11-1-2017

Developing Local Oral Reading Fluency Cut Scores for Predicting High-Stakes Test Performance

Sally Grapin
Montclair State University, grapins@montclair.edu

John H. Kranzler
University of Florida

Nancy Waldron
University of Florida

Diana Joyce-Beaulieu
University of Florida

James Algina
University of Florida

Received: 24 October 2016 | Revised: 13 March 2017 | Accepted: 16 March 2017

DOI: 10.1002/pits.22035

WILEY

RESEARCH ARTICLE

Developing local oral reading fluency cut scores for predicting high-stakes test performance

Sally L. Grapin[1] | John H. Kranzler[2] | Nancy Waldron[2] | Diana Joyce-Beaulieu[2] | James Algina[2]

[1]Montclair State University

[2]University of Florida

Correspondence
Sally L. Grapin, Department of Psychology, Montclair State University, 221 Dickson Hall, 1 Normal Avenue, Montclair, NJ 07043.
Email: grapins@mail.montclair.edu

Abstract

This study evaluated the classification accuracy of a second grade oral reading fluency curriculum-based measure (R-CBM) in predicting third grade state test performance. It also compared the long-term classification accuracy of local and publisher-recommended R-CBM cut scores. Participants were 266 students who were divided into a calibration sample ($n = 170$) and two cross-validation samples ($n = 46$; $n = 50$), respectively. Using calibration sample data, local fall, winter, and spring R-CBM cut scores for predicting students' state test performance were developed using three methods: discriminant analysis (DA), logistic regression (LR), and receiver operating characteristic curve analysis (ROC). The classification accuracy of local and publisher-recommended cut scores was evaluated across subsamples. Only DA and ROC produced cut scores that maintained adequate sensitivity (≥.70) across cohorts; however, LR and publisher-recommended scores had higher levels of specificity and overall correct classification. Implications for developing local cut scores are discussed.

KEYWORDS

curriculum-based measurement, local benchmarks, universal screening

The ability to provide effective early reading intervention is contingent on the use of efficient and accurate screening procedures. Often, screeners often are used to predict students' performance on a particular "gold standard" or outcome measure. The term "gold standard" is widely used in the classification analysis literature to describe a criterion or outcome measure that "represents the best possible way to know if a condition is truly present" (VanDerHeyden, 2010, p. 282). In schools, the "condition" of interest typically is an undesirable academic outcome, such as failure to attain reading proficiency (VanDerHeyden, 2011). For students in the latter elementary grades (i.e., grades 3–5), schools often select standardized tests of reading comprehension as their "gold standard" or "outcome" measure, given the emphasis of curricula in these grades on higher order reading skills.

Schools may use a variety of screening instruments to predict later academic outcomes. In particular, a vast array of research has explored the relationships of students' early cognitive abilities (e.g., phonological processing

and nonverbal reasoning abilities) and oral language skills to their later reading comprehension outcomes (Adolf, Catts, & Lee, 2010; Fuchs et al., 2012; Scarborough, 1998). For example, Fuchs et al. (2012) observed that students' scores on measures of phonological processing, nonverbal reasoning, and oral language measures in first grade correlated moderately with their scores on a measure of reading comprehension in fifth grade (correlations ranging from .43 to .56). Similarly, Adolf et al. (2010) found that students' performance on measures of nonverbal intelligence, picture vocabulary, and oral vocabulary in kindergarten correlated moderately with both their second grade and eighth grade reading comprehension performance (correlations ranging from .45 to .56). Overall, this research has identified phonological awareness, rapid naming, verbal working memory, and oral/language vocabulary as among the best cognitive predictors of later reading failure (Fletcher, Lyon, Fuchs, & Barnes, 2007; Steubing et al., 2015).

Additionally, a number of scholars have explored the relationships between students' basic reading skills and their later reading comprehension outcomes, arguing that cognitive measures have relatively less utility for treatment planning and predicting intervention response (e.g., Steubing et al., 2015). In particular, many of these studies have investigated the predictive utility of curriculum-based measures (CBMs) for identifying students at risk for poor performance on statewide tests of reading achievement. CBMs are brief, standardized assessments of academic skills that can be used to monitor students' progress over time. Although CBMs measuring a variety of early reading skills (e.g., knowledge of letter names and the alphabetic principle) can be used to predict reading outcomes, oral reading fluency measures (R-CBMs) have been shown to be the best indicators of overall reading competence and, in turn, the best predictors of students' state reading test performance (Goffreda & DiPerna, 2010). Previous research indicates that moderate to strong positive correlations (typically ranging from .60 to .90) exist between students' performance on early grade R-CBMs and state reading, depending on test and sample characteristics, duration between screener and state test administration, and other variables (Wood, 2006; Yeo, 2010).

R-CBM screeners can be used to classify students into one of two groups based on risk status: *test positives* (i.e., those who are at risk for poor performance on the gold standard measure) and *test negatives* (i.e., those who are not at risk). This can be accomplished by designating a screening cut score, which represents the critical number of words correct per minute (WCPM) that students must read to demonstrate that they are on track to meeting grade-level standards (Silberglitt & Hintze, 2005). Similarly, on state achievement tests (i.e., gold standard measures), cut scores can also be used to distinguish proficient from nonproficient readers. Outcomes on the screening and gold standard measures yield four different types of classification decisions: *true positives*, *true negatives*, *false positives*, and *false negatives*. High rates of false positives can result in unnecessary intervention referrals, which can inflate student–teacher ratios. Conversely, high rates of false negatives may result in the egregious error of failing to provide intervention to students who are truly at risk.

To evaluate the accuracy and efficiency of screeners in predicting state test performance, school professionals can use a variety of classification agreement indicators (CAIs). CAIs quantify the degree to which a decision based on a screener corresponds to an outcome on a subsequent gold standard measure (VanDerHeyden, 2010). For example, school psychologists can compute metrics such as sensitivity, specificity, and area under the curve (AUC) using an analytic procedure known as receiver operating characteristic curve analysis (ROC). Sensitivity refers to the likelihood that an individual who failed the gold standard test will fall below the specified cut score on the screener, whereas specificity refers to the likelihood that an individual who passes the gold standard test will score at or above the screening cut score. In ROC, the true positive (sensitivity) and false positive (1-specificity) rates are plotted on the *y*- and *x*-axes, respectively, for all possible scores on the screener. The area under the resulting curve describes the overall accuracy of the screener. AUC values range from .5 to 1.0. According to Cummings and Smolkowski (2015), if $A = .50$, the screener correctly predicts outcomes on the gold standard measure 50% of the time, which is no more accurate than flipping a coin. Conversely, if $A = 1.0$, then all students are correctly classified by the screener. Other CAIs include positive predictive power (PPP; probability that a positive test finding is truly positive), negative predictive power (NPP; probability that a negative test finding is truly

negative), and overall correct classification (OCC; percentage of students who are correctly classified as "at-risk" or "not at-risk").

