Screening in the Upper Elementary Grades: Identifying Fourth Grade Students At-Risk for Failing the State Reading Assessment

Shawn C. Kent, Ph.D.

University of Houston

Jeanne Wanzek, Ph.D.

Vanderbilt University

Joonmo Yun, M.S.

Florida State University

Published 2019

Kent, S. C., Wanzek, J., & Yun, J. (2019). Screening in the upper elementary grades: Identifying fourth grade students at-risk for failing the state reading assessment. *Assessment for Effective Intervention, 44*, 160-172.

Shawn C. Kent, Department of Educational Leadership & Policy Studies, University of Houston; Jeanne Wanzek, Department of Special Education, Vanderbilt University, jeanne.wanzek@vanderbilt.edu; Joonmo Yun, Florida Center for Reading Research and School of Teacher Education, Florida State University, jyun@fcrr.org.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A150269 to Vanderbilt University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Correspondence should be addressed to Shawn Kent, Department of Educational Leadership & Policy Studies, University of Houston, 3657 Cullen Blvd. Room 112, Houston, TX 77204

Email: sckent@uh.edu

Abstract

This study examined the predictive validity and classification accuracy of individual and group-administered screening measures relative to student performance on a year-end state reading assessment in two states. A sample of 321 students were assessed in the areas of word-level and text fluency, as well as reading comprehension in the fall of fourth grade. For individual measures, a group-administered reading comprehension test exhibited the highest classification accuracy (74-80%) for both outcomes though no screener demonstrated optimal sensitivity and specificity levels. Using a multivariate approach, logistic regression results revealed minimal to no increase in classification accuracy over the single comprehension measure. ROC curve analyses determined local cut scores in order to maintain sensitivity constant at .90 which resulted in a large number of false positives. The implications and limitations of these findings for screening at the upper elementary level are discussed.

Screening in the Upper Elementary Grades: Identifying Fourth Grade Students At-Risk for Failing the State Reading Assessment

At their essence, multi-tiered systems of support, such as Response to Intervention (RtI), with a focus on identifying students in need of support and providing targeted, data-driven intervention, provide a systematic framework designed to change the trajectory of reading outcomes for struggling readers at all levels (Johnson, Mellard, Fuchs, & McKnight, 2006). The foundation of successful implementation of RtI for ameliorating and preventing reading difficulties is the accurate and timely identification of students with or at-risk for reading difficulties in order that additional instruction/intervention can be provided (Glover & Albers, 2007). As such, universal screening processes have become nearly ubiquitous in schools (Fuchs, Fuchs, & Compton, 2012).

When considering the tools and procedures that are utilized in the universal screening process, it is imperative to strike a balance between consideration of the diagnostic accuracy and psychometric characteristics of specific measures and issues of practicality for schools. That is, the feasibility and efficiency of the administration of universal screening measures, processes for decision-making, and prudent allocation of instructional resources. From the measurement perspective, effective screening tools demonstrate high levels of *sensitivity* in correctly identifying those students who will actually encounter difficulties, as well as high levels of *specificity* in the accurate identification of those who are not likely to demonstrate reading difficulties (Zhou et al., 2002). Ultimately, the goal is to maximize *classification accuracy,* a summative measure of the overall proportion of students who were correctly identified as at-risk or not at-risk on a screening measure. If the goal of universal screening is to promote the early identification of reading difficulties (or potential reading difficulties) screening measures that detect a large proportion of at-risk students would be desirable so that appropriate remedial support can be provided to students in order to prevent further difficulties (Frances et al., 1996). With that in mind, researchers have argued that high levels of sensitivity are necessary for universal screening measures (Compton et al., 2006; Jenkins, Hudson, & Johnson, 2007). Though consensus has not been reached regarding optimal levels of sensitivity, acceptable sensitivity values noted in the literature range from .70 to .90 (e.g., Catts et al., 2009; Compton et al., 2006; Jenkins et al., 2007; Kilgus et al., 2014). Relatedly, specificity levels of at least .70 are generally considered adequate for screening measures.

The above-mentioned indices of diagnostic accuracy are considered population-based as they are properties of the measure(s) itself. Several researchers argue that sample-based indices of diagnostic accuracy should also be considered (e.g., Christ & Nelson, 2014; Petscher, Kim & Foorman, 2011). These include *positive predictive value* and *negative predictive value*. The positive predictive value can be represented as the proportion of students identified as at-risk on a screening measure who ultimately fail the outcome test/assessment, while negative predictive value is the proportion of students who were identified as not at-risk during screening and subsequently passed the outcome test/assessment. These values are influenced by the actual number of students (i.e., base rate) who demonstrate risk on the outcome measure. When considering both the population and sample-based statistics of diagnostic accuracy for a given screening measure(s), educators and researchers must balance these indices with the actual needs and aims for the school or district. If the goal is an accurate and efficient direct route screening process in order to provide immediate intervention, as is common in many schools (e.g., Fuchs,

Fuchs, & Compton, 2012), sensitivity may take precedence. Meanwhile, if the goal is to begin by ruling out students who are not at-risk, so as not to include in any further monitoring or screening, maximizing the negative predictive value would be important. In sum, the accuracy of a screening measure(s) has significant implications for ensuring that schools are able to allocate increasingly limited resources to those students who are most at-risk for poor outcomes.

Though early attempts in the identification of children at risk for reading difficulties during the initial stages of reading development yielded considerable classification errors (e.g., Fletcher & Satz, 1984; Jenkins & O'Connor, 2002; Scarborough, 1998), recent advances have been made for screening in the primary grades (Jenkins et al., 2007; Speece et al., 2010). Typical screening measures for kindergarten and first grade involve assessing critical precursor reading skills such as phonemic awareness and letter-sound knowledge, while in Grades 2-3 screening generally includes assessment of student's accuracy and fluency in reading words and/or connected text (Jenkins et al., 2007). Measures of oral reading fluency (ORF) are the most prevalent of screening tools, as they have demonstrated predictive validity to later reading outcomes, including performance on high stakes assessments, are sensitive to growth over time, and are relatively easy to administer (Atkins & Cummings, 2011; Deno, 2003; Petscher et al., 2013; Speece & Ritchey, 2005). Despite increased knowledge relative to effective screening measures for identifying students in need of intervention, and mounting research demonstrating the efficacy of intervention in the earliest grades, (e.g., Denton, Fletcher, Anthony, & Francis, 2006; Fletcher, Lyon, Fuchs, & Barnes, 2006; Jenkins & O'Connor, 2002), the prevalence of students in the upper elementary grades demonstrating poor reading outcomes is well documented (NAEP, 2015). In order to alter the trajectory in reading performance for this group of students, schools must be able to accurately and efficiently identify those students with reading deficits in order to provide the necessary supports (Glover & Vaughn, 2010). The purpose of this research is to examine the accuracy of potential screening measures in the identification of students with reading problems in the upper elementary grades, namely fourth grade.

