12-1-2011

Identifying Struggling Readers in Middle School with ORF, Maze and Prior Year Assessment Data

Jennifer R. Allison
Walden University

Evelyn S. Johnson
Boise State University

Identifying Struggling Readers in Middle School with ORF, Maze and Prior Year Assessment Data

Jennifer R. Allison
Richard W. Riley College of Education and Leadership, Walden University
155 Fifth Ave. South, Suite 100, Minneapolis, MN 55401, USA
Tel: 1-800-925-3368 E-mail: jennifer.allison@waldenu.edu

Evelyn S. Johnson (Corresponding author)
College of Education, Boise State University
1910 University Dr., MS 1725, Boise, ID, 83725, USA
Tel: 1-208-283-1820 E-mail: evelynjohnson@boisestate.edu

Received: October 16, 2011 Accepted: November 25, 2011 Published: December 30, 2011
doi:10.5539/jedp.v1n1p35 URL: http://dx.doi.org/10.5539/jedp.v1n1p35

Abstract

Response to Intervention (RTI) is a framework with the primary purpose of early identification and prevention of learning problems. Screening procedures identify students in need of targeted intervention, but current screening research is limited to the elementary grades. This study explored the use of screening measures: prior year assessment data, oral reading fluency (ORF), and maze, to predict performance on Georgia's Criterion-Referenced Competency Test (CRCT-8) for 236 eighth grade students from one district in Georgia. Logistic regression analyses compared the accuracy of the predictor variables. Overall classification accuracy was 96.6% for ORF and maze and 97.1% for CRCT-7; however, this was primarily due to the low base rate of poor performance on the CRCT-8 in the sample. A combination of screens did not significantly improve classification accuracy. A screening process that used CRCT-7 data followed by fall ORF resulted in 100% sensitivity and 90% specificity. Implications for practice are discussed.

Keywords: Screening, Reading, Middle school

1. Introduction

When a Response to Intervention (RTI) process is implemented with fidelity in the early grades, the anticipated outcome is that students who are struggling readers will be identified early and provided intervention to support their learning needs. Even with an effective RTI process in place at the elementary level however, there will continue to be students in the later grades that require intervention to become successful readers. Research on struggling readers at the middle grades suggests that although every individual student may have differences in their reading profiles, in general, struggling readers in the secondary grades will likely fall into one of the following categories:

1) Late-Emergent Reading Disabled Students. These are students who were able to keep up with early reading demands but for whom later demands become too great. Several research studies confirm that there is a category of students who will progress as typically developing students or respond positively to early intervention, only to develop reading problems in the later grades (Compton, Fuchs, Fuchs, Elleman & Gilbert, 2008; Leach, Scarborough, & Rescorla, 2003; Lipka, Lesaux, & Siegel, 2006).

2) Instructional Casualties. Although there has been a strong emphasis on improving reading instruction in the early grades, not all schools have strong reading programs in place. In the later grades, there will be students who have not been the recipients of strong reading instruction in the early grades who will require supports (Vaughn et al., 2008).

3) English Language Learners. In the last decade, the number of English language learners has increased by 57 percent (Maxwell, 2009). All schools will need to provide instruction and intervention to meet the needs of a growing ELL population.

4) Students Requiring Ongoing Intervention. Students who received intervention at the early grades may make progress, but perhaps not at a rate that is sufficient to be successful in the general education program without continued intervention. These students may require support in later grades before they are able to successfully perform at grade level benchmarks.

That there are groups of students who will continue to need intervention suggests a need for screening procedures to continue throughout middle school. A coordinated, effective screening system can identify struggling students early in the school year, allowing middle schools to further diagnose students to determine which type of struggling reader the student is, and then provide and tailor intervention resources to support continued literacy development. An effective screening system requires the identification of measures or indicators that are highly accurate in predicting the outcome of interest. However, the research base to identify measures that are predictive of poor reading outcomes at the secondary level is currently lacking.