Many published R-CBMs specify benchmarks, or cut scores, for gauging students' reading progress over time. For example, the 6th Edition Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency measure (DORF) (Good & Kaminski, 2002a) specifies recommended benchmarks, or cut scores, for classifying students' performance into one of three categories: "low risk," "some risk," and "at risk." Whereas some studies have indicated that the DORF publisher-recommended cut scores exhibit adequate levels of sensitivity and specificity in identifying students at risk for poor state test performance, others suggest that they produce too many false positives (e.g., Sandberg Patton, Reschly, & Appleton, 2014).

Alternatively, schools may wish to develop local cut scores that are tailored to specific end-of-year tests and student populations. Goffreda, DiPerna, and Pedersen (2009) found that the use of recalibrated, local R-CBM cut scores (rather than publisher-recommended cut scores) may increase classification accuracy in identifying students who are at risk for poor high-stakes test performance. Additionally, several studies have compared various methods of local R-CBM cut score development, including ROC, discriminant analysis (DA), and logistic regression (LR; e.g., Sandberg Patton et al., 2014; Silberglitt & Hintze, 2005). For example, Silberglitt and Hintze (2005) used each of these three methods to develop R-CBM cut scores for identifying first and second grade students who were at risk for poor performance on the third grade Minnesota Comprehensive Assessment. They found that each of these methods produced cut scores that met minimum sensitivity and specificity criteria (i.e., $\geq$.70). Given that the ROC and LR methods are generally most robust to violations of statistical assumptions (e.g., normality and equal variance assumptions) and produced cut scores with the greatest diagnostic accuracy, Silberglitt and Hintze concluded that these two methods were most appropriate for use in schools. Similarly, Sandberg Patton et al. (2007) found that R-CBM cut scores developed via LR and ROC exhibited superior classification accuracy, as compared with publisher-recommended DORF benchmarks, in identifying students at risk for poor performance on the Georgia Criterion-Referenced Competency Tests. These studies suggest the promise of using local cut scores for the early identification of students who are at risk for long-term reading difficulties.

1.1 | Limitations of previous research and aims of the current study

Most of the studies described above examined the diagnostic utility of R-CBM screeners in predicting outcomes on state tests administered in the same academic year. To date, relatively few studies have examined the diagnostic utility of screening measures administered more than 1 year in advance of the state test. Early prediction of students' state test performance is important, as these tests typically are administered for the first time at the end of third grade (by which time a critical window for early intervention has passed). Thus, one goal of this study was to examine the utility of second grade R-CBM screeners in predicting performance on a state reading test administered the following academic year.

Additionally, prior research has not examined the long-term classification accuracy of local cut scores over multiple years and student cohorts. Cross-validation of local cut scores on consecutive student cohorts is necessary to ensure the accuracy of resulting CAI values (VanDerHeyden, 2011). Therefore, the second purpose of this study was to compare the long-term classification accuracy of locally developed R-CBM cut scores with that of publisher-recommended cut scores. Specifically, we used cross-validation to examine the classification accuracy of local and publisher-recommended DORF cut scores in predicting state test performance across multiple student cohorts. We also examined how three statistical methods for cut score development (LR, DA, and ROC) function at the local school level. The following questions summarize the study's aims: (1) How accurate are second grade R-CBM screeners in predicting third grade state test performance? (2) Which statistical methods produce R-CBM cut scores with adequate long-term classification accuracy? and (3) How does the long-term classification accuracy of local and publisher-recommended cut scores compare?

TABLE 1 Percentages of students in various demographic categories by sample

Demographic		Total Sample ($N = 266$)	Sample 1 ($n = 170$)	Sample 2 ($n = 46$)	Sample 3 ($n = 50$)
Gender	Male	50.0	51.6	44.4	50.0
	Female	50.0	48.4	55.5	50.0
Race/ethnicity	Caucasian	45.8	47.1	40.0	46.8
	African American	21.8	21.7	27.5	17.0
	Hispanic	17.3	17.4	12.5	21.3
	Asian American	3.1	2.2	5.0	4.3
	Native American	.9	.7	2.5	.0
	Multiracial	11.1	10.9	12.5	10.6
Exceptionality	Specific learning disability	7.5	11.6	2.4	.0
	Language impairment	2.2	1.4	2.4	4.3
	Emotional/behavioral disability	.4	.7	.0	.0
	Other health impaired	.9	.0	4.9	.0
	504 Plan	6.2	5.1	9.8	6.4
	Gifted	17.7	25.4	7.3	4.3
	Multiple exceptionalities	3.1	4.3	.0	2.1
Lunch status	Paid status	73.3	74.2	68.1	75.0
	Free or reduced price	26.7	25.8	31.2	25.0

2 | METHODS

2.1 | Research setting and participants

Participants were 266 students enrolled in second grade between 2004 and 2010 at a university-affiliated, developmental research school in North Central Florida. All students enrolled in second grade during this time period were eligible for inclusion in the study. Specifically, the sample included 21 students enrolled during 2004–2005; 50 students during 2005–2006; 51 students during 2006–2007; 48 students during 2007–2008; 46 students during 2008–2009; and 50 students during 2009–2010. All participants also completed third grade at this school the following year. Demographic data for all samples are displayed in Table 1.

2.2 | Measures

2.2.1 | DORF, 6th edition

The DORF (Good & Kaminski, 2002a) is a standardized, individually administered test of reading rate and accuracy. Although the test's authors released a more recent version of this instrument (DIBELS Next) in 2010, many schools continue to use the 6th Edition for screening and progress monitoring (Smolkowski & Cummings, 2016). Students were presented with three brief grade-level passages and asked to read aloud as many words as possible in 1 minute. Of the three passages, the median number of WCPM was recorded for each assessment period (AP). The DORF has high alternate form reliability ($r_{AB} = .89–.94$), test–rest reliability ($r_{XX} = .92–.97$), and high moderate to high criterion-related evidence of validity with various nationally normed and state-standardized tests of reading comprehension ($r = .65–.80$) (Good & Kaminski, 2002b; Good, Kaminski, Smith, & Bratten, 2001).