Screening in Upper Elementary Grades

There may be several reasons reading problems in the upper elementary grades. As Vaughn and colleagues (2008) have pointed out, some students either are not identified as at-risk or fail to receive intervention in the early grades and thus, reading difficulties persist. Further, some K-3 students are provided intervention but such supports are insufficient for remediating difficulties and/or they experience a recidivism of reading difficulty as demands increase in the later elementary grades (O’Connor & Sanchez, 2011; Vaughn et al., 2008). O’Connor and Sanchez (2011) refer to some students demonstrating a pattern of being in and out of intervention. Finally, there is a group of students who did not previously exhibit a level of reading deficit that suggested a need for intervention prior to the upper elementary grades – these students are considered to have late-emerging reading disabilities (LERD). Originally discussed by Chall (1983), these are students demonstrating seemingly grade-appropriate reading levels until encountering the more complex text and academic vocabulary in fourth grade. The literature now contains several studies that report anywhere from 13% to 46% of students with reading difficulties are not identified until after the primary grades (e.g., Badian, 1999; Catts, Compton, Tomblin, & Bridges, 2012; Leach, Scarborough, & Rescorla, 2003). Of particular note with regards to LERD, Kieffer (2010) found that students from low socioeconomic status may be at substantial risk for developing reading difficulties in the upper elementary grades and middle school. These results highlight that in addition to timely early identification, there is a need for continuous attention to the reading performance of students in all grades in order to provide the necessary instruction and intervention. Universal screening represents a critical avenue for identification of older students with reading difficulties.

The screening of students in the upper elementary grades should ideally follow the same principles as screening in the lower grades, namely utilizing measures that maximize the identification of students who are ultimately at-risk for poor reading outcomes. At the upper elementary level however, there are two specific realities that influence thinking about universal screening measures. First, reading instruction is heavily focused on the application of fundamental reading skills in order to comprehend increasingly complex texts; this is notably reflected in state and national curriculum standards (e.g., Common Core State Standards [CCSS], 2010). Second, upper elementary students in all states participate in a year-end reading assessment with often high stakes implications (e.g., grade level promotion). Whether these year-end assessments are state-specific outcome measures or assessments created from the PARCC or Smarter Balanced Assessment Consortiums, screening measures that reliably predict student performance on such key outcome assessments are essential.

Jenkins and colleagues (2007) highlighted the relative dearth of studies addressing this very question and to date, much still remains to be learned. As noted, ORF has been widely used as a screening and progress monitoring measure given its predictive validity for later outcomes, as well as ORF’s ease of administration (Shapiro, Solari, & Petscher, 2008). However, Shinn et al. (1992) found that in explaining reading performance, in comparison to the unitary factor at third grade, the components of fluency and reading comprehension represented two distinct factors at fifth grade. More recent studies have demonstrated that as students get older, the sensitivity and specificity of ORF in predicting outcomes on standardized comprehension measures and state reading assessments decreases (e.g., Johnson, Jenkins, & Petscher, 2010; Park et al., 2011; Petscher & Kim, 2011). Further, while research exists demonstrating the validity of ORF in predicting student performance on state reading assessments in the upper elementary and middle grades (e.g., Keller-Margulis, Shapiro, & Hintze, 2008; Park et al., 2011; Stage & Jacobson, 2001; Vander Meer, Lentz, & Stollar, 2005), a recent meta-analysis of the literature noted that in general, ORF only minimally exceeded acceptable diagnostic accuracy standards (Kilgus, Methe, Maggin, & Tomasula, 2014). Finally, both reviews of screening research (Jenkins et al., 2007; Kilgus et al., 2014) highlighted the concern over the classification accuracy of single measures, such as ORF, as predictors of reading outcomes beyond the early grades and noted a need to explore multivariate approaches.

Research on screening measures beyond the earliest grades has shown that the inclusion of a standardized reading comprehension measure along with ORF may enhance the prediction of performance on a state reading assessment for students in third, fourth, and fifth grades (Shapiro et al., 2008). Even when predicting performance on the state assessment in third grade, Johnson et al. (2010) found that a comprehension measure was the best predictor. Similarly, Ardoin and colleagues (2004) noted that while ORF was an adequate predictor of basic reading performance for third-grade students, the inclusion of a standardized measure of comprehension was a better predictor of year-end reading comprehension. It is important to note that findings from the above-mentioned studies reveal that inclusion of a measure of reading comprehension, either individually or combined with an ORF measure, does not always result in optimal classification accuracy. More specifically, large numbers of students are identified as at-risk who do not in fact actually have reading difficulty (i.e., false-positive), which may result in the allocation of limited intervention resources to students who do not require such support.

More recently, research has examined the use of computer-adaptive tests (CAT) as an alternative method for universal screening in the upper elementary grades (Klingbeil et al., 2017; Van Norman et al., 2017). CATs are generally group-administered and computer scored assessments that can be administered in similar or less time than required for individually administered screening measures (Klingbeil et al., 2017). Klingbeil and colleagues (2017), examining a CAT, ORF, and a running record measure in Grades 3 to 5, found that the CAT demonstrated the most promise as a single predictor of performance on a version of the Smarter Balanced summative assessment. In general, the CAT alone performed similar to a multivariate screening battery, though, in third and fifth grade, sensitivity was increased when using a combination of measures. Van Norman and colleagues (2017) also examined a CAT and student’s performance on the previous year state assessment as predictors of the current year state assessment outcome. Their findings suggested that the CAT and previous year outcome performed similarly as screening tools, with generally adequate specificity and sensitivity levels. The authors in both the Klingbeil et al. and Van Norman et al. studies note the limitation that the research was conducted with students who were predominately middle to upper-class and majority caucasian and suggest further screening research with more diverse samples.

The growing body of literature on screening at the upper elementary level is certainly promising, particularly evidence for the utilization of multivariate approaches. However, the extant studies reviewed have included as screeners only measures of reading fluency (ORF) and/or standardized assessments of text comprehension and language/vocabulary, either in the form of paper and pencil tests or computer-adapted measures. Older students with reading difficulties, however, do not always exhibit comprehension deficits absent of word-level deficits. Some demonstrate difficulties with both lower and higher level skills, while others have primarily phonological processing deficits that impact word reading (e.g., Catts et al., 2002; Compton et al., 2008; Leach et al., 2003). Thus, it would reason that a heterogeneous, multivariate battery, addressing lower and higher-level reading skills may be beneficial.

To date, Speece and colleagues (2010) are the lone study to evaluate a multivariate screening battery for older elementary age students which included a combination of word-level and higher-level skills. They found that a group-administered comprehension measure, individually-administered word reading fluency measure, and teacher ratings were most effective in predicting reading status, though status was measured via a latent reading factor rather than performance on a state reading assessment as in other studies. These findings lend some initial support for a screening approach that specifically incorporates an assessment of word-level skills. However, as noted by Speece et al, much more research in this area is needed to validate and extend their findings. Furthermore, the sample utilized included predominately Caucasian students attending parochial schools and thus, research with a sample representing a more diverse population of students is warranted.

Summary and Research Questions

Screening and identification of students with/at-risk for reading difficulties represent an important first step in RtI models, including students in upper elementary grades where there is a particularly large percentage of struggling readers (e.g., NAEP, 2015). Available research suggests a need for multiple measures rather than a single screening instrument. Though ORF and comprehension measures represent likely candidates, findings on resulting classification accuracy are somewhat equivocal (e.g., Johnson et al., 2010; Klingbeil et al., 2017; Shapiro et al., 2008: Van Norman et al., 2017) and thus, the addition of measures assessing word-level reading may enhance diagnostic accuracy (Speece et al., 2010). The present study endeavors to add to the emerging screening literature within the upper elementary grades by building upon and extending the promising findings from Speece et al. in a couple specific ways. First, we examined screening measures with a relatively diverse sample comprised primarily of students with low SES and from culturally and racially diverse backgrounds who may be at particular risk for late-emerging reading problems in need of remediation (Kieffer, 2010). Second, we investigated the utility of measures of both lower and higher-level reading skills in predicting students' actual performance on their respective state reading assessment. Accurate prediction of which students are at significant risk for not passing such high-stakes assessments is particularly relevant in the current educational era as these tests impact decision making and are used to judge the quality of instruction for schools and teachers. In the current research, we were able to examine screening measures in relation to two separate state assessments. The primary research questions were:

(1) Using a direct-route screening approach, what is the classification accuracy for individual measures of reading comprehension, oral reading fluency, and word-level fluency for predicting the performance of fourth-graders from diverse backgrounds on their year-end state reading assessment?

