UNIVERSITY OF CALIFORNIA
RIVERSIDE

# Using Error Recognition to Identify Students at Risk of Reading Failure

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Arts

in

Education

by

Guadalupe Guzman

June 2017

Thesis Committee:
Dr. Rollanda O'Connor, Chairperson
Dr. Austin Johnson
Dr. Cathleen A. Geraghty

The Thesis of Guadalupe Guzman is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ABSTRACT OF THE THESIS

Using Error Recognition to Identify Students at Risk of Reading Failure

by

Guadalupe Guzman

Master of Arts, Graduate Program in Education
University of California, Riverside, June 2017
Dr. Rollanda O'Conner, Chairperson

The current trend in education points toward preventative methods in order to address the problem of reading failure. An essential component of a preventative method is early identification of students at risk. Effective screeners are those with high classification accuracy, efficiency, and usability. Though current screeners such as the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) have sufficient classification accuracy for their intended purpose, there is room for improvement in the accuracy and usability of these measures. The forthcoming study will identify differences in the accuracy of an experimental error recognition measure, DIBELS Oral Reading Fluency (DORF) measure, and DIBELS Daze measure for identifying 3rd grade students at risk for reading failure. Error recognition (ER) in this context is the identification of reading errors made by another individual. A total of 57 3rd grade students, 22 males, and 35 females, ranging in age from 7:11 to 9:1 participated in the study. Participants were predominantly of Hispanic/ Latino descent with 6 African-American and 2 Asian- American student participants. Of the 57 participating students 23 were English Learners with California English Language Development Test levels ranging from 1 – 5. At risk criteria for the DIBELS measures were based on appropriate benchmark goals while at risk criteria for

the ER task was defined as a score of less than 13 on the ER measure. Results indicated that the DORF measure had a slightly larger area under the curve (AUC) value followed by the ER measure and the Daze measure respectively. However, AUC value comparisons indicated there were no statistically significant differences between the AUC values obtained by any of the three measures. Limitations and future directions are discussed.

*Keywords*: Error recognition, early identification, reading, English Learners

Table of Contents

Using Error Recognition to Identify Students at Risk of Reading Failure

Early literacy skills are essential for students' success throughout their education (National Reading Panel, 2000). However, in 2015 only 35% of fourth-grade students and 34% of eighth-grade students performed at or above proficiency level in reading, while 31% of fourth-grade students and 24% of eighth-grade students performed below the basic level in reading (National Assessment of Educational Progress (NAEP), 2015). English Learners (ELs) differ from English-only students in that they must learn academic content while also acquiring knowledge of the English language. Research indicates that ELs currently represent 9.2% (i.e., approximately 4.4 million students) of the students enrolled in public schools in the United States (National Center for Educational Statistics, 2015). A large portion of ELs struggle academically with only 7% of fourth-grade and 4% of eight-grade ELs scoring at or above proficiency level on reading assessments.

The aforementioned statistics are alarming given that research has consistently indicated that children who get off to a poor start in reading rarely catch up. Torgesen and colleges (1988) indicated that students who do not learn how to read in the early grades have persistent reading difficulties. Moreover, the negative consequences of being a struggling reader continue to accumulate over time. Consequences can range from negative attitudes toward reading (Oka & Paris, 1986), to fewer opportunities for building vocabulary comprehension strategies (Brown, Palinscar, & Purcell, 1986), and/or to less actual practice in reading (Allignton, 1984). Learning to read in the early grades then, is critical to students' success.

Current trends intended to alleviate the problem of reading failure are embedded in early identification of students at risk. Early screening tools have allowed educators to identify students at risk of reading failure with a brief assessment. Screening measures have been developed to assess different reading skills based on students' grade levels. Early elementary screening measures assess letter-name knowledge, phonemic awareness, and letter-sound knowledge while later elementary screening tools measure fluency and reading comprehension. Screeners are typically administered school-wide and are primarily administered individually. Screening measures are intended to be brief (i.e., approximately 5 -10 minutes per student) and with adequate training can be administered by a variety of school personnel. Universal screening data provides the best results when administered periodically throughout the year and used to identify students who are at risk of not developing the necessary reading skills for their grade level (Ardoin & Christ, 2008).