2.2.2 | Florida Comprehensive Assessment Test (FCAT)

The FCAT is a criterion-referenced achievement test previously administered to students in grades 3–12 in Florida. For third graders, the reading section is predominantly a measure of comprehension and typically consists of 50–55 multiple choice questions posed in response to brief reading passages. FCAT Reading scaled scores are classified as falling within one of five achievement levels (Levels 1–5). Students who achieve a scaled score of 284 or higher (Level 3 or higher) are considered to be achieving at grade level. The internal consistency reliability coefficient for the third grade FCAT Reading section is .92 (Harcourt Assessment, 2007). Criterion-related evidence of validity with other measures of comprehension has been substantiated (Harcourt Assessment, 2007). For example, correlations between the FCAT and the Stanford Achievement Test, 10th Edition (SAT-10; Harcourt Brace, 2003) range from .70 to .81 (Crist, 2001).

2.3 | Procedure

Demographic and assessment data were collected from archival databases provided by the school. All participants completed the DORF in September (fall AP), January (winter AP), and May (spring AP) of second grade as well as the FCAT in May of third grade. To ensure the integrity of administration and scoring, examiners were trained thoroughly on test procedures and conducted multiple sessions with experienced examiners to establish inter-rater reliability.

2.3.1 | Data analysis

Participants were divided into three subsamples: one calibration subsample (S1) and two cross-validation subsamples (S2 and S3). S1 comprised 170 participants who completed second grade between 2004 and 2008. S2 comprised 46 students who completed second grade during the 2008–2009 school year, and S3 consisted of 50 students who completed second grade during the 2009–2010 school year. Descriptive statistics and Pearson correlations for scores on all measures were computed, and DA, LR, and ROC were conducted.

2.3.2 | DORF cut scores

DORF cut scores for identifying students who were at risk for poor performance on the FCAT were developed based on data from S1. Poor performance on the FCAT Reading section was defined as scoring below Level 3 (i.e., scaled score <284). Overall, 11 cut scores for each of the three R-CBM administrations (i.e., two using DA, one using LR, six using ROC, and two using DORF publisher recommendations) were examined.

2.3.3 | Discriminant analysis

DA involves using a predictor variable to determine the probability of membership in discrete groups of the dependent variable (Silberglitt & Hintze, 2005). In this study, the predictor variable was DORF scores. The groups were defined as those who met and those who did not meet the grade-level proficiency standard on the FCAT, respectively. To develop cut scores for predicting risk status, the following formula was applied:

$$C = \frac{\bar{X}_{PASS} + \bar{X}_{FAIL}}{2} + \ln\left(\frac{P_{FAIL}}{P_{PASS}}\right),$$

where C is the resulting DORF cut score; $\bar{X}_{PASS}$ is the mean DORF score for students who ultimately met the FCAT performance criterion; $\bar{X}_{FAIL}$ is the mean DORF score for students who ultimately did not meet the FCAT performance criterion; P_{PASS} is the prior probability of meeting the FCAT performance criterion; and P_{FAIL} is the prior probability of failing to meet the FCAT performance criterion.

In this study, DA was used to develop DORF cut scores in two ways. In the first approach, we assumed that the prior probabilities of meeting and failing to meet the performance criterion on the outcome test were equal. Cut scores developed via this approach are referred to as the DA(.50) cut scores. In the second approach, we estimated adjusted prior probabilities of meeting and failing to meet the performance criterion based on data from S1. Cut scores developed using this approach are referred to as the DA(S1) cut scores.

2.3.4 | Logistic regression

LR is a regression analysis procedure in which a continuous or categorical independent variable is used to determine probabilities of membership in each of the categories of the dependent variable (Silberglitt & Hintze, 2005). In this respect, it is similar to DA; however, LR does not require strict adherence to normality and equal variance assumptions. For these analyses, the continuous independent variable was the participants' DORF scores. The dependent variable was FCAT performance, wherein participants fell in one of two categories: students who met the grade-level proficiency standard and students who failed to meet the standard. DORF scores that corresponded to a .50 probability of failing to meet the performance standard on the FCAT were identified as the cut scores for determining risk status.

2.3.5 | Receiver operating characteristic curve analysis

ROC involves plotting the true-positive rate (sensitivity) on the *y*-axis and the false-positive rate (1-specificity) on the *x*-axis for each possible screening cut score. ROC was used to derive cut scores in two ways. First, cut scores with the optimal balance between sensitivity and specificity [ROC(Optimal)] were identified for the fall, winter, and spring DORF administrations. ROC(Optimal) cut scores were defined as the cut scores with the smallest difference between sensitivity and specificity such that sensitivity was either greater than or equal to specificity. Second, cut scores with sensitivity values as similar as possible to the following values were identified for each administration: .70 [ROC(.70)], .75 [ROC(.75)], .80 [ROC(.80)], .85 [ROC(.85)], and .90 [ROC(.90)]. When cut scores with these precise values could not be identified, the cut scores with the lowest sensitivity levels that still met the minimum criteria ($\geq$.70, $\geq$.75, etc.) were selected.

2.3.6 | Publisher-recommended cut scores

Two sets of publisher-recommended cut scores from the DORF test manual were used: one set of scores for distinguishing "at risk" from "some risk" students and one set for distinguishing "some risk" from "low risk" students. Each of these sets of cut scores is available for the fall, winter, and spring APs, respectively.

2.4 | Evaluating classification accuracy

For all samples, AUC values were computed to determine the accuracy of the fall, winter, and spring DORF screeners in predicting FCAT performance. According to Smolkowski and Cummings (2015, 2016), the accuracy of a screener may be characterized as *excellent* ($A \geq .95$), *very good* ($A = .85–.94$), *reasonable* ($A = .75–.84$), or *limited* ($A < .75$). As they stated, most reading screeners cannot be characterized as *excellent* and often exhibit values of *A* below .95. Screeners administered within the same academic year as the gold standard measure are likely to have higher levels of decision-making accuracy than screeners administered during the previous academic year. Thus, Swets (1988) suggested that, depending on the context in which screening occurs, "values of *A* between .70 and .90 represent accuracies that are useful for some purposes, and higher values represent a rather high accuracy" (p. 1292).