(2) Is classification accuracy improved through the use of a multivariate screening approach utilizing measures of lower and higher-level reading skill in fourth-grade?

Method

Participants

The participants included 321 students from 31 fourth-grade classrooms. These classrooms were located within 10 public elementary schools in four school districts in Florida (FL) and Texas (TX). At the FL site, there were 188 students, of which 17% identified as Hispanic ethnicity and two-thirds were minority status. More specifically, 54% of students were African American, 6% identified as American Indian, 3% were Asian, 3% were considered multi-racial, with less than 1 percent considered Pacific Islander. Further, 66% of the students were identified as low socioeconomic status via their participation in free or reduced price lunch programs, 5% as English Language Learners, and 17% were eligible for special education services. Female students comprised 52% of the sample in Florida.

The sample in TX was comprised of 133 students. Of the students in TX sample, 68% identified as Hispanic. With regards to race, 38% were Caucasian, 35% American Indian, 23% African American, and approximately 1% each were Asian and Pacific Islander. A vast majority (91%) were identified as from low-SES households, while 19% were English Language Learners and 5% eligible for special education. Females comprised 56% of the sample in TX.

All of the participants were part of a larger randomized control trial designed to examine the efficacy of a multi-component reading intervention program (Wanzek et al., 2016). This included students identified as struggling readers (*n* = 221), as evidenced by performance at or below the 30th percentile on a measure of reading comprehension (MacGinitie et al., 2006), as well as a random sample of 100 typical readers (i.e., performance above the 30th percentile).

Screening Measures

All students were assessed in early fall of their fourth-grade year with several reading measures across a variety of skills, including reading comprehension and word, sentence, and text-level reading fluency. As a key feature of a screening measure is its ability to accurately classify students as at risk or not at risk for poor outcomes while also being efficient (Jenkins et al., 2007), we specifically included group-level assessments and individual-level fluency measures, which could be administered without compromising too much instructional time. Each assessment administered has been purported by the respective publisher as appropriate for screening and identification of students in need of more intensive reading instruction.

Gates-MacGinitie Reading Test. (GMRT; **MacGinitie et al., 2006**). The GMRT is a group-administered, norm-referenced assessment of reading. The reading comprehension subtest, in which students are presented with multiple paragraph-length passages and required to respond to related multiple-choice questions, was administered. The reading passages include both narrative and expository text. Students are allowed a total of 35 min to take this assessment. For fourth-grade students, test-retest reliability coefficients are above .85; alternate-form reliability is .86 at this level. Construct validity estimates range from .79-.81.

Test of Silent Reading Efficiency and Comprehension (TOSREC; Wagner, Torgesen, Rashotte, & Pearson, 2010). The TOSREC is a brief, group or individually administered measure of silent reading efficiency (i.e., speed and accuracy) of connected text for comprehension. Students in this sample were administered the TOSREC in a group setting.

Students are given 3 min to silently read and verify the accuracy of as many sentences, ranging in length from 4-10 words and increasing in complexity of content, grammar, and vocabulary, as possible. Alternate form reliability for the TOSREC is .86 for fourth grade. Predictive validity of the TOSREC has been examined in relation to student performance on the Florida Comprehensive Assessment Test. At fourth grade, correlations of all forms of the TOSREC with the FCAT range from .55 - .73.

Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999). The TOWRE is a standardized, individually-administered timed test of single-word reading fluency wherein students are given 45 seconds to read a list of words. The Sight Word Efficiency (SWE) and Phonemic Decoding Efficiency (PDE) subtests were administered. SWE assesses real word reading while PDE measures reading of decodable nonsense words. A student's raw score for each subtest is the number of words read correctly within the allotted time. Test-retest reliability coefficients range from .83-.96 for fourth graders on the SWE and PDE. Concurrent validity for SWE and the Word Identification subtest of Woodcock Reading Mastery Tests-Revised (WRMT-R) is .89. For PDE and the word attack subtest of WRMT-R, concurrent validity is estimated at .86.

Dynamic Indicators of Basic Early Literacy Skills -6th Edition (DIBELS; Good & Kaminski, 2002). The oral reading fluency (ORF) subtest from DIBELS was administered to measure student's ability to read connected text with speed and accuracy. Students read three separate passages aloud for one minute each. The total number of correct words read per minute is recorded for each passage, and the median score of the passages is used to indicate the achieved level of fluency. Test-retest reliabilities for ORF with elementary students range from

.92 to .97; alternate-form reliability across passages from the same level is reported as .89 to .94. At Grade 4, concurrent validity was .74 with the Group Reading Assessment and Diagnostic Evaluation measure and .89 with the NAEP reading assessment.

Outcome Measure

All students participated in their respective state's reading achievement test administered in the spring of fourth grade. These measures were administered by school personnel.

Florida Comprehensive Assessment Test 2.0 (FCAT; Florida Department of Education, 2014). The FCAT 2.0 Reading assessment is Florida's state achievement test, administered in Grades 3 through 10, measuring students' attainment of established content standards. The FCAT 2.0 Reading test is a standardized, criterion-referenced test that presents students with both literary (50%) and informational (50%) passages and multiple-choice questions that evaluate comprehension of text and vocabulary. Reading passages at the fourth-grade level average approximately 500 words in length. Scores on the FCAT 2.0 are reported in terms of developmental scale scores (DSS) which allow for comparison of performance and progress in reading achievement across adjacent grade levels; at Grade 4, DSS range from 154 to 269. Developmental scale scores are translated into achievement levels that describe student's overall level of proficiency in meeting assessed content standards. Achievement levels on FCAT 2.0 range from Level 1 (lowest) to Level 5 (highest). Students must achieve Level 3 (DSS of at least 208) or higher in order to be considered proficient. Cronbach's alpha reliability coefficients for FCAT 2.0 Reading range between .89 and .93.

State of Texas Assessments of Academic Readiness (STAAR; Texas Education Agency, 2013). The STAAR Reading test is the state achievement test designed to assess student knowledge of the Texas Essential Knowledge and Skills (TEKS) standards. It is a standardized and criterion-referenced assessment of reading. The assessment involves presenting students with literary and informational tests and requires them to answer multiple-choice items tapping both comprehension and vocabulary. Raw scores on the STAAR tests are converted to scaled scores to allow for comparison across grade levels. Ultimately, student performance on the STAAR is translated into an academic performance standard that represents the degree to which expected standards have been met. Established performance standards are Level 1 (Unsatisfactory Academic Performance), Level 2 (Satisfactory), and Level 3 (Advanced). At Grade 4, the minimum scale score is 816 while the maximum score is 1922. A scaled score of at least 1422 is required for Level 2 performance. Internal consistency for the STAAR reading assessment at fourth-grade was reported to be .89.

