significant FRL \times ORF effect was observed ($z^2 = .02$, $p > .05$).

Testing of the English Language Learner Status (ELL) model with three predictors (ELL, ORF, ELL \times ORF) indicated a strong ORF effect was present ($z^2 = 80.94$, $p < .001$). When converted to an odds ratio, students who were designated as *at risk* on ORF were 53.95 times more likely to be at risk on the FCAT-SSS than students who were identified as *not at risk* on ORF, supporting the hypothesis that ORF scores significantly predict FCAT-SSS reading scores. No significant ELL ($z^2 = 3.34$, $p > .05$) or ELL \times ORF ($z^2 = .30$, $p > .05$) effects were observed.

Race/ethnicity was tested by individually comparing African American and Latino students with White students. Significant main effects for ORF were observed for both comparisons ($z^2 = 1596.38$, $p < .001$), with a small observed main effect for Latino students ($z^2 = 4.80$, $p < .05$) but not African American students ($z^2 = 2.97$, $p > .05$). This suggested that Latino students were .45 times more likely than White students to be at risk according to the FCAT-SSS. While African American students were descriptively more likely to be at risk than White students (.58), this difference was not significantly different ($p = .085$).

The interactions for both groups were not significant ($z^2 = -.27$, $p > .05$; $z^2 = -.02$, $p > .05$). Because the interactions between the demographic subcategories with the Winter 2 ORF Assessment were not significant and the most significant predictor of risk on FCAT-SSS was ORF, ORF did an equally good job identifying at risk readers regardless of their demographic characteristics. A summary of results for the logistic regressions is presented in Table 9.

Discussion

The purpose of the current study was to examine the relationship between DIBELS *Oral Reading Fluency* and reading comprehension achievement on the FCAT-SSS and the SAT-

10 measures of reading comprehension, and to estimate the predictive and concurrent validity of using ORF to identify students who are at risk for below grade-level reading achievement. In addition, ORF was evaluated for any predictive bias across different demographic subgroups. DIBELS ORF is currently used as a measure of students' progress in reading in all *Reading First* schools in the state of Florida and many other states, thus an analysis of the diagnostic accuracy of the current standards in an underachieving and particularly diverse population is necessary.

Performance of ORF with SAT-10 and FCAT-SSS

The correlations of ORF with both FCAT-SSS and SAT-10 were high ($r_s = .70-.71$) and consistent with previous findings about the relationships between oral reading fluency and reading comprehension (Good et al., 2001; Buck & Torgesen, 2003; Schilling, Carlisle, Scott, & Zeng, 2007). As expected, the relationships between the third administration (Winter 2 Assessment) for ORF and both FCAT-SSS and SAT-10 were the strongest observed correlations in both samples, corresponding to the concurrent time interval the tests were given. Thus, DIBELS is related equally well to a common measure of reading comprehension used across states as it is to a state-developed measure.

Recalibration of risk levels

ROC curve results for recalibrating the ORF risk-level cut scores suggested that a greater proportion of students identified as true positives (i.e., at risk and not at risk) could be identified using the recalibrated scores than the currently implemented standards. A reciprocal benefit of improving the sensitivity and specificity of diagnostic accuracy is a reduction in the number of students identified as not at risk by FCAT-SSS but at risk on ORF (false-positives) and the number of students identified as at risk by FCAT but not at risk on ORF (false-negatives). Across all assessments, sensitivity improved by an average of +5.8% with the highest gains observed during the Fall Assessment (+10%). Although specificity only slightly improved across most assessments ($\overline{SP} = +1.7\%$), this is largely due to the drop in efficiency during the Winter 1 Assessment; however, the reciprocal increase in false positive rates indicated that increased over-identification may occur during this time, but not at a significantly different rate than is already estimated using the current standards. With the adjusted cut scores, though there is a risk of over-identifying at risk students, improved efficiencies decrease the likelihood of missing students who truly are at risk (i.e., there is a decreased chance of excluding them from interventions), which counterbalances the risks of over-identification.