Using data from S1, the classification accuracy of cut scores for the fall, winter, and spring DORF administrations was evaluated by computing the following CAIs: sensitivity, specificity PPP, NPP, and OCC. Cut scores developed for the calibration sample (S1) were then cross-validated on two independent samples (S2 and S3). Again, the aforementioned CAIs were computed for all cut scores for these two samples. Because false negatives may have more serious consequences for individuals and systems than false positives, strongest consideration was given to the cut scores' observed levels of sensitivity over multiple years (Smolkowski & Cummings, 2016). Only DORF cut scores that maintained sensitivity levels of at least .70 across subsamples were considered to have retained their classification accuracy over multiple years. In a recent diagnostic test accuracy meta-analysis of R-CBM properties, Kilgus, Methe, Maggin, and Tomasula (2014) found that screeners administered more than 12 months in advance of the criterion measure had relatively lower sensitivity values (i.e., .70–.80) than screeners administered fewer than 12 months earlier (sensitivity $\geq$ .80). Given that all DORF administrations occurred at least 13 months prior to FCAT administration, a minimum

TABLE 2 Publisher-recommended and locally developed DORF cut scores

	FCAT		
Cut Score	**Fall**	**Winter**	**Spring**
Publisher recommended			
DORF "at risk"	26	52	70
DORF "some risk"	44	68	90
Discriminant analysis (DA)			
DA(.50)	67	81	100
DA(S1)	66	79	99
Logistic regression	N/A	28	47
Receiver operating characteristic (ROC)			
Curve analysis			
ROC(.70)	62	75	98
ROC(.75)	69	76	102
ROC(.80)	74	78	103
ROC(.85)	80	99	104
ROC(.90)	95	105	115
ROC(optimal)	62	75	94

Note. Values are expressed in terms of words read correctly per minute (WCPM); N/A indicates cut score that fell outside the DORF score range for all samples.

sensitivity value of .70 was deemed appropriate for this study. For cut scores with consistently acceptable sensitivity values, additional CAIs (specificity, PPP, NPP, and OCC) also were considered.

3 | RESULTS

3.1 | Cut score values and descriptive statics

Table 2 displays both the publisher-recommended and local cut scores for the fall, winter, and spring DORF administrations (developed using S1 data). As shown in the table, the publisher-recommended cut scores were among the lowest examined (with the exception of the LR cut scores). The LR procedure did not result in a viable fall DORF cut score (i.e., a cut score within the range of observed scores for S1). Cut scores developed via the two DA methods were similar, with a difference of no more than two points between cut scores developed for the same AP.

Table 3 displays descriptive statistics for the DORF and FCAT for all participant samples as well as for statewide and national samples of second and third graders. Means for the three DORF administrations were fairly similar between S1 and S2 but somewhat higher for S3. Mean FCAT scores were generally comparable between S1 and S3 but somewhat lower for S2. Generally, mean scores for the DORF and FCAT were higher than those seen in larger, more diverse samples; however, score ranges were somewhat smaller than in state and national samples.

Table 4 displays FCAT pass and fail rates across groups. Although pass rates were comparable, they were somewhat lower for S2. Generally, pass rates were higher than in state and national samples. For example, approximately 69–75% of third grade students in Florida earned passing scores on the FCAT between 2006 and 2011 (Florida Department of Education [FLDOE], 2011), whereas more than 75% of students in this study earned passing scores.

TABLE 3 Descriptive statistics for DORF and FCAT performance across samples

Sample	Measure	Mean	*SD*	Range
Total sample ($N = 266$)	Fall DORF	76.6	37.7	12–183
	Winter DORF	90.4	37.3	15–201
	Spring DORF	109.0	37.1	33–218
	FCAT	332.0	55.0	532–762
S1 ($n = 170$)	Fall DORF	73.5	36.4	21–181
	Winter DORF	88.8	35.3	15–189
	Spring DORF	108.7	36.6	42–218
	FCAT	333.4	54.6	147–500
S2 ($n = 46$)	Fall DORF	72.7	38.2	12–166
	Winter DORF	86.7	39.2	26–184
	Spring DORF	101.6	35.9	33–187
	FCAT	318.2	54.5	201–437
S3 ($n = 50$)	Fall DORF	90.6	39.3	27–183
	Winter DORF	99.4	41.4	32–201
	Spring DORF	116.7	38.9	48–218
	FCAT	340.0	55.9	222–500
DORF, 6th Ed. (norm sample)[a]	Fall DORF ($N = 637,017$)	56.37	33.4	0–256
	Winter DORF ($N = 615,480$)	84.94	37.8	0–275
	Spring DORF ($N = 608, 782$)	98.13	37.8	0–247
FCAT-statewide sample[b]	FCAT-statewide sample ($N = 4,645$)	301.4	59.2	N/A[c]

Note. [a]Data as reported by Cummings et al. (2011).
[b]Data as reported by the Human Resources Research Organization (2002).
[c]Data were not available to the authors.

TABLE 4 Pass and fail rates for the FCAT across subsamples

Sample	n_{Pass}	Percentage (Pass)	n_{Fail}	Percentage (Fail)
Whole sample	215	80.8%	51	19.2%
Sample 1	138	81.2%	32	18.8%
Sample 2	35	76.1%	11	23.9%
Sample 3	42	84.0%	8	16.0%

Note. n_{Pass} indicates number of participants who passed the standardized test; n_{Fail} indicates number of participants who failed the standardized test.

3.2 | Correlations

Table 5 presents correlations between DORF and FCAT scores. As can be seen here, correlations between all measures were statistically significant ($p < .01$).[1] High correlation coefficients (ranging from .88 to .96) were observed between all DORF administrations, with the highest correlations observed between consecutive administrations (e.g., fall and winter administrations). Correlations between the DORF and FCAT scores ranged from .49 to .74 and were somewhat lower, but generally comparable to, correlations between R-CBMs and state achievement tests that have been observed in previous research (Yeo, 2010). As expected, correlations between the DORF measures and FCAT generally increased as the length of time between their administrations decreased.

[1] All correlations remained statistically significant when the Holm–Bonferroni multiple comparison procedure was applied.

TABLE 5 Pearson correlations between DORF and FCAT measures

Sample	Administration	Fall DORF[b]	Winter DORF[b]	Spring DORF[b]
Total sample	Winter DORF	.91[a]		
$N = 266$	Spring DORF	.88[a]	.94[a]	
	FCAT	.55[a]	.59[a]	.62[a]
Sample 1	Winter DORF	.89[a]		
$n = 170$	Spring DORF	.88[a]	.92[a]	
	FCAT	.49[a]	.51[a]	.55[a]
Sample 2	Winter DORF	.93[a]		
$n = 46$	Spring DORF	.89[a]	.96[a]	
	FCAT	.72[a]	.77[a]	.74[a]
Sample 3	Winter DORF	.95[a]		
$n = 50$	Spring DORF	.90[a]	.96[a]	
	FCAT	.61[a]	.67[a]	.68[a]

Note. [a]$p < .01$.
[b]All results remained significant when the Holm–Bonferroni multiple comparison procedure was applied.