1.1 Screening at Secondary Levels

Recommendations for practice at the secondary level include using prior year state assessment data to identify an initial risk pool of poor readers followed by subsequent assessment to determine the nature of the student's difficulties (Torgesen & Miller, 2009). But few studies have examined reading screening tools at the secondary level, and not all students will have state assessment data from the previous year, so it is likely that middle schools will require a screening measure that can be administered in the beginning of the school year to screen all current students for reading problems. Two promising candidates are oral reading fluency (ORF) and maze. For earlier grade levels, research supports the use of read aloud measures (Jenkins, Hudson & Johnson, 2007; Wayman, Wallace, Wiley, Ticha & Espin, 2007), but the validity of read aloud for older grades is not well established (Wayman et al., 2007). At the middle grades maze measures may serve as a better predictor of reading performance (Espin & Foegen, 1996; Wiley & Deno, 2005). Torgesen, Nettles, Howard, & Winterbottom (2005) examined the relationship between ORF and maze and the Florida Comprehensive Assessment Test (FCAT) with students at the fourth, sixth, eighth, and tenth grades. Results indicated that maze was more highly correlated (.67) with F-CAT performance than ORF (.59) for sixth-grade students. At the eighth-grade level, maze and ORF demonstrated similar correlations with F-CAT performance, .63 and .62 respectively. Fore, Boon and Martin (2007) also found maze (.439) and ORF (.397) performance to be moderately correlated with concurrent performance outcomes on Georgia's Criterion-Referenced Competency Test (CRCT). However, their sample was relatively small (N = 50) and consisted only of students identified with behavior disorders.

1.2 Requirements for Screening Measures

Although these studies demonstrate a moderate correlation between the predictor and outcome variables, documenting a relationship between a screening and outcome measure provides only weak evidence regarding the utility of a measure for screening purposes (Jenkins, 2003). In a preventive service delivery model such as RTI, screening results are used to efficiently determine which students are at risk for reading failure. Therefore, it is important to determine the effectiveness and efficiency of a screening measure prior to implementation in practice. A number of statistics beyond predictive or concurrent validity coefficients are used to evaluate screening measures. Classification accuracy, or how well a screen accurately sorts students into *at-risk* or *not at-risk* categories, is an important criterion for effective screens (Jenkins, Hudson, & Johnson, 2007). Classification accuracy however, must be evaluated by considering the base rate of the population with the condition being predicted (Meehl & Rosen, 1955; Wilson & Reichmuth, 1985). If the base rate of the condition being identified is very low, high classification accuracy can be achieved simply by assuming that nobody is at risk (Johnson, Jenkins, Petscher & Catts, 2009).

In addition to classification accuracy, sensitivity and specificity are two key statistics that help to describe how well a screening measure works. Sensitivity is the probability that a screening test will be positive when the student is at-risk. Because the goal in an RTI framework is to provide intervention early to those students who need it, researchers have suggested that sensitivity rates of screening processes reach 90% or higher (Compton et al, 2006; Jenkins, 2003). Specificity is the probability that a screening test will be negative when the student is in fact not at-risk. Specificity must also be high in order to avoid over-identifying students as at-risk and over-taxing school resources (Jenkins & Johnson, 2008). Sensitivity and specificity are helpful statistics to determine the general efficacy of a screen, but are difficult to interpret for use in practice.

1.3 Receiver-Operating Characteristic (ROC) Curves

Receiver-operating characteristic (ROC) curves are useful graphs that provide a complete representation of classifier performance (Krzanowski & Hand, 2009). The ROC curves are a generalization of the set of potential combinations of sensitivity and specificity possible for predictors (Pepe, Janes,Longton, Leisenring, &

Newcomb, 2004) and provide information about cut scores and associated levels of sensitivity and specificity. For example, if a desired level of sensitivity is 90%, the ROC analysis identifies the cut score that results in that level of sensitivity, and also reports the associated level of specificity. An overall indication of the diagnostic accuracy of a ROC curve is the area under the curve (AUC). The AUC values closer to 1 indicate the screening measure reliably distinguishes among students with satisfactory and unsatisfactory reading performance, whereas values at .50 indicate the predictor is no better than chance (Zhou, Obuchowski, & Obuchowski, 2002). A main advantage of ROC curve analysis over other screening statistics is the ability to identify cut scores associated with various levels of sensitivity and specificity. If the cut scores are cross-validated, ROC analysis provides practitioners with information that directly informs implementation.