Procedures

Data Collection. The data from the proposed screening measures utilized in the present study were collected by trained research staff in September/October (~ 5th to 8th week of school) of Grade 4. Assessment order was counterbalanced and all staff were required to demonstrate 100% accuracy in administration and scoring prior to actual field testing. Each measure was also double-scored by another member of the research staff. In general, assessment took place in two separate sessions. The state assessments of reading (i.e., FCAT 2.0, STAAR) were administered between mid-March to early April.

Data Analyses. Due to the differences in the outcome measure across states, all data analyses described below were conducted separately for the samples in FL and in TX. Initial analysis of these data was conducted using logistic regression to investigate each of the screening measures as an individual predictor of student performance (pass-fail) on their respective state year-end reading assessment. These analyses provided specificity and sensitivity levels, as well as overall classification accuracy, or the percentage of the total sample correctly identified, for each of the potential screening measures (i.e., predictors). Additionally, an analysis was conducted predicting the outcome by entering all of the predictors simultaneously into the logistic regression model; such a model allowed for the determination of whether a combination of screening measures would improve classification accuracy.

While general classification accuracy of an individual, or combined, predictor(s) is important to determine, ultimately the primary goal of any academic screening process is to maximize the extent to which students actually in need of intervention are identified. Thus, further data analyses were conducted to identify specific cut points associated with desired sensitivity levels. Given the potential poor outcomes for older students whose reading difficulties are not remediated (e.g., Frances et al., 1996), it was felt that optimizing sensitivity at .90 would maximize the probability of identifying struggling readers. Receiver operating characteristic (ROC) curve analysis was utilized for this purpose. ROC curves provide a plot of the true positive rate against the false positive rate so that potential combinations of sensitivity and specificity can be analyzed (Pepe et al., 2004). The area under the curve (AUC) value, a probability index, is generated during ROC curve analysis and provided an indicator of general diagnostic accuracy. As values of the AUC approach 1.0, the screening measure can be said to reliably discriminate between students with satisfactory (i.e., passing) and unsatisfactory performance on the outcome while AUC values near .50 indicate no better than chance prediction (Zhou et al., 2002). Compton and colleagues (2006) suggest AUC values above .90

represent excellent diagnostic accuracy, between .80 and .90 good, .70 to .80 fair, and values under .70 are considered poor. We generated ROC curves for each individual predictor, identified cut scores associated with .90 sensitivity level, and subsequently, determined the number of students classified as true positives, false negatives, true negatives, and false positives. In addition to computing overall classification accuracy, specificity, the positive predictive value, and negative predictive value were also calculated. In order to generate a ROC curve for the combined measures model, the predicted probabilities from the logistic regression analysis were utilized. This procedure was used in order to address the inherent difficulty with interpretation of multiple cut scores from combined measures in the screening process. We set alpha at .05 for the analyses within each site (individual regressions and multiple regression). To adjust for multiple comparisons when running the logistic regression for each individual predictor, we applied a Bonferroni correction such that p-values below .01 (alpha = .05 / 5 comparisons) were considered significant. Analyses were conducted using SPSS Version 24.

Results

Descriptive statistics, by state, for each of the predictor measures and their bivariate correlations are provided in Tables 1. With regards to the respective end-of-year state reading assessments, the base rate of students not achieving the state-specified proficiency levels was 58% and 43% respectively, within the FL sites and TX sites. It is important to note that these figures are not directly comparable given the potential differences in the reading assessments across the two states.

Classification Accuracy of Individual & Combined Measures

FL site. Using logistic regression, each of the individual screening measures was a significant predictor of the FCAT (*p*s < .01). Using the GMRT resulted in the highest classification accuracy (74.4%). Taken individually, all of the other measures resulted in classification accuracy values of less than 70%; the PDE measure of the TOWRE demonstrated the lowest accuracy (61.1%). Across these predictors, only the GMRT demonstrated an optimal sensitivity rate (.90), while the other measures had sensitivity rates ranging from .75 to .78. However, all individual screening measures had poor levels of specificity with rates of .38 to .55. In order to determine whether a multivariate approach would improve diagnostic accuracy, all predictor measures were included in a logistic regression model. In this model, only the GMRT and the ORF measure were significant (*p* < .05). The resulting classification accuracy (75.9 %) showed marginal improvement over the GMRT alone; this represented a 1.5% increase in accuracy using a multivariate approach. While sensitivity using this approach was acceptable at .88, the corresponding specificity level of .61 was below recommended values. In summary, whether employing a single screening measure or combined measures, between 24 and 39% of students were not properly classified when using the chosen screening instruments. See Table 2 for a summary of findings.

TX site. Results from the logistic regression analysis using both an individual and multivariate approach are provided in Table 2. Again, each of the individual measures was a significant predictor (*p* < .01) of the end-of-year STAAR measure. Similar to the findings from the FL sites, the GMRT demonstrated the strongest classification accuracy (80.5%) with generally adequate sensitivity and specificity levels. The ORF measure also demonstrated minimally acceptable sensitivity (.70) and adequate specificity (.74), while correctly classifying75% of students. The classification accuracy of the SWE measure was 71.2% while both the TOSREC (69.5%) and PDE (67.4%) had accuracies below 70%. In contrast to the FL site, these individual measures showed higher levels of specificity (.78 - .84) in predicting performance on the state reading assessment, with lower levels of sensitivity (.54 - .75).

When using logistic regression to examine a multivariate approach, only the GMRT measure was significant. The resulting classification accuracy was 78.9%, which actually represented a slight decrease (1.6%) in comparison to using the GMRT individually as a screener. Sensitivity level using this combined approach was .73 with specificity at .83.

Classification Accuracy with Maximized Sensitivity Levels

Presented in Table 3 are results, by site, from the ROC curve analyses, along with cut scores associated with a sensitivity level of ~ 90%, and multiple indices of classification accuracy. Results are presented for individual and combined screening measures. Figure 1 presents the resulting ROC Curves, for individual measures and the multivariate approach, across the two sites.

FL site. With this optimal sensitivity, specificity rates for individual predictor measures ranged from .15 (PDE) to .53 (GMRT). The AUC values for the individual measures ranged from .642 to .806. The GMRT demonstrated the most optimal classification accuracy followed by the ORF measure, though both were below 75%. On the GMRT, the cut score corresponding to a sensitivity of .90 was the 28th percentile. Meanwhile, on the SWE (SS ≤ 102), PDE (SS ≤ 103) and the TOSREC (SS ≤ 100), cut scores for establishing risk were above the 50th percentile for each respective measure. Further, to achieve .90 sensitivity using the ORF measure, all students reading at or below 116 correct words per min would be at-risk; the published norms for

DIBELS ORF indicate the fall benchmark for students in fourth grade is 93 correct words read per min. As a result of setting sensitivity at .90, though we are able to identify nearly all students at-risk for not passing the end-of-year state reading assessment; large numbers of those identified at-risk actually passed the state assessment (19% to 34% across measures). The issue of false positive is also indicated in the range of PPV from .59 to .72. We also conducted ROC curve analysis with the combined predictors. The AUC value was .826. Setting sensitivity at .90, the resulting specificity level was .59 and this approach correctly classified just over three-quarters (77.1%) of students. This approach surpassed the accuracy of all single screening measures though only by about 2.7% over the GMRT alone; the PPV of .74 suggests that approximately one-quarter of students that would be identified as at-risk via the multivariate approach actually demonstrated proficiency on the FCAT 2.0.