TABLE 6 Area under the curve values for whole sample and subsamples

Sample	Receiver Operating Characteristic (ROC) Curve	Area Under the Curve (AUC)[c]	Standard Error (*SE*)	95% Confidence Interval	Adjusted Confidence Interval[d]
Whole sample	Fall DORF-FCAT	.737[b]	.034	.671–.802	.640–.834
	Winter DORF-FCAT	.807[b]	.031	.747–.867	.718–.896
	Spring DORF-FCAT	.781[b]	.030	.721–.840	.695–.867
Sample 1	Fall DORF-FCAT	.649[a]	.046	.558–.740	.517–.781
	Winter DORF-FCAT	.731[b]	.045	.644–.819	.602–.860
	Spring DORF-FCAT	.712[b]	.043	.628–.796	.589–.835
Sample 2	Fall DORF-FCAT	.873[b]	.051	.773–.973	.726–1.000
	Winter DORF-FCAT	.913[b]	.041	.832–.994	.796–1.000
	Spring DORF-FCAT	.856[b]	.057	.745–.967	.693–1.000
Sample 3	Fall DORF-FCAT	.857[a]	.066	.728–.986	.668–1.000
	Winter DORF-FCAT	.936[b]	.035	.868–1.000	.836–1.000
	Spring DORF-FCAT	.903[b]	.044	.818–.989	.777–1.000

Note. [a]$p < .01$;
[b]$p < .001$.
[c]All results remained significant when the Holm–Bonferroni multiple comparison procedure was applied.
[d]Confidence intervals were adjusted by using the Bonferroni critical value to account for multiplicity.

3.3 | Research questions: Classification agreement analyses

3.3.1 | Question 1: Assessing overall screener accuracy

Table 6 displays the AUC values (including standard error values and confidence intervals) for the DORF across samples. For the whole sample, values of *A* ranged from .737 to .807, indicating *reasonable* to *very good* classification accuracy. Across the three subsamples, values of *A* ranged from .649 to .936, indicating *limited* to *very good* classification accuracy. All values of *A* were statistically significant ($p < .01$), meaning that they predicted risk status significantly

TABLE 7 Fall publisher-recommended and local DORF cut scores for predicting FCAT performance: results of classification agreement analyses

Cut Score	Sensitivity	Specificity	PPP	NPP	OCC
DORF "at risk"	.03/.27/NA	.99/1.00/NA	.33/1.00/NA	.81/.81/NA	.81/.83/NA
DORF "some risk"	.25/.64/.50	.80/.80/.91	.23/.50/.50	.82/.88/.91	.70/.76/.86
DA(.50)	.72/1.00/.88	.51/.69/.81	.25/.50/.47	.89/1.00/.97	.55/.76/.84
DA(S1)	.72/1.00/.75	.51/.71/.81	.25/.52/.43	.89/1.00/.96	.55/.78/.82
LR	N/A	N/A	N/A	N/A	N/A
ROC(.70)	.72/.91/.75	.57/.74/.86	.28/.53/.50	.90/.96/.95	.60/.78/.86
ROC(.75)	.78/1.00/.88	.48/.57/.79	.28/.42/.44	.90/1.00/.97	.54/.67/.82
ROC(.80)	.81/1.00/.88	.45/.51/.72	.25/.39/.37	.91/1.00/.97	.52/.63/.76
ROC(.85)	.88/1.00/.88	.41/.40/.58	.25/.34/.28	.93/1.00/.96	.49/.54/.64
ROC(.90)	.91/1.00/1.00	.32/.34/.47	.24/.32/.26	.94/1.00/1.00	.43/.50/.56
ROC(optimal)	.72/.91/.75	.57/.74/.86	.28/.53/.50	.90/.96/.95	.60/.78/.86

Note. In each cell, values of classification agreement indicators are displayed for all three subsamples and are recorded in the following format: sample 1/sample 2/sample 3.

TABLE 8 Winter publisher-recommended and local DORF cut scores for predicting FCAT performance: results of classification agreement analyses

Cut Score	Sensitivity	Specificity	PPP	NPP	OCC
DORF "at risk"	.16/.63/.50	.91/.89/.98	.28/.64/.80	.82/.89/.91	.76/.83/.90
DORF "some risk"	.53/.91/.88	.76/.77/.88	.34/.56/.58	.88/.96/.97	.72/.80/.88
DA(.50)	.84/1.00/1.00	.60/.63/.71	.33/.46/.40	.94/1.00/1.00	.65/.72/.76
DA(S1)	.84/1.00/1.00	.62/.63/.76	.34/.46/.44	.94/1.00/1.00	.66/.72/.80
LR	.00/.09/NA	.99/1.00/NA	.00/1.00/NA	.81/.78/NA	.81/.78/NA
ROC(.70)	.72/1.00/.88	.67/.69/.81	.34/.50/.47	.91/1.00/.97	.68/.76/.82
ROC(.75)	.75/1.00/.88	.66/.69/.81	.34/.50/.47	.92/1.00/.97	.68/.76/.82
ROC(.80)	.81/1.00/1.00	.62/.66/.79	.33/.48/.47	.93/1.00/1.00	.66/.74/.82
ROC(.85)	.88/1.00/1.00	.37/.49/.48	.24/.38/.27	.93/1.00/1.00	.46/.61/.56
ROC(.90)	.91/1.00/1.00	.33/.43/.43	.24/.35/.25	.94/1.00/1.00	.44/.57/.52
ROC(optimal)	.72/1.00/.88	.67/.69/.81	.34/.50/.47	.91/1.00/.97	.68/.76/.82

Note. In each cell, values of classification agreement indicators are displayed for all three subsamples and are recorded in the following format: sample 1/sample 2/sample 3.

better than chance.[2] Generally, higher values of *A* were observed for DORF administrations that occurred closer in time to the FCAT administration (i.e., winter and spring administrations). With the exception of the Fall DORF screener in S1, all screening measures yielded values of *A* greater than .70. Standard error values ranged from .031 to .066 across screening administrations.

3.3.2 | Question 2: Comparing methods of local cut score development

Tables 7–9 display sensitivity, specificity, PPP, NPP, and OCC values for each set of 11 DORF cut scores used to predict FCAT performance. Each table also displays cut scores and CAIs for one of the three APs. For each cut score, CAI values in S1, S2, and S3, respectively, are displayed. Of the three methods of cut score development, only two (DA and ROC) produced cut scores that consistently met the minimum sensitivity criterion of .70 across samples.