Because screening is a high stakes decision (Davis, Lindo, & Compton, 2007), screening measures need to be evaluated accordingly. Screening tools that do not improve the classification and identification of at-risk students tax limited resources (e.g., time, money, personnel) and threaten the potential efficacy of the RTI process. Although ORF and maze have both demonstrated moderate correlations with general reading outcomes at the older grades, very few studies examining their classification accuracy, sensitivity, and specificity for use within an RTI framework have been conducted. Despite the research demonstrating high correlations of ORF with overall reading ability, ORF may fail to predict the reading performance of older students when more complex reading tasks such as comprehension and analysis of text are required (Wiley & Deno, 2005). In addition to the lack of research on ORF as an effective middle school screening instrument, its use at older grade levels also raises practical concerns. Although ORF consists of a one-minute read aloud, it must be administered individually (Fuchs & Fuchs, 1992). Alternatively, maze tasks can be given in a group setting (Fuchs & Fuchs, 1992). Efficiency becomes even more crucial in the upper grades when resources, including personnel and funding, often become more limited than they are at elementary level.

1.4 Research Purpose and Questions

As middle schools move forward with RTI implementation, they require information about the utility of various instruments as potential screening measures. The purpose of this study is to evaluate the classification accuracy of three potential screening tools for predicting performance on the eighth-grade reading subtest of one state's reading assessment; prior year state assessment data, ORF, and maze. The specific research questions addressed in this study are:

1) What are the classification accuracy, sensitivity, specificity and ROC AUC of each of the three predictor variables (prior year state assessment, ORF and maze) in predicting 8th grade reading performance?

2) Does a combination of predictors result in improved classification accuracy compared to a single predictor?

3) Does a screening system that uses prior year assessment data to identify an initial risk pool followed by additional screening measures as recommended by Torgesen and Miller (2009) improve classification accuracy over a single predictor?

2. Research Methods

2.1 Participants

Participants included 236 eighth-grade students from one suburban middle school (grades 6 – 8) in Georgia, where the lead author serves as the district RTI coordinator. Of the 236 students, 49% were girls. The racial composition of the sample was 88% white, 9% African American, 2% Hispanic, 1% other. Nine percent of students received special education services.

2.2 Measures

2.2.1 Oral Reading Fluency (ORF)

ORF is designed to measure an individual's accuracy and rate of reading. The probes used in this study are one component of the System to Enhance Educational Performance (STEEP, 2007), a package designed to screen and progress monitor students in reading and math. Length, content, and format for STEEP ORF are based on CBM research. The Spache and Dale Chall formulas were used to ensure appropriate level of readability. In studies conducted with STEEP ORF progress monitoring, test-retest reliability ranged from .91 to .95 (STEEP, 2007). Alternate form reliability ranged from .83 to .88 (STEEP, 2007). It should be noted that data from the STEEP technical manual were derived from progress monitoring probes for grades 1-5; as data for probes at the higher grades were not available. STEEP ORF probe administration for the purpose of screening involves the use of one probe as opposed to three. Ardoin et al. (2004) found that three probes versus one did not statistically improve the predictive validity of a screen for classifying students at risk for failure on the criterion measure.

For ORF screens, students have 1 minute to read a grade-level passage to a test administrator who then calculates rate and accuracy. The score is represented by the number of words read correctly in one minute (wcpm). Established winter ORF cut scores for each performance level at the eighth grade are as follows: frustrational: < 110 wcpm, instructional: 110 - 129 wcpm, mastery: > 129 wcpm. For this study, a team of school and district personnel were the test administrators. Directions and procedures for STEEP ORF are standardized and all testers completed a training session, which included practice in administration and scoring.

2.2.2 Sentence Maze

The sentence maze is used for screening and progress monitoring purposes within the STEEP assessment system. This sentence completion task is a variation on the traditional passage maze and is considered to be a general measure of reading performance and comprehension. Students have 3 minutes to complete the sentence maze probe, which consists of circling the correct answer from three choices to complete a sentence. The score is computed as the total number of correct answers. Established winter maze cut scores for each performance level at the eighth grade were as follows: frustrational: < 17, instructional: 17 - 24, mastery: > 24. Reliability of the middle school passages is high, with median test-retest reliability of .90; alternate form reliability of .87 and coefficient alpha (internal consistency) of .92 (STEEP, 2007).