TX site. In TX, when sensitivity of the individual measures was optimized at .90, specificity rates for individual predictor measures ranged from .34 (SWE) to .55 (GMRT). The AUC values for the individual measures were between .736 and .846. Similar to FL, using the GMRT individually resulted in the highest classification accuracy (69.9%); all individual measures had accuracy levels below 70%. On the GMRT, the cut score corresponding to a sensitivity of .90 was the 31st percentile. Cut scores on the other norm-referenced measures and the ORF assessment were such that performance in the average range (or in the benchmark range on ORF) could classify a student as at-risk. The resulting PPV ranged from .49 to .60, while NPP values were all above .86 with the exception of SWE (.79). When we conducted ROC curve analysis with the combined predictors, the resulting AUC value was .861. The specificity level was .73 using the multivariate approach. Using this approach, there was an increase in classification accuracy of nearly 10% over using any of the individual predictors; accuracy of this combined approach was 79.7%. Of those identified at-risk, 70% (PPV = .70) demonstrated a lack of proficiency on the STAAR. Conversely, NPV was .90 suggesting that 90% of students considered not at-risk using the multivariate approach were in fact, proficient on the STAAR at the end of fourth grade.

Discussion

In this retrospective study we sought to examine the diagnostic accuracy, both individually and as part of a multivariate screening battery, of word level and text level reading screening measures for upper elementary students. We examined the classification accuracy of these beginning of the year measures on end of the year state reading comprehension outcomes for a diverse sample of students in two states. Results demonstrated that all individual measures were significant predictors of student outcomes of both state reading tests. However, despite optimization of sensitivity of measures, other indices (i.e., specificity, PPV, NPV, CA) were indicative of generally inadequate diagnostic accuracy. The GMRT, a group-administered reading comprehension measure, had the highest overall classification accuracy across FL and TX though again, with poor specificity levels. The NPVs (.80- .88) for the GMRT across states suggested it performed reasonably well in determining students who were not at-risk for failing their respective year-end, high-stakes reading assessment. Adding other measures, specifically those assessing word and text-level fluency, to the prediction model that optimized sensitivity improved overall classification accuracy slightly over the GMRT alone, but specificity remained poor in FL (.59) and was at a minimally acceptable level in TX (.73). Once again, NPVs were promising, but PPVs of < .75 suggested many students identified as at-risk via this multivariate battery would actually go on to pass the state assessments (i.e., false positives). In the multivariate logistic regression models, only GMRT was significant for both samples; ORF was also significant in FL while not other individual measures were significant in TX.

Adequacy of Screening with Single Measures

Given that schools commonly employ only a single measure within the universal screening process (Jenkins et al., 2013), it is important to consider whether the present findings lend support for one specific measure at the upper elementary level. The relative strength of the GMRT measure for predicting the state reading comprehension tests may be expected given that both tests are measuring the same construct, reading comprehension, in a similar way with group-administered independent reading of passages followed by reading comprehension questions. Despite vastly different samples, this finding in our study is in line with recent studies by Klingbeil et al. (2017) and Van Norman et al. (2017), which found that a CAT assessing reading comprehension and related skills was an adequate predictor of year-end state assessments of reading. Of note, both sensitivity and specificity of the CAT in the above studies was above .80 while specificity of the GMRT in this study was .53 to .55. The diagnostic accuracy when using ORF as a single screener was limited, particularly when utilizing a cut point that optimized sensitivity. Specificity of ORF was markedly better in TX (.54) in comparison to FL (.31), as was NPV. Our findings regarding ORF as a single predictor was somewhat surprising given that the outcome measure required students to read lengthy passages within a timed assessment and the fact that previous screening research demonstrating this measure as having at least minimal diagnostic accuracy when predicting to year-end state assessments (Kilgus et al., 2014). Nonetheless, the limited diagnostic accuracy of ORF as found in this study does seem to highlight potential concerns with using it as a single screening measure at this level (e.g., Jenkins et al., 2007). A relatively clear finding in this study was the fact that measures of single word reading efficiency do not represent viable options as individual screeners in upper elementary. These measures has the lowest values for AUC, specificity, PPV, and NPV. This is not surprising given the nature of what skills are specifically assessed in year-end state assessments and in fact, the primary goal for their inclusion in this study was to examine whether such measures increased diagnostic accuracy in a multivariate approach (see below).

In making sense of the findings for individual measures for screening at this level, particularly with regards to the GMRT, several considerations must be taken into account. First, an argument can certainly be made that it would actually be less costly to over-identify students as at-risk, thereby potentially providing unnecessary reading intervention, then to not identify a student who is likely to fail a subsequent high-stakes assessment (and remains at increased risk for poor educational outcomes [Frances et al., 1996]). However, our results suggest that upwards of 30 to 40% of students would be misidentified as at-risk on the GMRT. For many schools and districts, this may serve to tax already limited intervention resources and may create problems for implementation, which is critical given the document challenges in implementing Tier 2 interventions (e.g., Hoover et al., 2008). On the other hand, the NPVs for GMRT suggest that this measure has potential for ruling out students who are not at-risk. So, while using this measure in a direct route approach to screening is of some concern, employing the GMRT as a first step in “gated” screening may have potential. Gated screening processes have been the subject of recent research (Klingbeil et al., 2017; Van Norman et al., 2016) though findings suggest that this process may result in unacceptable numbers of false negatives, or students truly at-risk who are not identified as such.

Multivariate Screening

Given the established concerns regarding single screeners, including the present findings, it is important to consider the diagnostic accuracy when using multiple measures for universal screening, as well as potential implications. When sensitivity rates were set high, including measures of word and text-level fluency along with the GMRT, classification accuracy was improved, primarily via the reduction in the number of students misidentified as at-risk. This was mostly evident in TX as demonstrated by a nearly 10% improvement in classification accuracy when all measures were included: conversely, the increase in classification accuracy (2.7%) for predicting outcomes for the FL test were negligible in comparison to GMRT alone. The result for using a multivariate approach in TX should not be understated, as these findings suggest that all resulting diagnostic accuracy indices would be considered within acceptable standards.

Our results seem to align with previous research that has reported mixed findings on the degree of improved accuracy in identifying students as at-risk or not when employing a multivariate approach (e.g., Johnson et al., 2010; Klingbeil et al, 2017). As we specifically aimed to extend the work of Speece and colleagues (2010), it is important to consider the present findings within that context. Again, findings are mixed, with only the analyses in TX confirming the potential importance of adding word-level screening measures. Notably, our screening was in the early fall as compared to the late fall screening conducted in Speece et al., which also allowed them to include a teacher rating of the students as part of their model because teachers had seen the students in several months of instruction and possibly assessment.

So, while we may conclude that in FL the GMRT measure alone would be just as accurate within the context of universal screening as a multivariate battery, whether such a battery would ultimately be most beneficial to schools in TX requires further consideration. Namely, including multiple measures requires more time be set aside for screening and the inherent difficulty in interpreting student performance across several measures. With regards to time, the GMRT alone can be administered to large groups of students in approximately 35 to 40 minutes and can be machine scored, making it a very efficient screener. While the TOSREC can be group-administered and takes approximately five minutes, adding the individual measures would require an additional 10 minutes per student plus time for scoring. Once again, given these potential concerns, utilization of the GMRT first, within a gated approach may be warranted. This is especially true given the fact that the multivariate approach does no better in ruling out students who are not at-risk across either FL or TX when compared to GMRT alone.