[2] All AUC values remained statistically significant when the Holm–Bonferroni procedure was applied.

TABLE 9 Spring publisher-recommended and local DORF cut scores for predicting FCAT performance: results of classification agreement analyses

Cut Score	Sensitivity	Specificity	PPP	NPP	OCC
DORF "at risk"	.19/.45/.25	.91/.94/.93	.32/.71/.40	.83/.85/.87	.77/.83/.82
DORF "some risk"	.53/.91/1.00	.73/.66/.81	.31/.45/.50	.87/.96/1.00	.69/.72/.84
DA(.50)	.72/1.00/1.00	.56/.54/.74	.27/.50/.42	.90/.83/1.00	.59/.67/.78
DA(S1)	.72/1.00/1.00	.58/.57/.74	.28/.50/.42	.90/.81/1.00	.61/.67/.78
LR	.03/.09/NA	1.00/1.00/NA	1.00/1.00/NA	.82/.78/NA	.82/.78/NA
ROC(.70)	.72/1.00/1.00	.58/.57/.76	.28/.53/.44	.90/.81/1.00	.61/.70/.80
ROC(.75)	.75/1.00/1.00	.53/.54/.71	.27/.58/.40	.90/.90/1.00	.57/.67/.76
ROC(.80)	.81/1.00/1.00	.52/.51/.71	.28/.58/.40	.92/.90/1.00	.58/.67/.76
ROC(.85)	.88/1.00/1.00	.51/.51/.71	.29/.45/.40	.95/1.00/1.00	.58/.63/.76
ROC(.90)	.94/1.00/1.00	.44/.43/.60	.28/.45/.32	.97/1.00/1.00	.53/.63/.66
ROC(optimal)	.66/.91/1.00	.65/.60/.76	.30/.42/.44	.89/.95/1.00	.65/.67/.80

Note. In each cell, values of classification agreement indicators are displayed for all three subsamples and are recorded in the following format: sample 1/sample 2/sample 3.

Often, the LR method did not yield viable cut scores for one or more of the subsamples. When viable, the corresponding sensitivity levels of LR cut scores were exceptionally low across subsamples (≤.09). Conversely, specificity values for these cut scores were very high (≥.99) across samples. OCC values also were somewhat more promising for the LR cut scores. When viable, LR cut scores correctly classified approximately 80% of students in predicting FCAT outcomes.

Generally, CAI values were fairly similar for cut scores generated via the two DA methods. As shown in Tables 7–9, the DA(.50) cut scores consistently exhibited sensitivity values greater than or equal to .70 across subsamples. Similarly, for the DA(S1) cut scores, adequate levels of sensitivity were observed across APs in predicting FCAT performance. Given the relatively high sensitivity levels observed for the DA cut scores, it is not surprising that their specificity levels were somewhat lower. For cut scores developed via both DA methods, values of specificity ranged from .51 to .81 across APs and subsamples, with most of these values falling below .70. For the DA(.50) and DA(S1) cut scores, values of OCC ranged from .55 to .84, indicating that these sets of cut scores correctly classified between 55% and 84% of students.

For both the DA(.50) and DA(S1) cut scores, estimates of PPP ranged from .25 to .52 across APs and subsamples in predicting FCAT outcomes. More specifically, the probability that a given individual who was identified as "at risk" on the DORF ultimately did not perform at grade level on the FCAT was often below chance (i.e., <.5). Estimates of NPP were much higher than estimates of PPP and ranged from .81 to 1.00 for both sets of DA cut scores.

Most ROC cut scores met the minimum sensitivity criterion when cross-validated on S2 and S3. Generally, the ROC cut scores maintained their targeted sensitivity levels over time, with few exceptions. For example, the ROC(.80) cut scores consistently maintained sensitivity levels greater than .80, across subsamples and APs. As expected, however, cut scores with higher targeted sensitivity levels [e.g., the ROC(.85) and ROC(.90) cut scores] had lower levels of specificity. For example, specificity values for the ROC(.70) cut scores ranged from .57 to .86, whereas they ranged from .32 to .60 for the ROC(.90) cut scores.

Values of OCC varied for the ROC cut scores. Generally, the highest OCC values were observed for cut scores with lower targeted levels of sensitivity. For example, OCC values for the ROC(.70) cut scores ranged from .60 to .86, whereas they ranged from .43 to .66 for the ROC(.90) cut scores. For the ROC cut scores, PPP values were relatively low across APs, ranging from .24 to .34 for S1, from .32 to .58 for S2, and from .25 to .50 for S3. Conversely, NPP values were much higher and ranged from .89 to .97 for S1, from .81 to 1.00 for S2, and from .95 to 1.00 for S3, respectively.

3.3.3 | Question 3: Comparing locally developed and publisher-recommended cut scores

As indicated in Tables 7–9, sensitivity levels for the "at risk" publisher-recommended cut scores were generally low, ranging from .03 to .63 across APs. As a result, none of the fall, winter, and spring cut scores met the minimum sensitivity criterion in predicting FCAT performance. These cut scores did, however, have consistently high levels of specificity, ranging from .89 to 1.00 across subsamples. As compared with the "at risk" cut scores, the DORF "some risk" cut scores had somewhat higher levels of sensitivity, ranging from .25 to 1.00 across subsamples. Nevertheless, sensitivity levels for these cut scores were generally less than .70. Similar to the DORF "at risk" cut scores, specificity levels for the "some risk" cut scores were typically above .70, ranging from .66 to .91.

Sensitivity levels generally were higher for the local cut scores (specifically the DA and ROC cut scores) across samples. However, values of specificity were generally lower for the local cut scores than for the publisher-recommended cut scores. Moreover, the publisher-recommended cut scores had among the highest values of OCC. Values of OCC for the DORF "at risk" cut scores ranged from .76 to .90 across subsamples, with most of these values greater than or equal to .80. Similarly, for the "some risk" cut scores, values of OCC ranged from .69 to .88. In comparison, OCC values for the ROC and DA cut scores generally fell below .80. Thus, the publisher-recommended cut scores (especially the "at risk" cut scores) correctly classified larger percentages of students across samples than the DA and ROC cut scores.