2.2.3 Georgia Criterion-Referenced Competency Test, Reading (CRCT)

In this study, we obtained the students' 7th and 8th grade CRCT scores. The CRCT reading subtest consists of reading passages and multiple choice questions designed to evaluate student skills in the following domains: reading skills and vocabulary acquisition, literary comprehension, and information and media literacy. The reliability coefficient, Cronbach's alpha, for the 2008 eighth-grade reading subtest is .87. Reliability for the 2007 seventh-grade reading subtest was .88. Georgia's performance levels on the CRCT are defined as "does not meet expectations", (< 800), "meets expectations" (800 - 849), and "exceeds expectations" (> 849). Conditional standard errors of measurement for each of the state's performance level cut scores are as follows: "meets expectations": 7; "exceeds expectations": 10. Procedures to establish validity, defined as the test's ability to assess the intended measure (e.g., the state of Georgia's curriculum standards), are implemented throughout the entire test construction. The test development process consisted of several phases. Educators from the state are involved in each step beginning with a determination of the standards to be measured on the test. A test blueprint and test specifications are produced in the next phase. Drafting of actual test items occurs next followed by field testing. Data from field testing are examined for evidence of bias.

2.3 Procedures

Students completed the ORF and maze screening measures in the winter of the 2007/2008 school year. Probes were administered by trained professionals according to standardized procedures as described above. ORF probes were administered individually and maze was given in a group setting by class. Students complete the CRCT in the spring of the school year. The CRCT was administered according to standardized procedures and was scored electronically at the state level. Results are provided to each school through a state database. The CRCT results for student performance in 7th grade (2006-07 school year) and 8th grade (2007-08 school year) were obtained from existing school records.

2.4 Data Analysis

Predictor variables included winter ORF and maze raw scores and seventh grade CRCT (CRCT-7) scale scores. The outcome variable was the eighth-grade reading CRCT (CRCT-8) subtest. Scale scores on the 8th grade CRCT were recoded into a dichotomous pass/fail variable according to state performance standards. Pearson's correlations were conducted to determine the relationship between the predictor and outcome variables. Next, we conducted a logistic regression analysis using each of the screening measures as single predictors of a dichotomized CRCT score. This provided an overall classification accuracy of the predictors when accuracy was maximized through statistical analyses. Next, we conducted ROC curve analyses on each of the predictors to obtain the AUC and to identify cut scores on the predictors associated with 90% sensitivity levels. Then, using the cut scores associated with 90% sensitivity in the ROC analysis, we identified the number of students correctly identified as at-risk or not at risk and computed a resultant classification accuracy level.

Because screening research conducted at the elementary level has indicated that combinations of predictors result in higher levels of classification accuracy (Compton et al., 2006; Johnson, Jenkins & Petscher, 2009), we investigated whether the combination of predictors would result in greater accuracy than single predictors. We entered all of the predictors into a single logistic regression analysis to determine if the combination of measures would result in improved classification accuracy. Finally, because best practice recommendations for screening at the secondary levels suggests that prior state assessment data be considered (Torgesen & Miller, 2009), we repeated the logistic regression analyses using the students' CRCT score from their 7th grade year. We then examined whether a screening system that used CRCT-7 to identify an initial risk pool followed by confirmation of risk status with screening measures (ORF and maze) in the following school year (8th grade) would result in high classification accuracy levels.

3. Results

Table 1 presents the means, standard deviations (SD) and correlations of variables. Medium correlations (Cohen, 1988) were demonstrated between ORF and maze raw scores and CRCT-8 standard score (r = .501 and .507 respectively). Results indicated small but significant correlations between ORF and maze raw scores and the dichotomized pass/fail CRCT variable (r = .278 and .231 respectively). The CRCT-7 had the highest correlation with CRCT-8 (r = .767).

The logistic regression analyses of each of the single predictors yielded high levels of overall classification accuracy (96.6% for ORF and maze; 97.1% for CRCT). However, classification accuracy is highly dependent upon the base rate of a condition in a population (Meehl & Rosen, 1955). As seen in Table 2, neither ORF nor maze identified any students at-risk (0% sensitivity), and CRCT-7 had an associated sensitivity level of only 29%. As seen in Table 3, the base rate of students with poor CRCT-8 outcomes in this sample was very low; only 8 of the 236 participants (3.4%) failed to meet expectations. The classification accuracies of ORF and maze (96.6%) are equal to the percentage of students who passed the CRCT-8. In other words, high classification accuracy was obtained simply by assuming that no students were at-risk. To be diagnostically efficient, a screening measure must increase prediction above and beyond what could be predicted by base rates alone (Meehl & Rosen, 1955). Therefore, in this case, neither ORF nor maze was an efficient predictor for state test performance. CRCT-7 faired only slightly better, resulting in 97.1% classification accuracy and 29% sensitivity. As anticipated by the results of the single predictors, when a combined model with all three predictors was conducted, CRCT-7 was the only variable that was significant ($p < .01$), classification accuracy of the combined model remained at 97.1%.