Limitations and Future Research

The present findings should be interpreted in light of several limitations. First, given the retrospective nature of this study, we were only able to examine measures administered during the original study. Specifically, this precluded us from capturing information on students' achievement on the previous year (i.e. Grade 3) state reading assessment. Though a limitation of previous research as well, two recent screening studies in upper elementary and middle school grades found that this data demonstrated adequate diagnostic accuracy in predicting student's current year performance (Denton et al., 2011; Van Norman et al., 2017). Thus, future research should include this data whenever possible to further clarify how such information can add to prediction models whether in a multivariate approach and/or as a first step in a gated screening approach. While such information ideally increases the efficiency of the screening process given that state assessment data for students would be readily available and require no additional time for assessment, there is at least one important caveat. In a diverse sample of students and schools, such as found in this study, there are many students for whom earlier grade reading achievement information is not available due to movement between districts or states.

Second, as with any large sample, the core instruction and interventions that students struggling with reading received throughout the year varied within and across schools. Thus, it is not possible to add specific instructional variables to the models, though researchers must acknowledge that this practical limitation creates unknown noise in school decision making.

Third, though assessments aligned with the CCSS (2010) are being developed through state-level consortiums, more than half (27) of states continue to use their own state developed, high stakes assessments. Thus, while we were able to examine data in two states, classification accuracy will likely differ in other states even when the same construct of reading comprehension is being measured. Additional research examining factors in state tests that can make classification accuracy more direct and straightforward for schools would make a significant practical contribution in the field. For example, state assessments may differ in the extent to which they directly assess specific subskills under the umbrella of reading comprehension (e.g., main idea, cause and effect, etc.) and thus, more targeted rather than broad screening measures of comprehension may serve as better predictors of student performance on these high-stakes measures. The less than optimal results of this study and others particularly in the area of specificity, suggest the need for continued research on identification of older students with reading difficulties. A multitude of factors including previous experiences as well as differing areas of reading strength and weakness can contribute to reading difficulties as students progress through the grades. Continued research in this area with the goal of providing educators effective and efficient methods for identifying students in need of intervention as well as informing instructional decisions is needed.

Summary of Implications

The current study was unable to identify any single screening measure with acceptable levels across all diagnostic indices, including sensitivity, specificity, PPV, and NVP, in predicting student outcomes of pass or fail on two different state reading comprehension tests for upper elementary students. The GMRT did consistently provide the strongest prediction with fair to good sensitivity, good AUC, and overall classification accuracy at 70% or above. As such, it, may provide an efficient and feasible screening method for schools at this time, particularly when the focus is on maximizing the identification of students truly at-risk. Nonetheless, the potential for misallocation of scare resources must be considered when using the GMRT as a direct route screener. Preliminary evidence provided support for a multivariate screening approach including measures of comprehension along with word and text-level fluency in TX though must be considered in light of issues with time/efficiency and interpretation of multiple measures. Ultimately, it is imperative that schools purposefully reflect on the primary purpose of their universal screening efforts, the population of students being served, and available resources when considering the recommendations and implications from the present findings.

References

Ardoin, S. P., Witt, J. C., Suldo, S. M., & Connell, J. E. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psychology Review*, *33*, 218.

Atkins, T. A., & Cummings, K. D. (2011). Utility of oral reading and retell fluency in predicting proficiency on the Montana comprehensive Assessment System. *Rural Special Education Quarterly*, *30*, 3–12.

Badian, N. A. (1999). Reading disability defined as a discrepancy between listening and reading comprehension. *Journal of Learning Disabilities*, *32*, 138–148. doi:10.1177/002221949903200204

Catts, H. W., Compton, D., Tomblin, J. B., & Bridges, M. S. (2012). Prevalence and nature of late-emerging poor readers. *Journal of Educational Psychology*, *104*, 166–181. doi:10.1037/a0025323

Catts, H., Fey, M., Tomblin, J., & Zhang, X. (2002). A longitudinal investigation of reading outcomes in children with language impairments. *Journal of Speech, Language and Hearing Research*, *45*, 1142–1157. doi:10.1044/1092-4388(2002/093)

Catts, H. W., Petscher, Y., Schatschneider, C., Sittner Bridges, M., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities*, *42*, 163-176. doi:10.1177/0022219408326219

Christ, T. J., & Nelson, P. M. (2014). Developing and evaluating screening systems: Practical and psychometric considerations. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A.

Feeney-Kettler (Eds.), Universal screening in educational settings: Evidence-based decision making for schools (pp. 79–110). Washington DC: American Psychological Association.

Chall, J. S. (1983). *Stages of reading development*. New York: McGraw- Hill.

Common Core State Standards Initiative. (2010). Common core state standards for English language arts and literacy in history/social studies, science, and technical subjects. Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, *98*, 394–409. doi:10.1037/0022-0663.98.2.394

Compton, D.L., Fuchs, D., Fuchs, L.S., Elleman, A.M., & Gilbert, J.K. (2008). Tracking children who fly below the radar: Latent transition modeling of students with late-emerging reading disability. *Learning and Individual Differences, 18*, 329-337. doi:10.1016/j.lindif.2008.04.003

Deno, S.L. (2003). Developments in curriculum-based measurement. *Journal of Special Education, 37,* 184-192.

Denton, C. A., Barth, A. E., Fletcher, J. M., Wexler, J., Vaughn, S., Cirino, P. T., ... & Francis, D. J. (2011). The relations among oral and silent reading fluency and comprehension in middle school: Implications for identification and instruction of students with reading difficulties. *Scientific Studies of Reading*, *15*, 109-135.

Denton, C.A., Fletcher, J.M., Anthony, J.L. & Francis, D.J. (2006). An evaluation of intensive interventions for students with persistent reading difficulties. *Journal of Learning Disabilities, 39*, 447-466.

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2006). *Learning disabilities: From identification to intervention.* New York: Guilford Press.

Fletcher, J., & Satz, P. (1984). Test-based versus teacher-based predictions of academic achievement: A three-year longitudinal study. *Journal of Pediatric Psychology, 9*, 193-203.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88*, 3-17. doi:10.1037/0022-0663.88.1.3

Fuchs, D., Fuchs, L. S., & Compton, D. L. (2012). Smart RTI: A Next-Generation Approach to Multilevel Prevention. Exceptional Children, 78, 263–279.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. Journal of School Psychology, 45, 117-135.

Glover, T. A., & Vaughn, S. (Eds.). (2010). *The promise of response to intervention: Evaluating current science and practice*. Guilford Press.

Good, R. H., & Kaminski, R. A. (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Hoover, J. J., Baca, L., Wexler-Love, E., & Saenz, L. (2008). National implementation of response to intervention (RTI). Retrieved from http://www .nasde.org/Portals/0/NationalImplementationof RTI-ResearchSummary.pdf

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*, 582-599.

Jenkins, J. R., & O'Connor, R. E. (2002). Early identification and intervention for young children with reading/learning disabilities. In R. Bradley, L. Danielson, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 99–149). Mahwah: Erlbaum.