4 | DISCUSSION

This study examined the utility of second grade R-CBM screeners in predicting performance on a third grade statewide reading test. Specifically, we compared the long-term classification accuracy of local DORF cut scores developed via three statistical methods (LR, DA, and ROC) with that of 6th Edition DORF publisher-recommended cut scores, in predicting FCAT performance. Results of AUC analyses indicated that, with the exception of the S1 fall screening, values of *A* for all DORF administrations were above .70, and the majority of these values were above .75. Given that the lengths of time between screener and state test administrations in this study (i.e., 12–20 months) were longer than typically seen in previous studies, it is noteworthy that the DORF measures generally had adequate utility in predicting FCAT performance. These results suggest that early screenings can provide important information about a student's preliminary risk for poor state test performance well in advance of third grade. Of course, this information must be supplemented with ongoing progress monitoring data throughout second and third grades, as students' skill sets evolve in response to instruction.

Regarding the DORF cut scores in this study, LR and publisher-recommended cut score values were generally lower than DA and ROC cut score values. Only the DA and ROC cut scores consistently maintained the minimum acceptable level of sensitivity (i.e., ≥.70) when cross-validated on independent samples. Across multiple years, ROC cut scores consistently maintained the minimum levels of sensitivity specified for their development. In contrast, when viable, the LR and publisher-recommended cut scores had exceptionally low sensitivity levels but high OCC levels. Ultimately, while the DA and ROC cut scores correctly identified more of the truly at-risk students, the LR and publisher-recommended cut scores correctly classified greater numbers of students overall. This may be due to the fact that these cut scores were lower and thus more conservative in identifying students as "at risk" than the DA and ROC cut scores.

Across all sets of cut scores, low levels of PPP and high levels of NPP were observed. These findings likely are attributable to the low base rates of the target condition (i.e., poor FCAT performance) observed in this study. When base rates of the target condition are relatively low, estimates of PPP tend to be lower, while estimates of NPP tend to be higher. Thus, caution should be exhibited in comparing PPP and NPP values across samples with different base rates.

Overall, results of this study corroborate findings from prior research. For example, we found that cut scores developed via ROC and DA had higher values than LR and publisher-recommended cut scores (cf. Goffreda et al., 2009; Silberglitt & Hintze, 2005). Moreover, similar to previous studies, we found that local cut scores had higher

sensitivity levels in identifying students at risk for poor state test performance. In particular, results of this study indicated that the ROC cut scores had considerable decision-making utility in identifying at-risk students, which substantiates findings from Silberglitt and Hintze (2005) and Sandberg Patton et al. (2014).

However, the results of this study corroborated only some of the findings from Silberglitt and Hintze (2005), who found that cut scores developed via all three of the aforementioned methods (i.e., LR, DA, and ROC) exhibited adequate sensitivity levels (i.e., ≥.70). Based on their findings, Silberglitt and Hintze recommended that practitioners use LR to develop initial cut scores and establish appropriate sensitivity and specificity criteria. They also recommended that practitioners adjust these cut scores via ROC to maximize OCC. Silberglitt and Hintze cautioned against the use of DA to develop cut scores, due to its rigid underlying distributional assumptions.

In contrast, we found that only the DA and ROC methods consistently produced cut scores with adequate sensitivity levels, whereas the LR cut scores did not (nor were they viable decision points for all subsamples). Although comparisons between the two studies must be made cautiously (due to differences in participant samples and measures), there are a number of possible explanations for these discrepant results. One possible explanation concerns differences in the implementation of the LR method. In this study, a value of .50 was inputted as the probability of failing the outcome measure in the LR formula. In contrast, Silberglitt and Hintze did not specify this value. Other explanations for these discrepancies may pertain to differences in sample sizes, participant characteristics, and R-CBM characteristics, all of which can impact CAI values. Notably, Silberglitt and Hintze did not cross-validate their cut scores in independent samples. Thus, it remains unclear whether these cut scores would have maintained adequate levels of classification accuracy across multiple cohorts. Nevertheless, both studies suggest the value of using local cut scores to improve the identification of at-risk students.

4.1 | Limitations

Several limitations to this study should be acknowledged. First, some students in each subsample received secondary and tertiary reading interventions during second and third grades, and decisions to initiate and terminate these interventions were based partially on their DORF performance. This interference, along with systems-level changes in instructional practices over time (e.g., implementation of response to intervention), could have attenuated the relationship between scores on the DORF and FCAT, thereby affecting the predictive power of the screeners (Silberglitt & Hintze, 2005). It is noteworthy that, despite these potential interferences, the DORF screeners in this study maintained acceptable levels of classification accuracy over time. Second, the results of this study may have limited generalizability to settings with different student populations (e.g., linguistic backgrounds and educational achievement); nevertheless, it generally is expected that local decision rules will have the greatest utility in the settings for which they were initially designed. Moreover, this sample did not include English language learners (as none were enrolled during these years). It was, however, diverse with respect to students' racial, ethnic, socioeconomic, and disability backgrounds.

Another possible limitation of this study concerns the size of the sample (S1) used to develop the local cut scores. Estimates of sensitivity and specificity may be biased when cut scores are developed with smaller samples, due to the potential impact of random sampling error (Leeflang, Moons, Reitsma, & Zwinderman, 2008). VanDerHeyden (2011) recommended using a calibration sample size of at least 200 participants when developing local cut scores. This recommendation is based on the work of Leeflang et al., who conducted a series of data simulations to estimate sensitivity and specificity values for a given screening cut score using a variety of sample sizes. They found that, when sample size increased from 40 to 200 participants, the accuracy of sensitivity and specificity estimates improved significantly. Nevertheless, Leeflang et al. did not calculate sensitivity and specificity values for all possible sample sizes between 40 and 200; thus, it remains unclear whether a somewhat smaller sample (i.e., $n = 170$) would yield sufficiently precise estimates of these metrics.

Finally, an important characteristic of this study is that students in S3 were administered a different version of the FCAT than students in the S1 and S2 subsamples. In spring 2011, the FLDOE transitioned from the FCAT to the FCAT 2.0, and the S3 participants were the first cohort to complete this new edition of the test. Despite this transition, local DORF cut scores maintained adequate classification accuracy in predicting students' risk status for both versions of

the test. Although potentially disruptive to a school's attempts to coordinate screening procedures, modifications to statewide testing practices inevitably occur over time. When these changes occur, it is important for practitioners to consider how they might impact the integrity of screening procedures and decision rules. In the present study, it is noteworthy that local screening cut scores maintained their classification accuracy despite the fact that gradual modifications were made to state testing practices. Given that both Florida's state test (now the Florida Standards Assessments) and DORF have been updated because this study was conducted, further research is necessary to explore relationships between these two tests. As in any study of screening accuracy, generalizability of the present findings may be limited due to differences in test characteristics, length of time between screener and state test administration, and population characteristics across settings. Ultimately, it is incumbent on school personnel to evaluate the accuracy and utility of screeners in their local contexts. Nevertheless, the present results suggest that second grade R-CBM screeners can be used to predict third grade state test performance and that local cut scores may maintain adequate classification accuracy across multiple cohorts.