Logistic regression analyses report classification results based on optimal statistical analysis. Because the goal in an RTI framework is to identify and intervene for students at-risk for poor learning outcomes, the next question explored whether an adjustment in cut scores would be appropriate for achieving sensitivity levels closer to the recommended 90% (Jenkins, 2003). Recognizing that no screening process is perfect, in practice, erring on the side of the over identification of students (e.g., increasing the number of false positives) is preferred to not identifying students who are truly at risk (e.g., false negatives). At the same time, because of limited resources to provide intervention, and because interventions are more effective with small teacher to student ratios (Johnson et al, 2009), the over-identification of students (false positives) must be kept to manageable levels. ROC curve analyses were conducted to evaluate sensitivity, specificity and related classification accuracy levels when sensitivity levels were set as close to 90% as possible, when test-developer recommended cut scores were used, and when the cut scores associated with the logistic regression analyses were used (see Table 3).

Although CRCT-7 has the highest AUC of all predictors (.962), when setting sensitivity levels at 90%, ORF was the only predictor that resulted in high sensitivity and high specificity (88% and 82% respectively), missing only one student who was not successful on CRCT-8, and over-identifying 40 students (17% of the sample) as potentially at risk. Similarly, maze resulted in a high level of sensitivity (88%) and moderate specificity (72%), with 64 students (27% of the sample) overidentified.

Because the logistic regression analysis using a combined model did not result in higher levels of classification accuracy than was achieved by CRCT-7 alone, we did not run a ROC analysis on a combined model. However, following the guidance of Torgesen and Miller (2009), we used prior year state assessment data to identify an initial risk pool, then confirmed results with the following year's screening results. Figure 1 presents a flowchart of the process. First, we identified the students who were not successful on CRCT-7, and then examined their ORF and maze scores to determine if their performance on those screens fell under the cut scores (those related to obtaining 90% sensitivity). For this analysis, we did not use the exact cut point of 800 on CRCT-7, but added the SEM and used the cut score of 807 (Georgia SDE, 2008). Our rationale for adjusting cut scores was to determine a practical way for schools to make decisions that account for measurement error and for students who fall close to the cut score. Additionally, our ROC analysis demonstrated 100% sensitivity for this cut score on CRCT-7. For the entire sample, CRCT-7 scores were available on 208 of 236 students. Thirty-six students had CRCT-7 scores that fell below 807. One of the 8 students in the entire sample (n = 236) who did not pass the CRCT-8 did not have CRCT-7 data. The remaining 7 students who did not pass CRCT-8 in the reduced sample (n = 208) did not pass CRCT-7, according to the adjusted score of 807. The resulting sensitivity of CRCT-7 (based on n = 208) is 100%, with an associated specificity level of 82%.

We next looked at the performance of the 36 students with CRCT-7 scores below 807 on ORF. Twenty of 36 students also scored below 151 on ORF, and 29 of 36 had scores below 18 on Maze. We ran a crosstabs to determine the number of students who did not pass CRCT-7 (N = 36) who also had ORF scores below 151 and their outcomes on CRCT-8, then repeated this analysis for maze, using the cut score of 18. Finally, we also examined the ORF and maze performance of the 28 students from the original sample who did not have CRCT-7 data. Of the 28, 10 had ORF scores below 151, and 9 had maze scores below 18 (the cut scores identified by our ROC analysis). Adding the 28 students for whom CRCT-7 data was not available to the 36 students who did not successfully pass CRCT-7, the initial risk pool was a total of 64. Following the process of identifying students at risk based on CRCT-7 scores, followed by poor performance on ORF, and then using poor performance on ORF for students for whom CRCT-7 data was not available, 30 of the 64 students were identified as at risk and all 8 students who failed CRCT-8 were identified as at-risk. This would result in 100% sensitivity and 90% specificity. Repeating this analysis with maze in place of ORF, 38 of 64 students would have been identified as at risk and all 8 students who failed CRCT-8 would have been identified. This would result in 100% sensitivity and 87% specificity.