Jenkins, J. R., Schiller, E., Blackorby, J., Thayer, S. K., & Tilly, W. D. (2013). Responsiveness to intervention in reading: Architecture and practices. Learning Disability Quarterly, 36, 36-46. doi:10.1177/0731948712464963

Johnson, E. S., Jenkins, J. R., & Petscher, Y. (2010). Improving the accuracy of a direct route screening process. *Assessment for Effective Intervention*, *35*, 131-140. doi:10.1177/1534508409348375

Johnson, E., Mellard, D. F., Fuchs, D., & McKnight, M. A. (2006). *Responsiveness to intervention (RTI): How to do it.* National Research Center on Learning Disabilities. Retrieved from http://search.proquest.com/docview/62012767?accountid=4840

Kieffer, M. J. (2010). Socioeconomic status, english proficiency, and late-emerging reading difficulties. *Educational Researcher, 39*, 484–486. doi:10.3102/0013189X10378400

Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review*, *37*, 374.

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of school psychology*, *52*, 377-405.

Klingbeil, D.A., Nelson, P.M., Van Norman, E.R., & Birr, C. (2017). Diagnostic accuracy of multivariate universal screening procedures in upper elementary grades. Remedial and Special Education, 38, 308-320. Doi:10.1177/0741932517697446

Leach, J., Scarborough, H., & Rescorla, L. (2003). Late-emerging reading disabilities. *Journal of Educational Psychology*, *95*, 211–224. doi:10.1037/0022-0663.95.2.211

MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L. G., & Hughes, K. E. (2006). *Gates-MacGinitie Reading Tests* (4th ed.). Rolling Meadows, IL: Riverside Publishing.

National Assessment of Educational Progress. (2015). *The nation's report card*. Retrieved from http://www.nationsreportcard.gov/ reading_math_2015/#reading?grade=4

O'Connor, R.E. & Sanchez, V. (2011). Responsiveness to intervention models for reducing reading difficulties and identifying learning disability. In Kauffman, J.M. & Hallahan, D.P. (Eds.), Handbook of Special Education (pp. 123-133). New York: Routledge.

Park, B. J., Anderson, D., Irvin, P. S., Alonzo, J., & Tindal, G. (2011). *Diagnostic efficiency of easyCBM reading: Oregon* (Technical Report No. 1106). Eugene, OR: Behavioral Research and Teaching, University of Oregon

Pepe, M., Janes, H., Longton, G., Leisenring, W., & Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*, 159, 882-890.

Petscher, Y., Cummings, K. D., Biancarosa, G., & Fien, H. (2013). Advanced (Measurement) applications of curriculum-based measurement in reading. *Assessment For Effective Intervention*, *38*, 71-75. doi:10.1177/1534508412461434

Petscher, Y., & Kim, Y.-S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology*, 49, 107–129. doi:10.1016/j.jsp. 2010.09.004

Petscher, Y., Kim, Y. S., & Foorman, B. R. (2011). The importance of predictive power in early screening assessments: Implications for placement in the response to intervention framework. *Assessment for Effective Intervention*, *36*, 158-166. doi:10.1177/1534508410396698

Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), *Specific reading disability: A view from the spectrum* (pp. 75–119). Timonium: York Press.

Shapiro, E., Solari, E., & Petscher, Y. (2008). Use of a measure of reading comprehension to enhance prediction on the state high stakes assessment. *Learning & Individual Differences, 18*, 316-328. doi:10.1016/j.lindif.2008.03.002

Shinn, M. R., Good, R. H., & Knutson, N. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*, 459−479.

Speece, D. L., & Ritchey, K. D. (2005). A longitudinal study of the development of oral reading fluency in young children at risk for reading failure. *Journal of Learning Disabilities, 38*, 387-399.

Speece, D. L., Ritchey, K. D., Silverman, R., Schatschneider, C., Walker, C. Y., & Andrusik, K. N. (2010). Identifying children in middle childhood who are at risk for reading problems. *School Psychology Review*, *39*, 258–276.

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*, 407-419.

Texas Education Agency. (2013). *State of Texas Assessments of Academic Readiness (STAAR) Standard Setting Technical Report*. Retrieved from www.tea.state.tx.us.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficency*. Austin, TX: Pro-Ed.

Vander Meer, C. D., Lentz, F. E. & Stoller, S. (2005). *The relationship between oral reading fluency and the Ohio proficiency testing in reading* (Technical Report). Eugene, OR: University of Oregon.

Van Norman, E. R., Nelson, P. M., & Klingbeil, D. A. (2017). Single measure and gated screening approaches for identifying students at-risk for academic problems: Implications for sensitivity and specificity. *School Psychology Quarterly, 32*, 405. doi:10.1037/sp10000177

Vaughn, S., Fletcher, J.M., Francis, D.J., Denton, C.A., Wanzek, J., Wexler, J., … Romain, M.A.. (2008). Response to intervention with older students with reading difficulties. *Learning and Individual Differences, 18,* 338-345. doi:10.1016/j.lindif.2008.05.001

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. (2010). *Test of silent reading efficiency and comprehension.* Austin, TX: PRO-Ed.

Wanzek, J., Al Otaiba, S., Petscher, Y., Kent, S.C., Schatschneider, C., Haynes, M., .. & Jones, F.G. (2016). Examining the average and local effects of a standardized treatment for fourth graders with reading difficulties. Journal of Research on Educational Effectiveness, 9, 45-66. doi:10.1080/19345747.2015.111603

Zhou, X. H., Obuchowski, N. A., & Obuchowski, D. M. (2002). *Statistical methods in diagnostic medicine.* New York, NY: Wiley.

Table 1. *Descriptive Statistics and Correlations for Measures*

		Florida Sample				
Measure	GMRT	TOSREC	ORF	SWE	PDE	FCAT
GMRT	1.0					
TOSREC	.69	1.0				
ORF	.61	.64	1.0			
SWE	.53	.59	.87	1.0		
PDE	.48	.50	.76	.78	1.0	
FCAT	.67	.57	.55	.45	.36	1.0
Mean	455.36	87.38	92.77	90.49	86.89	204.80
SD	31.62	14.24	30.07	13.76	14.52	18.25
n	188	181	186	186	186	176
		Texas Sample				
	GMRT	TOSREC	ORF	SWE	PDE	FCAT
GMRT	1.0					
TOSREC	.77	1.0				
ORF	.74	.71	1.0			
SWE	.59	.57	.87	1.0		
PDE	.62	.57	.84	.85	1.0	
STAAR	.65	.59	.62	.53	.52	1.0
Mean	460.85	86.35	90.93	91.86	89.81	1453.88
SD	37.81	15.06	35.02	14.60	15.83	119.12
n	133	128	132	132	132	133

Note. GMRT = Gates-MacGinitie Reading Test; TOSREC = Test of Silent Reading Efficiency and Comprehension; ORF = oral reading fluency; SWE = Sight Word Efficiency; PDE = Phonemic Decoding Efficiency; FCAT = Florida Comprehensive Assessment Test 2.0; STAAR = State of Texas Assessments of Academic Readiness.