Negative predictive power estimates were larger across all assessments for the recalibrated scores than the current indicators of risk, reaching near perfect predictive power. Conversely, the positive predictive power, although larger for the recalibrated cut score during the Fall Assessment, was in the moderate range across all assessments. These findings reflect similar patterns in predictive power estimated by Hintze et al. (2003), who assessed diagnostic accuracy of DIBELS subtests other than ORF and found low to moderate positive predictive power but near unity negative predictive power with the criterion measure. Current findings indicated that while a student identified as low risk on

ORF may reliably be considered truly not at risk, a positive at risk identification on ORF has only moderate consistency in its ability to identify at risk readers. Although the degree of confidence one is able to have in the predictive reliability of ORF is improved using the recalibrated scores, the magnitude of the positive predictive power at the Winter 1 and 2 Assessments is below a standard threshold of .75 and limits the accuracy in identifying students who are at risk readers.

Nevertheless, the strengths of the recalibrated scores suggest that the implementation of new standards may prove beneficial to educators in the state of Florida. Congruence between the proportion of students in each risk category meeting the end of year benchmark in our sample and Good et al.'s (2002) was observed. A portion of Good et al.'s guidelines in developing risk scores was to have at least 80% of students who were at grade level (i.e., low risk) meeting end of year goal, and 20% or fewer of students who were identified as being at risk to achieve a similar outcome. With the exception of the proportion of students who were low risk during the Winter 1 Assessment (<80%), the recalibrated scores met both criteria for the initial standards across each assessment. Furthermore, across all indices of efficiency, the recalibrated scores generally provide cross-validated evidence of improved classification. Although the positive predictive power values for the Winter 1 and Winter 2 Assessments are moderate in their magnitude, both positive predictive power and negative predictive power are dependent on the base rate of risk in the sample. Since ORF in the context of *Reading First* is used for screening students who are at disproportionate risk for reading difficulties, it is more important to identify the efficiency measures that are not affected by base rates (i.e., sensitivity, specificity, LR+, LR-, AUC). Given these unaffected indices, there is strong evidence that the recalibrated scores capture the students who are truly at risk and truly not at risk.

Analysis of predictive bias

There was no evidence of predictive bias across several demographic groups in the logistic regression analyses. The most significant predictor of risk on FCAT-SSS was ORF, and the interactions between races/ethnicities, levels of socio-economic status, and language status with ORF were not significant contributors to FCAT-SSS risk. Hence, the ability of ORF to accurately identify at risk readers of varying demographic characteristics also was not significantly different. Our findings are consistent with those of Hintze et al. (2002) and Hixson and McGlinchey (2004), in that oral reading fluency measured by ORF is not biased in the prediction of comprehension outcomes for African American and White students. Our results reveal that ORF also is not biased in predictions for Latino students. The finding that ORF was not biased in predicting the reading comprehension outcomes of English Language Learners, however, contradicts those of Klein and Jimerson (2005). They found predictive bias in the SAT-9 outcomes of Latino students with Spanish as their home language (compared to White students with English as their home language) using oral reading fluency probes (other than DIBELS ORF). They found no bias, however, when comparing Latino and White students, both with English as their home language. In our study, the data set did not provide information on which home languages were spoken, collapsing into one group all students who were identified as English Language Learners regardless of the language spoken at home.

Policy implications

If DIBELS ORF is considered a precise short assessment that predicts later reading comprehension performance (on both the FCAT-SSS and the widely-used SAT-10, as demonstrated by this study), then its use in relation to appropriate risk-level cut scores is critical to the monitoring of student learning and subsequent adjustment of instruction. From a policy perspective, it is essential that educators be provided with precise student achievement data and benchmarks if the rigorous grade-level reading standards set forth in accountability policies are to be met by all students. The FCAT-SSS is often regarded among educators as one of the more difficult of the state assessments used to measure grade-level performance in reading. As such, evaluating the utility of DIBELS ORF data for monitoring learning and predicting FCAT-SSS outcomes, which serves as a gateway for promotion to fourth grade, is particularly important. A general limitation, however, is that the recalibrated cut scores for ORF are only as good as the cut scores on the FCAT. While there is no public information about the basis for FCAT cut scores, they have large practical implications for children's progression. Therefore, looking at how well we can predict students' achievement on the different FCAT levels has implications for the validity of the use of such measures, particularly as it relates to social consequences (Messick, 1989).