4. Discussion

The purpose of this study was to compare the classification accuracy, sensitivity, specificity and ROC AUC of three predictor variables (ORF, maze and prior year state assessment data) in predicting 8th grade reading performance, both as single predictors and as a combined prediction model. Additionally, this study examined the results of employing a system consistent with best practice recommendations for identifying struggling adolescent readers (Torgesen & Miller, 2009). These recommendations include using prior state assessment data to identify an initial risk pool, then confirming results with additional screening measures in the beginning of the next school year.

4.1 Accuracy of Single Predictors

In evaluating screens, the base rate of the predicted condition (e.g., poor reading outcomes) is an important consideration in judging a screen's overall classification accuracy (Wilson & Reichmuth, 1985). Specifically, focusing on a screen's overall classification accuracy can be misleading when the base rate of the predicted condition occurs in a relatively small percentage of the overall population. In such situations, a relatively high overall accuracy can be obtained by skipping screening altogether and assuming that no students are at risk. It has been argued that to be efficient, a screening measure must result in a higher number of correct classification decisions than could be obtained in terms of base rate alone (Meehl & Rosen, 1955). However, in a preventive service model, one could also argue that if the condition being predicted is of significant concern but occurs in a relatively small percentage of the population, the costs associated with screening to detect those at risk may be worthwhile.

In this study, the base rate of students with poor reading outcomes was low (3.4%). When using a logistic regression analysis that relies on the use of statistically optimal cut points, no single predictor resulted in classification accuracy rates better than the 96.6% that is achieved simply by assuming no student is at risk. This sample is not very different from the overall base rate in Georgia, which reported a 94% pass rate for 8th grade reading during the 2007-08 school year (Georgia Department of Education, 2008). However, in an RTI model, screening is a purposeful activity – its first job is to identify all or nearly all students at risk for poor future outcomes (Johnson et al., 2009). This led us to select cut points that resulted in 90% sensitivity, allowing us to compare the relative accuracy of screening measures according to their level of over-prediction, or specificity. When sensitivity levels were set at 90% for ORF and maze, both resulted in over-identification rates that would still be unmanageable to serve within a tiered intervention system. ORF over-identified 40 students (about 20% of the sample), and maze 64 (about 25% of the sample). Providing high-quality, effective, intensive services for this many students is not possible for schools. A combination of predictors entered simultaneously did not improve the classification accuracy.

4.2 Using Existing Data to Identify an Initial Risk Pool

At the secondary level, schools will have available data on many of their students from prior assessments of reading performance. If extant data is accurate in identifying students who are likely to be at risk, then the costs associated with screening could be greatly reduced. In this study, 88% of the original sample had performance data from their seventh grade year (CRCT-7). In analyzing CRCT-7 as a predictor variable, the best prediction was obtained when we added the SEM to the cut score that separates "meets expectations" from "does not meet expectations". Twenty-nine students were over-identified using this procedure. When ORF or maze was used to further screen these students, the rate of over-identification was reduced to 14 for ORF and 23 for maze.

4.3 Developing a Screening System

Although prior year assessment data worked well as a screener, there were 28 students (12%) who did not have CRCT-7 data available. When we used ORF and maze to screen these 28 students, 10 of them had ORF scores below the cut point, and 9 students had maze scores below the cut point. Adding those numbers to the students already identified by unsatisfactory CRCT-7, 24 total students were confirmed as at-risk by the all ORF measure; a total of 32 were confirmed as at-risk using maze. Both groups contained all 8 of the students who did not successfully pass the CRCT-8, resulting in 100% sensitivity for the gated measures and specificity levels of 90% (CRCT-7 + ORF), and 87% (CRCT-7 + maze). These numbers meet the requirements for successful screening systems within an RTI framework and maintain efficiency by requiring additional screening data on a much smaller pool of students.

4.4 Limitations

While this study presents important findings for informing the practice of screening at middle school levels, we note several limitations that restrict the generalizability of our study. First, the base rate of poor reading outcomes in this sample was extremely low, which limits the ability of any screening tool to identify students as at-risk without over-identifying a large percentage of students. Schools that serve a student population with higher percentages of students at-risk for poor learning outcomes will most likely need to develop their own decision rules and cut points. Second, limiting the sample to one district and one state test limits the generalizability of our findings outside of Georgia. Finally, creating decision rules *post hoc* allows researchers to determine accurate cut scores for a particular sample. The decision rules developed on this sample may not hold up during a cross validation study. Our next steps for continued research include testing the decision rules developed in this study on the next year's 8th graders to determine if similar results can be obtained.