Early identification of reading failure is essential given that the majority of students with a learning disability are primarily impaired in reading (Fletcher, Lyon, Fuchs, & Barnes, 2007). Early identification of reading failure allows for the identification of students in need and consequently students not in need of additional supports to develop the necessary skills and knowledge to succeed in reading. Thus, identifying students who are at risk allows for appropriate allocation of reading intervention resources. In addition, the literature base surrounding intervention studies indicates that a large majority of students can gain the necessary reading skills needed to prevent long-term reading failure when provided with early and intensive reading

interventions (Wanzek & Vaughn, 2007). Early reading interventions have been shown to reduce the number of students at risk for reading problems and those who qualify for special education (Mathes & Denton, 2002; O'Connor et al. 2013) and have demonstrated effectiveness for both English Only (EO) students and English Learners (ELs) (Linan-Thompson, Vaughn, Hickman-Davis & Kouzekanani, 2003; Vaughn et al., 2003; Wanzek & Vaughn, 2007). As such, early identification provides educators a means of allocating resources appropriately and a means of assisting students with intensive intervention in order to prevent long-term reading failure and consequently minimize the negative effects experienced by struggling readers.

**Response to Intervention**

Currently, school-wide screening assessments are used to identify students at risk of reading failure. Universal screening is an integral part of the Response to Intervention (RTI) framework, which is associated with positive student outcomes. Torgesen (2009) found that implementing an RTI model is associated with a decrease in special education referrals and an increase in overall academic achievement. In an RTI framework student learning is improved using a multi-tiered (i.e., Tier 1, Tier 2, and Tier 3) instructional and service delivery model (Jenkins & Johnson, 2008). Tier 1 instruction is provided to all students which includes high-quality instruction and periodic screening to identify students who are at risk. In Tier 2, students who are not making sufficient academic progress are provided intensive interventions matched to their individual needs and periodically monitored to inform the effectiveness of the intervention. In Tier 3, students are provided with increasingly intensive interventions that target specific skill deficits.

RTI is designed to provide a means of intervening early with students at risk for academic difficulty, allocating resources based on students' needs, and differentiating between students whose academic difficulties are attributed to instructional and experiential deficits as opposed to a learning disability (Jenkins & Johnson, 2008).

RTI uses a prevention framework to provide students with the necessary skills to succeed academically. Screening is a critical part of the prevention framework. In a district where RTI is fully implemented students are screened for reading, writing, math, and behavior/ mental health problems in order to identify students who are at risk in one or more of these areas. Students who are identified as at risk are then provided with Tier 2 or Tier 3 instructional and/ or behavioral support. To fulfill the conditions of an RTI framework, certain properties are desired in screening tools. Glover and Albers (2007) suggest screeners should have high technical adequacy, efficiency, and be appropriate for their intended use.

**Classification Adequacy.** The idea of classification adequacy was introduced in a study by O'Connor and Jenkins (1999), where they described the types of calculation errors that would impact classification error. The authors described under prediction and over prediction as errors that would reduce the accuracy of predicting a reading disability in Kindergarten and 1st grade students. Jenkins and Johnson (2008) defined classification adequacy of a screener as the accuracy with which the measure distinguishes between students who need intervention services and those who do not, meaning that the classification accuracy (hit rate) of a screener is the proportion of students correctly identified by the screener. Jenkins and Johnson indicate that a screener should correctly

identify at least 80% of students who are at risk. A screener can identify students by correctly identifying students at risk (i.e., true positives) and by correctly identifying students not at risk (i.e., true negatives). A screener can be incorrect by incorrectly identifying students at risk (i.e., false positives) and by incorrectly identifying students not at risk (i.e., false negatives). A screener's accuracy is identified through the sensitivity and specificity values. Sensitivity refers to the proportion of students correctly identified as at risk while specificity refers to the proportion of students correctly identified as not at risk.

Currently, no screener exists that has 100% classification accuracy; as such there is a trade-off in education between these values. The sensitivity of a screener can be easily manipulated by changing the values used to determine at risk status; however, as the sensitivity of a screener increases the specificity of the screener decreases (i.e., the screener will over-identify students at risk). A screener with low specificity will result in under-identification of students at risk. Thus, the classification adequacy of a screener can have direct consequences on the parties involved. Low specificity values will overlook students who require Tier 2 or Tier 3 services to succeed academically, while high sensitivity values will require that resources be provided to a larger portion of the school population. High sensitivity values may result in ineffective interventions due to the large number of students included in Tier 2 and Tier 3 services. Thus, high classification adequacy of screeners is necessary in order to escape these negative consequences.