Cut scores specific to the Florida *Reading First* population could be used to provide progress monitoring data tailored to Florida's unique population, thus improving the precision of information provided to teachers and administrators for programmatic decision-making. Further, even small improvements in the precision of ORF cut scores for the Florida *Reading First* population would result in the correct identification of many students due to the extremely large population size. Other states with similar student demographics may also find it useful to make similar adjustments to their cut scores, not only because of the specific features of their student populations, but also because of the specific characteristics of their outcome tests for reading comprehension. For instance, it is likely that states such as California, Texas and New York could realize substantial improvements in the accuracy of student classification that would result in many students being served with additional precision. Additionally, when considering the appropriateness of DIBELS ORF for widespread use in a large population of diverse students, it is important to determine if demographic variables have an impact on the utility of the measure. By showing that there is no demographic bias inherent in DIBELS ORF, our results further substantiate the validity of this measure within the Florida *Reading First* population.

Our findings still need to be replicated in a new sample of third graders to further validate the recalibrated cut scores. As a part of this validation, we also plan to examine how the recalibrated cut scores function for all third graders, not just those from *Reading First* schools. Without doing so, the generalizability of the results are limited; mainly due to the impact of base rates in the population. Given the proportion of all third grade students who are identified as at risk on the FCAT, some of the diagnostic efficiency indices (i.e., PPP and NPP) would change. We would expect that a higher prevalence rate in the population would result in larger PPP and smaller NPP values, while lower base rates would result in smaller PPP and larger NPP values.

Conclusion

Even given the strong predictive utility of using the *Oral Reading Fluency* progress monitoring measure to identify students at risk for poor performance on the FCAT-SSS reading comprehension measure, the promising results of the newly recalibrated cut scores for the DIBELS ORF in Florida, and the lack of predictive bias across various demographic groups, research remains to be done in the area of evaluating the implications of using the cut scores in school contexts. In particular, we agree with Pressley (2006) that more research is needed on the impact of DIBELS mandates on teachers' instruction and student motivation and learning.

References

- Allington, R. L. (1983). Fluency: The neglected reading goal. *The Reading Teacher*, 36, 556–561.
- Baker, S. K., & Good, R. (1995). Curriculum-based measurement of English reading with bilingual Hispanic students: A validation study with second-grade students. *School Psychology Review*, 24, 561–578.
- Barger, J. (2003). *Comparing the DIBELS Oral Reading Fluency indicator and the North Carolina end of grade reading assessment* (technical report). Ashville, NC: North Carolina Teacher Academy.
- Breznitz, Z. (1987). Increasing first graders' reading accuracy and comprehension by accelerating their reading rates. *Journal of Educational Psychology*, 79, 236–242.
- Breznitz, Z. (1991). The beneficial effect of accelerating reading rate on dyslexic readers' reading comprehension. In M. Snowling & M. Thomson (Eds.), *Dyslexia: Integrating theory and practice* (pp. 235–243). London: Whurr.
- Buck, J., & Torgesen, J. (2003). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test* (Tech. Rep. No. 1). Tallahassee, FL: Florida Center for Reading Research. Available at <http://www.fcrr.org/technicalreports/TechnicalReport1.pdf>
- Chard, D. J., Vaughn, S., & Tyler, B. J. (2002). A synthesis of research on effective interventions for building reading fluency with elementary students with learning disabilities. *Journal of Learning Disabilities*, 35, 386–406.
- Children's Educational Services (1987). *Test of Reading Fluency*. Minneapolis, MN: Author.
- CTB/McGraw-Hill (2004). *TerraNova*, The (2nd ed.) (CAT/6) Monterey, CA: CTB/McGraw-Hill.
- Daane, M.C., Campbell, J.R., Grigg, W.S., Goodman, M.J., & Oranje, A. (2005). *Fourth-Grade Students Reading Aloud: NAEP 2002 Special Study of Oral Reading* (NCES 2006-469). U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. Washington, DC: Government Printing Office.
- Deno, S. L. (1989). Curriculum-based measurement and special education services: A fundamental and direct relationship. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 1–17). New York: The Guildford Press.
- Deno, S. L., Marston, D., Shinn, M. R., & Tindal, G. (1983). Oral reading fluency: A simple datum for scaling reading disability. *Topics in Learning and Learning Disabilities*, 2(4), 53–59.
- Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, 49, 36–45.
- Dowhower, S. L. (1987). Effects of repeated reading on second-grade transitional readers' fluency and comprehension. *Reading Research Quarterly*, 22, 389–406.
- Dowhower, S. L. (1991). Speaking of prosody: Fluency's unattended bedfellow. *Theory into Practice*, 30, 165–175.
- Educational Data Systems (2005). *The California reading first year 3 evaluation report*. San Francisco, CA: Author.
- 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.
- Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review*, 21, 45–58.

- Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20–28.
- Fuchs, L. S., Fuchs, D., Hosp, M. D., & Jenkins, J. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*, 239–259.
- Good, R. H., & Kaminski, R. A. (1996). Assessment for instructional decisions: Toward a proactive/prevention model of decision-making for early literacy skills. *School Psychology Quarterly, 11*, 326–336.
- Good, R. H., Kaminski, R. A., Smith, S., & Bratten, J. (2001). *Technical adequacy of second grade DIBELS oral reading fluency passages* (Technical Report, No. 8). Eugene, OR: University of Oregon.
- Good, R. H., Kaminski, R. A., Smith, S., Laimon, D., & Dill, S. (2001). *Dynamic Indicators of Basic Early Literacy Skills* (5th Ed.). Eugene: University of Oregon.
- Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5*, 257–288.
- Good, R. H., Wallin, J., Simmons, D. C., Kame'enui, E. J., & Kaminski, R. A. (2002). *Systemwide percentile ranks for DIBELS benchmark assessment* (Technical Report, No. 9). Eugene, OR: University of Oregon.
- Goodman, K. (2006). A critical review of DIBELS. In K. Goodman (Ed.), *Examining DIBELS: what it is and what it does* (pp. 1–32). Brandon, VT: Vermont Society for the Study of Education.
- Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristics curves derived from the same cases. *Radiology, 148*, 839–843.
- Harcourt Brace (2003a). *Stanford Achievement Test* (10th ed.). San Antonio, TX: Author.
- Harcourt Brace (2003b). *Stanford Achievement Test* (10th ed.): *Technical data report*. San Antonio, TX: Author.
- Hasbrouck, J., & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59*, 636–644.
- Hintze, J. M., Callahan, J. E., III, Matthews, W. J., Williams, S. A. S., & Tobin, K. G. (2002). Oral reading fluency and prediction of reading comprehension in African American and Caucasian elementary school children. *School Psychology Review, 31*, 540–553.
- Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the dynamic indicators of basic early literacy skills and the comprehensive test of phonological awareness. *School Psychology Review, 32*, 541–556.
- Hixson, M. D., & McGlinchey, M. T. (2004). The relationship between race, income, and oral reading fluency and performance on two reading comprehension measures. *Journal of Psychoeducational Assessment, 22*, 351–364.
- Hudson, R. F., Lane, H. B., & Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher, 58*, 702–714.
- Kame'enui, E. J., Fuchs, L., Francis, D. J., Good III, R., O'Connor, R. E., Simmons, D. C., et al. (2006). The adequacy of tools for assessing reading competence: A framework and review. *Educational Researcher, 35*, 3–11.
- Klein, J. R., & Jimerson, S. R. (2005). Examining ethnic, gender, language, and socioeconomic bias in oral reading fluency scores among Caucasian and Hispanic students. *School Psychology Quarterly, 20*, 23–50.
- Marston, D. B. (1989). A curriculum-based measurement approach to assessing: What is it and why do it? In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18–78). New York: The Guildford Press.
- Mathson, D. V., Allington, R. L., & Solic, K. L. (2005). Hijacking fluency and instructionally informative assessments. In T. Rasinski, C. Blachowitz, & K. Lems (Eds.), *Fluency instruction: Research-based best practices* (pp. 106–119). New York: Guilford.
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 3*, 195–216.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*, 5–11.
- Meyer, M. A., & Felton, R. H. (1999). Repeated reading to enhance fluency: Old approaches and new directions. *Annals of Dyslexia, 49*, 283–306.
- National Reading Panel (2000). *Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: NIH.
- Perfetti, C. A., & Hogaboam, T. (1975). Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology, 67*, 461–469.