5. Conclusion

This study provides supporting evidence for the best practice recommendations offered by Torgesen and Miller (2009) to identify struggling adolescent readers. Using extant data, middle schools may be able to identify an initial risk pool, then confirm preliminary screening data with subsequent screening tools the following school year. This process resulted in high sensitivity and specificity without requiring a high level of additional resources. In this sample, ORF was found to be a better predictor than maze. These measures require further investigation as well as cross validation to continue to inform the most efficient and effective screening process for identifying adolescent students who are at risk for poor reading outcomes.

References

Ardoin, S. P., Witt, J. C., Suldo, S. M., Connell, J.E., Koenig, J. L., & Resetar, J. L., et al. (2004). Examining the incremental benefits of administering a maze and three versus on curriculum-based measurement reading probes when conducting universal screening. *School Psychology Review,* 33(2), 218-233.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

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. http://dx.doi.org/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. http://dx.doi.org/10.1016/j.lindif.2008.04.003

Davis, G. N., Lindo, E. J., & Compton, D. (2007). Children at-risk for reading failure: Constructing an early screening measure. *Teaching Exceptional Children,* 39 (5), 32-39.

Espin, C. A., & Foegen, A. (1996). Validity of general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children,* 62, 6, 497-514.

Fore, C., Boon, R. T., & Martin, C. (2007). Concurrent and predictive criterion-related validity of curriculum-based measurement for students with emotional and/or behavioral disorders. *International Journal of Special Education,* 22 (2), 24 – 31.

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School*

Psychology Review, 21(1), 45-58.

Jenkins, J. R. (2003, December). Candidate measures for screening at-risk students. Paper presented at the NRCLD responsiveness-to-intervention symposium, Kansas City, MO. [Online] Available: http://www.nrcld.org/symposium2003/jenkins/index.html (April 3, 2008)

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for service delivery in an RTI framework: Candidate measures. *School Psychology Review,* 36, 582-99.

Jenkins, J. R., & Johnson, E. S. (2008). *Universal screening for reading problems: Why and how should we do this?* RTI Action Network. [Online] Available: http://www.rtinetwork.org/Essential/Assessment/Universal/ar/ReadingProblems (April 16, 2008)

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research and Practice,* 24(4), 174-194. http://dx.doi.org/10.1111/j.1540-5826.2009.00291.x

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

Lipka, O., Lesaux, N. K., & Siegel, L. S. (2006). Retrospective analyses of the reading development of grade 4 students with reading disabilities: Risk status and profiles over 5 years. *Journal of Learning Disabilities,* 39, 364−378. http://dx.doi.org/10.1177/00222194060390040901

Maxwell, L. (2009, January 8). Immigration transforms communities. *Education Week*. [Online] Available: http://www.edweek.org (April 25, 2009)

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin,* 52(3), 194-216. http://dx.doi.org/10.1037/h0048070

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. http://dx.doi.org/10.1093/aje/kwh101

Torgesen, J. K., & Miller, D. H. (2009). *Assessments to guide adolescent literacy instruction*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Torgesen, J., Nettles, S., Howard, P., & Winterbottom, R. (2005). *Brief Report of a Study to investigate the relationship between several brief measures of reading fluency and performance on the Florida Comprehensive Assessment Test-Reading in 4th, 6th, 8th, and 10th grades. (Technical Report #6)*. Tallahassee, FL: Florida Center for Reading Research.

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

Wayman, M. M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measures in reading. *The Journal of Special Education,* 2(41), 85-120. http://dx.doi.org/10.1177/00224669070410020401

Wiley, H.I., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education,* 26(4), 207-214. http://dx.doi.org/10.1177/07419325050260040301

Wilson, B. J., & Reichmuth, M. (1985). Early-screening programs: When is predictive accuracy sufficient? *Learning Disability Quarterly,* 8 (3), 182-188. http://dx.doi.org/10.2307/1510892

Zhou, X. H., Obuchowski, N. A., & Obuchowski, D. M. (2002). *Statistical methods in diagnostic medicine.* Wiley & Sons: New York. http://dx.doi.org/10.1002/9780470317082

Table 1. Descriptive statistics and correlations

Measure	CRCT - 7	ORF	Maze	CRCT – 8	CRCT-8 (p/f)
CRCT – 7					
ORF	.514				
Maze	.591	.725			
CRCT – 8 (SS)	.767	.501	.507		
CRCT-8 (p/f)	.326	.278	.231	.442	
Mean (SD)	830.28 (23.19)	176.87 (31.89)	21.41 (5.90)	833.35 (19.14)	

Note. All correlations were significant, p < .01 (2-tailed).