Johnson, Jenkins, Petscher, and Catts (2009) examined the accuracy of the

Dynamic Indicator of Basic Literacy Skills (DIBELS) measures in predicting which students were at risk for reading failure in first grade. Results indicated that DIBELS Oral Reading Fluency (DORF) measures had higher classification accuracy than other DIBELS measures; however, the classification accuracy was higher when assuming that no students were at risk. In addition, different cut scores were needed for ELs in order to obtain comparable classification accuracy when using DORF measures. Similarly, Goffreda, Diperna, and Pedersen (2009) investigated the predictive validity of 1st grade DIBELS measures and identified DORF scores as the only first grade probes that significantly predicted reading proficiency in later grades. Parker et al. (2015) compared the diagnostic accuracy of DORF measures and an informal reading inventory for identifying students considered at risk for failing a district-wide assessment for second and third grade students. Findings indicated that DORF measures demonstrated higher diagnostic accuracy for correctly identifying at-risk students and resulted in 80% correct classification of students at risk. Yeo (2009) conducted a meta-analytic review of the research on the predictive validity of Curriculum Bases Measures (CBMs) on state achievement tests in grades one through eight. Findings indicated a .69 correlation coefficient between CBM probes and state achievement tests. However, moderator analyses indicated that lower correlations between CBM probes and statewide reading tests were found when larger proportions of the study sample size had a disability or were identified as ELs. As such, research demonstrates both a need for high classification accuracy and the need to identify differential classification accuracy of reading screening measures for ELs.

**Reading Screeners**

Currently in schools, particularly in elementary schools, the DIBELS probes are widely used as screening and progress monitoring tools in the RTI framework. The DIBELS measures are used for assessing students' acquisition of early literacy skills for grades K – 6. There are six DIBELS measures (i.e., First Sound Fluency, Letter Naming Fluency, Phoneme Segmentation Fluency, Nonsense Word Fluency, *DIBELS* Oral Reading Fluency, and Daze) that serve as indicators of phonemic awareness, alphabetic principle, accuracy and fluency with connected text, and reading comprehension (Good et al., 2013).

DORF screeners measure students' advanced phonics and word attack skills, accurate and fluent reading of connected text, and reading comprehension (Good et al., 2013). The DORF measure is administered individually and is recommended for Fall, Winter and Spring screening for grades 1 – 6. Approximate administration time per student is 6 minutes including directions and retell procedures. The Daze screening measure is administered in a group format in the Fall, Winter, and Spring for grades 3 - 6. Approximate administration time per student is 5 minutes for the group administration and 1 – 2 minutes per student for the scoring procedure. Good et al. (2013) argue that the Daze screener is a measure of student reading comprehension; however, the Maze measure (a measure which for all intent and purposes is equivalent to the Daze measure) has also been viewed as a group administered fluency test (Deno et al., 2009).

Although the DIBELS measures have demonstrated a degree of accuracy for screening purposes (Goffreda, Diperna & Pedersen, 2009; Schilling et al., 2007), the

measures are far from perfect. For instance, studies indicated a high degree of false positives particularly for the DORF measure (Schilling et al., 2007). Increasing the classification accuracy, efficiency, and usability of screening measures can minimize the negative consequences that stem from over- and under-identification of students at risk. Currently, for a third-grade teacher with 25 students, conducting the DORF and Daze screening procedures would take approximately 3 hours and 40 minutes to complete. The time to screen students can be diminished with the inclusion of more staff members participating in the screening task. However, in many real-life circumstances, staff members are spread thin in their responsibilities and the task falls primarily on teachers. With a fully implemented RTI framework screening takes place three times a year and can take a substantial amount of time. As such, this is one area where reading screening measures can be improved. In addition, improving the efficiency of reading screening measures may diminish push back to screening from teachers.