4.2 | Implications for schools and directions for future research

Overall, the results of this study underscore the value of using local R-CBM cut scores to predict students' subsequent performance on high-stakes achievement tests. Specifically, both the DA and ROC methods for developing local benchmarks may be especially promising. One drawback to the DA procedure, however, is that it requires greater attention to the distributional properties of datasets. ROC, on the other hand, may require greater expertise in signal detection methodology, although it affords the greatest flexibility in cut score selection and is robust to a number of common statistical assumption violations (Silberglitt & Hintze, 2005). Given an appropriate sample size, ROC may be ideal because it allows practitioners to choose cut scores with specified values of sensitivity and other metrics. (It should be noted, however, that cross-validation of these cut scores is still necessary to ensure that CAIs are estimated accurately.)

Future research should explore the utility of the DA, ROC, and LR methods of cut score development in schools with varied student populations, instructional practices, and outcome measures. In particular, this research should explore the utility of local cut scores in predicting state test performance for diverse groups of students. Scholars also may benefit from exploring additional predictors of state test performance as well as the utility of ROC and DA methods for developing cut scores for these measures. A related question centers on how scholars can make these statistical methods more accessible to practitioners. The methods described above are especially promising, as all of them can be conducted using basic analytic software readily available in most schools (e.g., Microsoft Excel). Further research in these areas may increase schools' capacity to meet the needs of all learners in a more efficient manner.

REFERENCES

Adolf, S., Catts, H., & Lee, J. (2010). Kindergarten predictors of 2nd versus 8th grade reading comprehension impairments. *Journal of Learning Disabilities*, *43*, 322–345.

Crist, C. (2001). *FCAT briefing book*. Tallahassee: Florida Department of Education.

Cummings, K., Otterstedt, J., Kennedy, P., Baker, S., & Kame'enui, E. (2011). *DIBELS Data System: 2009–2010 percentile ranks for DIBELS 6th Edition benchmark assessments* (Technical Report 1102). Eugene, OR: University of Oregon.

Cummings, K. D., & Smolkowski, K. (2015). Selecting students at risk of academic difficulties. *Assessment for Effective Intervention*, *41*, 55–61.

Fletcher, J., Lyon, G., Fuchs, L., & Barnes, M. (2007). *Learning disabilities: From identification to intervention*. New York, NY: Guilford.

Florida Department of Education (FLDOE). (2011). State level report: Florida Comprehensive Assessment Test. Retrieved from http://fcat.fldoe.org/results/default.asp

Fuchs, D., Compton, D., Fuchs, L., Bryant, J., Hamlett, C., & Lambert, W., (2012). First-grade cognitive abilities as long-term predictors of reading comprehension and disability status. *Journal of Learning Disabilities*, *45*, 217–231.

Goffreda, C., & DiPerna, J. (2010). An empirical review of psychometric evidence for the DIBELS. *School Psychology Review*, *39*, 463–483.

Goffreda, C. T., DiPerna, J. C., & Pedersen, J. A. (2009). Preventive screening for early readers: Predictive validity of the DIBELS. *Psychology in the Schools*, *46*, 539–552.

Good, R. H., & Kaminski, R. A. (2002a). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: University of Oregon.

Good, R., & Kaminski, R. A. (2002b). *DIBELS oral reading fluency passages for first through third grades* (Report No. 10). Eugene, OR: University of Oregon.

Good, R., Kaminski, R., Smith, S., & Bratten, J. (2001). *Technical adequacy of 2nd grade DIBELS Oral Reading Fluency passages* (Report 8). Eugene, OR: University of Oregon.

Harcourt Assessment. (2007). *Reading and mathematics: Technical report for 2006 FCAT test administrations*. San Antonio, TX: Author.

Harcourt Brace. (2003). *Stanford achievement test* (10th ed.). San Antonio, TX: Author.

Human Resources Research Organization. (2002). *Florida Comprehensive Assessment Test (FCAT) for reading and mathematics: Technical report for tests administrations of FCAT 2002*. San Antonio, TX: Author/Harcourt Educational Measurement.

Kilgus, S., Methe, S., Maggin, D., & Tomasula, J., (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting universal screening. *Journal of School Psychology*, *52*, 377–405.

Leeflang, M. M. G., Moons, K. G. M., Reitsma, J. B., & Zwinderman, A. H. (2008). Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: Mechanisms, magnitude, and solutions. *Clinical Chemistry*, *54*, 729–737.

Sandberg Patton, K. L., Reschly, A. L., & Appleton, J. (2014). Curriculum-based measurement as a predictor of performance on a state assessment: Diagnostic efficiency of local norms. *Educational Assessment*, *19*, 284–301.

Scarborough, H. (1998). Predicting the future achievement of second graders with reading disabilities: Contributions of phonemic awareness, verbal memory, rapid naming, and IQ. *Annals of Dyslexia*, *48*, 115–136.

Silberglitt, B., & Hintze, J. (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.

Smolkowski, K., & Cummings, K. D. (2015). Evaluation of diagnostic systems: The selection of students at risk of academic difficulties. *Assessment for Effective Intervention*, *41*, 41–54.

Smolkowski, K., & Cummings, K. (2016). Evaluation of the DIBELS (6th Ed.) diagnostic system for the selection of native and proficient English speakers at risk of reading difficulties. *Journal of Psychoeducational Assessment*, *34*, 103–118.

Steubing, K., Barth, A., Trahan, L., Reddy, R., Miciak, J., & Fletcher, J. (2015). Are cognitive characteristics strong predictors of responses to intervention? A meta-analysis. *Review of Educational Research*, *85*, 395–429.

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*, 1285–1293.

VanDerHeyden, A. (2010). Use of classification agreement analyses to evaluate RtI implementation. *Theory into Practice*, *49*, 281–288.

VanDerHeyden, A. (2011). Technical adequacy of response to intervention decisions. *Council for Exceptional Children*, *77*, 335–350.

Wood, D. E. (2006). Modeling the relationship between oral reading fluency and performance on a statewide reading test. *Educational Assessment*, *11*, 85–104.

Yeo, S. (2010). Predicting performance on statewide achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education*, *31*, 412–422.

How to cite this article: Grapin SL, Waldron N, Joyce-Beaulieu D, Algina J. Developing local oral reading fluency cut scores for predicting high-stakes test performance. *Psychol Schs*. 2017;54:932–946. https://doi.org/10.1002/pits.22035