Table 2. *Logistic Regression of Single and Combined Measures*

	FLORIDA					TEXAS				
Predictor	*B (SE)*	*Wald*	Sensitivity	Specificity	CA	*B (SE)*	*Wald*	Sensitivity	Specificity	CA
GMRT	-.052 (.009)	31.26	.90	.53	74.4	-.043 (.009)	23.82	.75	.84	80.5
TOSREC	-.066 (.014)	22.93	.78	.53	67.1	-.086 (.019)	21.53	.58	.78	69.5
ORF	-.036 (.007)	26.57	.78	.55	68.6	-.041 (.008)	26.62	.70	.79	75.0
SWE	-.062 (.014)	18.63	.75	.49	64.0	-.076 (.017)	20.83	.57	.82	71.2
PDE	-.038 (.012)	9.92	.78	.38	61.1	-.064 (.015)	19.34	.54	.78	67.4
Combined Predictors	*B (SE)*	*Wald*	*p*	Sensitivity/ Specificity	CA	*B (SE)*	*Wald*	*p*	Sensitivity/ Specificity	CA
GMRT	-.045 (.011)	15.57	< .001			-.037 (.013)	7.98	.005		
TOSREC	.003 (.02)	.031	.861			-.018 (.025)	.513	.474		
ORF	-.039 (.015)	6.57	.01			-.020 (.018)	1.25	.263		
SWE	.005 (.031)	.031	.861			-.003 (.038)	.006	.938		
PDE	-.039 (.022)	3.05	.081			.015 (.028)	.278	.598		
				.88 / .61	75.9				.73 / .83	78.9

Note. CA = classification accuracy. See the note under Table 1 for abbreviation definitions for predictors.

Table 3. *Results of the ROC Curve Analyses with Single and Combined Measures*

Predictor	Cut Score	AUC [CI]	SE	SP [CI]	PPP [CI]	NPP [CI]	CA
GMRT	28th %ile	.806 [.737, .875]	~ .90	.53 [.41, .64]	.72 [.67, .77]	.80 [.68, .88]	74.4
TOSREC	SS = 100	.722 [.646, .799]	~ .90	.36 [.26, .48]	.65 [.60, .69]	.73 [.58, .84]	66.5
ORF	116 cwpm	.756 [.685, .828]	~ .90	.31 [.21, .43]	.64 [.60, .67]	.70 [.54, .82]	65.1
SWE	SS = 102	.702 [.623, .781]	~ .90	.26 [.16, .37]	.61 [.57, .65]	.58 [.42, .72]	60.5
PDE	SS = 103	.642 [.559, .724]	~ .90	.15 [.08, .25]	.59 [.57, .62]	.55 [.35, .74]	58.8
Combined		.826 [.762, .891]	~ .90	.59 [.47, .71]	.74 [.69, .79]	.83 [.72, .90]	77.1

Predictor	Cut Score	AUC [CI]	SE	SP [CI]	PPP [CI]	NPP [CI]	CA
GMRT	31st %ile	.846 [.776, .916]	~ .90	.55 [.43, .67]	.60 [.54, .66]	.88 p.76, .94]	69.9
TOSREC	SS = 92	.787 [.707, .866]	~ .90	.46 [.35, .58]	.53 [.48, .59]	.88 [.75, .94]	64.1
ORF	103 cwpm	.809 [.733, .885]	~ .90	.54 [.42, .65]	.59 [.52, .65]	.87 [.75, .94]	68.9
SWE	SS = 102	.749 [.663, .835]	~ .90	.34 [.24, .46]	.49 [.45, .54]	.79 [.63, .89]	56.8
PDE	SS = 100	.736 [.649, .824]	~ .90	.42 [.31, .54]	.54 [.48, .59]	.86 [.73, .94]	64.8
Combined		.861 [.793, .928]	~ .90	.73 [.62, .83]	.70 [.61, .77]	.90 [.81, .95]	79.7

Note: AUC = Area under the curve; CI = 95% confidence interval; SE = Sensitivity; SP = specificity; PPP = Positive predictive power; NPP = Negative predictive power; CA= Classification accuracy; SS = Standard Score

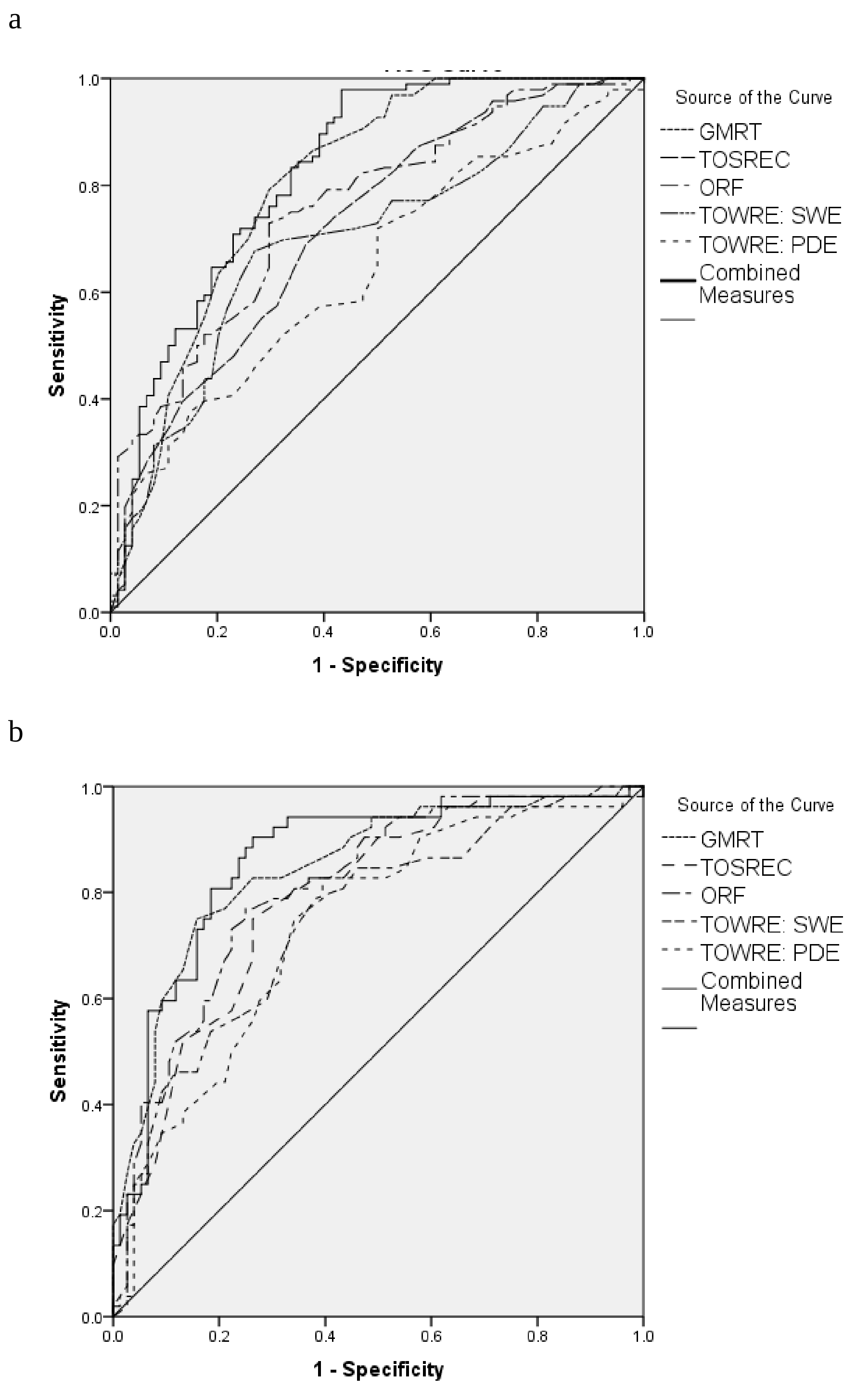

Figure 1. ROC curves for individual screening measures and combined measures for FL site (a) and TX site (b)