**Error Recognition**

Research indicates that effective readers are able to monitor their own reading comprehension and apply corrective procedures when comprehension falters, while less skilled readers are not (Brown, 1980; Brown 1982). Garner and Taylor (1982) conducted a study in which they presented 4$^{th}$ and 6$^{th}$ grade students with passages that contained inconsistencies in the text. Their findings revealed that students who scored lower on reading comprehension measures were not able to identify the inconsistencies presented in the passages even after the inconsistencies were pointed out to them. Research also indicates that student reading fluency scores are reliably predictive of student reading

achievement (Moon, 2016; Schilling et al., 2007), indicating that effective readers are more likely to be able to read text fluently. In addition, mastering fluency skills leads to the development of proficient reading comprehension skills (Jefferson, Grant & Sander, 2016). That is, fluent readers decode accurately and quickly, theoretically allowing for more cognitive space which can presumably be devoted to the comprehension of text (Kuhn, 2004). As such, it is reasonable to assume that students who have a higher degree of reading achievement, fluency, and comprehension will be more successful at identifying word reading errors made by another individual. Identifying words read incorrectly by another individual requires that students be able to identify words in print and interpret auditory information accurately and fluently, while also having sufficient cognitive space to identify the words read incorrectly. Essentially, while the DORF measure is a single component measure of reading fluency, the error recognition measure, like the Daze measure, is a multi-component measure. The error recognition measure combines various aspects of reading (i.e., phonemic awareness, phonological decoding, and phonics) while the Daze measure combines fluency and attention to meaning. Students who have low phonemic awareness, phonological decoding, or phonics skills may be at a disadvantage in identifying words read incorrectly, thereby allowing for differentiation among at risk and not at risk readers. However, this assumption has not yet been confirmed through empirical research.

The current study introduces an experimental error recognition measure in which students are instructed to follow along on their own copy of the reading material while marking the reading errors made in an audio recording of the principal investigator

reading the passage. The experimental measure is designed to help identify students at risk for reading failure more efficiently by diminishing the time required for screening substantially given the group based administration format. Using error recognition tasks to identify students at risk for reading failure may be one way of improving classification accuracy and efficiency of reading screeners.

**Research Questions**

1. What is the accuracy rate of using error recognition to identify students at risk (i.e., standard score ≤ 89 on WIAT-III) for reading failure?
2. Is the accuracy rate of error recognition improved over DORF and Daze screeners?
3. Does the accuracy rate of error recognition differ for students classified as ELs?

## Method

**Participants and Setting**

Upon institutional review board approval from the university and the participating school district, including the school's principle and all four 3rd grade teachers at the school site, 3rd grade students were recruited for participation. Parent consent to participate in the study was obtained for 57 of 97 students, 58.8%. Upon parental consent student assent was obtained for all 57 students. A total of 22 male and 35 female students participated in this study, 49 of which were Latino/a, 6 African American and 2 Asian American. Student ages ranged from 7:11 to 9:1 with a median age of 8:7. Of the 57 students, 23 students were identified as ELs and had California English Language Development Test (CELDT) levels ranging from 1 to 5. Individual measures were administered only to those students for whom parent consent was provided. However, the

DORF and Daze measures were administered to all students as these measures were routinely used as reading screeners by all four 3rd grade teachers. Study sessions took place in the participating student's general education classroom's during two 30-min blocks of time. As such, the only difference between participants with and without consent was the administration of the individual assessment.

**Measures**

**Wechsler Individual Achievement Test-III.** The Wechsler Individual Achievement Test- Third Edition (WIAT-III) is a frequently used standardized test of academic achievement in the areas of Reading, Written Language, Mathematics, and Oral Language. The WIAT-III is an individually administered, comprehensive instrument for assessing achievement for Pre-K - 12 students. A criterion measure that can reliably identify students with poor reading skills is necessary to perform accuracy analyses for the measures used in this study. The focus of this study was on the total reading composite; a combination of the Word Reading, Pseudoword Decoding, Reading Comprehension, and Oral Reading Fluency subtests. The WIAT-III total reading composite offers a means of assessing students' academic skills in the area of reading and is an instrument with demonstrated reliability and validity.

Reliability refers to the consistency of scores across replications of a testing procedure (AERA, APA & NCME, 1999). Salvia, Ysseldyke, and Bolt (2010) recommend standards of reliability between .6 - .9 depending on the purpose of the assessment. For test scores used to inform on groups of individuals a reliability of .6 is appropriate, for progress monitoring purposes a reliability of .7 is appropriate, for