Table 2. ORF, Maze and 7th CRCT logistic regression summary

Predictor	B	SE	Wald	Classification Accuracy	Sensitivity	Specificity
ORF	.060	.017	12.586	96.6	0	100
Maze	.326	.099	10.837	96.6	0	100
CRCT-7	.164	.052	9.867	97.1	29	99

Table 3. Classification indices for predictors at 90% sensitivity, at published cut scores, and based on logistic regression analysis

Measure	Sen	Spec	Cut Score	ROC AUC	TP	FP	TN	FN	Classification Accuracy
ORF	88	82	151	.898	7	40	188	1	82.6
	25	99	109		2	2	226	6	96.6
	0	100	95		0	0	228	8	96.6
Maze	88	72	18	.856	7	64	164	1	72.5
	75	76	17		6	55	173	2	75.8
	0	100	7		0	0	228	8	96.6
CRCT-7	100*	89	807	.962					
	71	96	800		5	7	194	2	95.6
	28	99	780		2	1	200	5	97.0

Note. Sen = Sensitivity; Spec = Specificity; ROC AUC = ROC area under the curve; TP = true positives; FP = false positives; TN = true negatives; FN = false negatives; Row 1 presents classification measures obtained after setting sensitivity levels as close as possible to 90%. Row 2 presents actual statistics from the study sample based on published cut scores. Row 3 presents statistics for cut scores resulting in maximum overall classification accuracy.

*Sensitivity levels for CRCT-7 in the ROC analysis were 71, with the next reported level of 100.

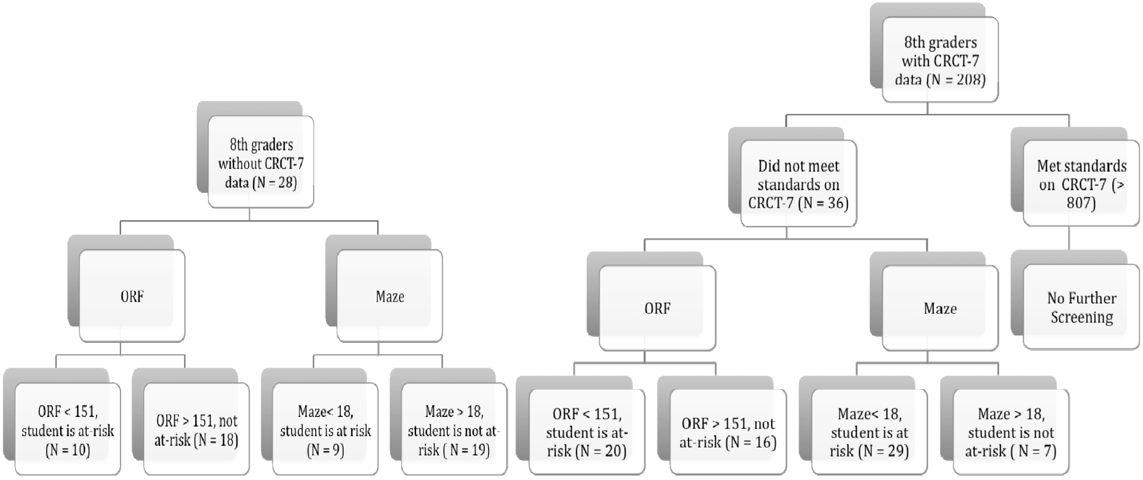
8th graders without CRCT-7 data (N = 28)
ORF
Maze
ORF < 151, student is at-risk (N = 10)
ORF > 151, not at-risk (N = 18)
Maze< 18, student is at risk (N = 9)
Maze > 18, student is not at-risk (N = 19)
8th graders with CRCT-7 data (N = 208)
Did not meet standards on CRCT-7 (N = 36)
Met standards on CRCT-7 (> 807)
ORF
Maze
No Further Screening
ORF < 151, student is at-risk (N = 20)
ORF > 151, not at-risk (N = 16)
Maze< 18, student is at risk (N = 29)
Maze > 18, student is not at-risk (N = 7)

Figure 1. Flowchart of screening process following Torgesen & Miller's (2009) best practices