- Pressley, M. (2006, April). *What the future of reading research could be*. A paper presented at the International Reading Association's Reading Research, Chicago, Illinois.
- Pressley, M., Hilden, K., & Shankland, R. (in press). An evaluation of end-grade-3 Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Speed reading without comprehension, predicting little. Retrieved October 1, 2006, from the (NOT the Official) DIBELS Clearinghouse Web site: <http://vsse.net/dibels/node/73>
- Rasinski, T. V. (1989). Fluency for everyone: Incorporating fluency instruction in the classroom. *The Reading Teacher*, 42, 690–693.
- Rasinski, T. V. (1990). Investigating measures of reading fluency. *Educational Research Quarterly*, 14(3), 37–44.
- Rasinski, T. (2006). Reading fluency instruction: Moving beyond accuracy, automaticity, and prosody. *The Reading Teacher*, 59, 704–706.
- Rasinski, T., Blachowitz, C., & Lems, K. (Eds.). (2006). *Fluency instruction: Research-based best practices*. New York: Guilford.
- Reynolds, C. R. (2000). Methods for detecting and evaluating cultural bias in neuropsychological tests. In E. Fletcher-Janzen, T. L. Strickland & C.R. Reynolds (Eds.), *Handbook of cross-cultural neuropsychology* (pp. 249–285). New York: Kluwer Academic/Plenum Publishers.
- Samuels, S. J. (2006a). Reading fluency: Its past, present, and future. In T. Rasinski, C. Blachowitz & K. Lems (Eds.), *Fluency instruction: Research-based best practices* (pp. 7–20). New York: Guilford.
- Samuels, S. J. (2006b). Toward a model of reading fluency. In S. J. Samuels & A.E. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 24–46). Newark, DE: International Reading Association.
- Samuels, S. J., & Farstrup, A. E. (2006). *What research has to say about fluency instruction*. Newark, DE: International Reading Association.
- 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). Tallahassee, FL: Florida Center for Reading Research. Available at http://www.fcrr.org/TechnicalReports/Multi_variate_study_december2004.pdf
- Schreiber, P. A. (1980). On the acquisition of reading fluency. *Journal of Reading Behavior*, 7, 177–186.
- Schreiber, P. A. (1991). Understanding prosody's role in reading acquisition. *Theory into Practice*, 30, 158–164.
- Schwanenflugel, P. J., Hamilton, A. M., Kuhn, M. R., Wisenbaker, J. M., & Stahl, S. A. (2004). Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers. *Journal of Educational Psychology*, 96, 119–129.
- Shilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *The Elementary School Journal*, 107, 429–448.
- Shaw, R., & Shaw, D. (2002). *DIBELS Oral Reading Fluency-based indicators of third grade reading skills for Colorado State Assessment Program (CSAP)* (Technical Report). Eugene, OR: University of Oregon.
- Shinn, M. R. (1989). Identifying and defining academic problems: CBM screening and eligibility procedures. In M. R. Shinn (Ed.), *Curriculum based measurement: Assessing special children* (pp. 90–129). New York: The Guilford Press.
- Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review*, 21, 459–479.
- Silberglitt, B., & Hintze, J. M. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment*, 23, 304–325.
- Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic screening tests. *Journal of Personality Assessment*, 81, 209–219.
- Swets, J. A. (1988). Measuring the diagnostic accuracy of diagnostic systems. *Science*, 240, 1285–1293.
- Tenenbaum, H. A., & Wolking, W. D. (1989). Effects of oral reading rate on intraverbal responding. *The Analysis of Verbal Behavior*, 7, 83–89.
- Tindal, G. A., & Marston, D. B. (1990). *Classroom-based assessment: Evaluating instructional outcomes*. Columbus, OH: Merrill Publishing Company.
- Torgesen, J. K., Rashotte, C. A., & Alexander, A. (2001). Principles of fluency instruction in reading: Relationships with established empirical outcomes. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 333–355). Parkton, MD: York Press.
- University of Oregon Center on Teaching and Learning (2006). *DIBELS benchmark levels*. Eugene, OR: Author Retrieved on August 2, 2006, from <http://dibels.uoregon.edu/benchmarkgoals4x.pdf>

- U.S. Department of Education (2006a). *Proven methods: Early reading first and reading first*. Retrieved October 30, 2006, from <http://www.ed.gov/nclb/methods/reading/readingfirst.html>
- U.S. Department of Education (2006b). *Reading first implementation evaluation: Interim report*. Retrieved August 3, 2006, from <http://www.ed.gov/about/offices/list/oepd/ppss/index.html>
- Vander Meer, C. D., Lentz, F. E., & Stollar, S. (2005). *The relationship between oral reading fluency and Ohio proficiency testing in reading* (Technical Report). Eugene, OR: University of Oregon.
- Wilson, J. (2005). *The relationship of Dynamic Indicators of Basic Early Literacy Skills (DIBELS) oral reading fluency to performance on Arizona Instrument to Measure Standards (AIMS)* (Technical Report). Tempe, AZ: Assessment and Evaluation Department, Tempe School District No.3.