screening purposes a reliability of .8 is appropriate and for important individual decision making a reliability of .9 should be the standard. All WIAT-III composite scores had split-half reliability scores ranging from .91 to .98, indicating the measure has excellent reliability (Vaughan-Jensen, Adame, McLean, & Gámez,
, 2009). In addition, reliability estimates for word reading, pseudoword decoding, reading comprehension and oral reading fluency subtests ranged from .90 to .94. Validity refers to the degree to which evidence and theory support the intended use of a measure and the interpretation of the test scores (AERA, APA & NCME, 1999). Content validity for the WIAT-III was established with a comprehensive literature and expert review (Vaughan-Jensen et al., 2009). The items included in the WIAT-III were thoroughly examined in order to ensure that the final items included in the measure closely aligned with and measure the intended construct. Convergent validity was derived by reporting relationships with other measures designed to measure the same or similar constructs (Vaughan-Jensen et al., 2009). The relationship between the WIAT-II and the WIAT-III was relatively high with coefficients ranging from .69 and .73. WISC-IV and WIAT-III correlations ranged from.41 to .76.

**DORF and DAZE.** The DORF measure is a frequently used reading screener which measures students' advanced phonics and word attack skills, accurate and fluent reading of connected text while Daze is commonly viewed as a standardized measure of reading comprehension (Good et al., 2013). The DORF measure is administered individually while the Daze measure can be administered in large groups, small groups, or individually.

The DIBELS Next Technical Manual reports alternate-form, test-reset, and inter-rater reliability estimates for DORF measures (Good et al., 2013). Alternate-form reliability estimates for students across 1st – 6th grade on DORF measures were all above .90; above the criterion set by Salvia, Ysseldyke, and Bolt (2010) for individual decision making purposes. Alternate form reliability estimates for the Daze adjusted score are .75, .81, and .83 for 3rd, 4th, and 5th grade students respectively indicating sufficient reliability of screening decisions. DORF test-retest reliability estimates for 1st – 5th grade exceed .90 as do inter-rater reliability estimates for 2nd – 6th grade. Test-retest reliability estimates for Daze measures were not provided; however, inter-rater reliability estimates exceeded .90 for 3rd – 6thgrade.

Content validity for DORF and Daze measures was established through measure design. That is, DIBELS Next measures were designed as indicators of overall performance in specified skill domains and linked to foundational early literacy skills. Criterion validity for DORF was established by correlating DORF measures with the Group Reading Assessment and Diagnostic Evaluation (GRADE), the Standard 4th Grade Reading Passage, and other DIBELS measures. The criterion validity coefficients for DORF measures fall within the moderate to strong range with other DIBELS scores (.70), GRADE (.66), and the Standard 4th Grade Reading Passage (.96) for 3rd grade students. The predictive and concurrent validity coefficients for the Daze adjusted score with the GRADE Total Test fall in the moderate to strong range (i.e., .50 to .69) which the authors report are indicative of the Daze measures ability to accurately measure reading comprehension. However, no other support for this notion is provided.

**Error Recognition Task.** The experimental error recognition task was adapted from the DORF progress monitoring passage 16. The passage was randomly selected in order to remove researcher selection bias. The passage contained a total of 254 words and 24 intentional errors were incorporated into the passage. Of the 24 intentional errors there were 5 preposition errors, 4 verb errors, 5 noun errors, 5 pronoun errors and 5 adjective errors. Intentional errors were incorporated approximately every 10th word. Hasbrouck and Tindal (2006) investigated the national performance norms for oral reading fluency measures for grades 1- 8. Results indicated that the average ORF rate for 3rd grade students in the 90th, 75th, 50th, 25th and 10th percentile were 146, 120, 92, 62, 36, and 43 words read correct per minute respectively. The error recognition task was designed to allow students in the upper and lower ranges of reading ability to experience some success with the exercise while also providing a means of identifying students who were at risk for reading failure. As such, the passage was read at a rate of 43 words per minute allowing for a one second pause after each word and a five second pause after each period. Students were instructed to listen carefully to the audio recording and circle any words that were read incorrectly. An adjusted score compensating for guessing was calculated based on the number of correct and incorrect responses.

The same passage was used for test-retest sessions, which occurred one week after the initial error recognition task was administered. The test- retest reliability estimate obtained was .75 indicating that the experimental error recognition task demonstrates stability over time (see Table 1). Convergent validity for the experimental error recognition task was established by correlating the scores obtained on the error

recognition task with scores obtained on the WIAT-III, DORF, and Daze measures. The error recognition task is expected to correlate with these measures given that the WIAT-III is a measure of total reading and the DIBELS probes are measures associated with different reading skills while the error recognition task was designed to identify students at risk for reading failure (students with low reading skills). The validity estimates ranged from .46 for the Daze measure to .72 for the DORF measure (see Table 1). The correlation value for the experimental error recognition measure and the WIAT-III was 0.65, while the correlation for the Daze measure and the WIAT-III was 0.59 (see Table 1). These values indicate that the experimental error recognition task appears to be a better measure of reading comprehension when compared to the Daze measure.

**Data Analysis Method**

The receiver operating characteristics (ROC) curve is widely used in evaluating the performance of many educational assessments. The experimental error recognition task introduced in this paper is designed as a diagnostic tool aimed at differentiating between students who are at risk and not at risk for reading failure. Test accuracy determination is accomplished using a gold standard criterion, which allows discrimination of participants into two groups (Fischer, Bachmann, & Jaescke, 2003). Students scoring in the average range or better (i.e., standard scores of 90 or above) on the WIAT-III total reading measure were classified as not at risk for reading failure (0 = not at risk). Students scoring below the average range (i.e., standard scores of 89 or below) were classified as at risk for reading failure (1 = at risk).

Screening measure utility was assessed using the area under the curve (AUC)

values as well as classification accuracy values. A ROC curve is defined as a plot of test sensitivity (true positive rate) as the y coordinate versus its 1-specificty (false positive rate) as the x coordinate (Fawcett, 2006). Essentially, two types of ROC curves can be derived when a binomial distribution is assumed: parametric and non-parametric. Although parametric and non-parametric ROC curve methods lead to statistically insignificant differences (Hajian-Tilaki, Hanley, Joseph, & Collet, 1997), when using parametric methods normality assumptions must be met and for this reason non-parametric methods are preferred. Non-parametric ROC curves were derived using the SPSS statistical package for all analysis included in this manuscript.

ROC curves typically fall on a diagonal line, above a diagonal line, or below a diagonal line. A ROC curve that falls on a diagonal line (sensitivity = 1- specificity) represents an uninformative test, which is equal to that of randomly guessing (Fawcett, 2006). ROC curves that are above the diagonal line represent better tests with higher curves representing more informative tests. ROC curves that are below the diagonal line represent uninformative tests that are worse than guessing. Although ROC curves can be visually analyzed to determine whether the diagnostic test was informative, a comparison of ROC curves often encompasses comparing the AUC. AUC values fall between 0 and 1 with AUC values equal to 1 indicating perfect accuracy (Hanley & McNeil, 1982). Generally speaking, larger AUC values are indicative of greater accuracy rates of the diagnostic classifier. A rough rule of thumb is that tests with AUC values greater than 0.90 have high accuracy, while AUC values between 0.70 and 0.90 indicate moderate accuracy, and AUC values between 0.50 and 0.70 indicate low accuracy (Swets, 1988).

Classification accuracy is based on one specific cut point and is a metric of the screeners raw ability to perform a hard classification (i.e., at risk or not at risk). In comparison, the area under the ROC curve is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006). In order to provide information regarding the screener's hard classification and ranking ability both metrics will be reported. In addition, statistical comparisons of AUC values using the DeLong, DeLong, and Clark-Pearson (1988) non-parametric method of comparing AUC values are presented.

**Procedures**

The current study was conducted in four sessions: (1) two error recognition assessment sessions; (2) an individual reading assessment session (i.e., WIAT-III administration); and (3) a standard reading assessment session (i.e., DORF and Daze administration) spanning across 7 weeks. All assessments were conducted during regularly scheduled class time. The error recognition assessment and the Daze assessment were administered in a class wide format while the WIAT-III and the DORF measures were administered in an individual format. All assessments were administered in a standardized format as indicated in the WIAT-III and DIBELS Next administration procedures.

The experimental error recognition assessment was administered at the onset of the Winter screening period. Student materials and number two pencils were distributed to all students. Students were instructed to listen carefully to the pre-recoded reading of the passage and were told to circle any words that were read incorrectly. Students were

assured that the activity would not count toward their grade and were told to keep their eyes on their own paper. At the conclusion of the measure student materials were collected and the error recognition activity was concluded. Test-retest sessions occurred one week after the initial error recognition task and was administered in an identical fashion. The pencils provided at the onset of the error recognition activity were gifted to all students including students who did not participate in the study.

Teachers were asked to complete a questionnaire for each student participating in the error recognition assessment. The teacher questionnaire consisted of the following four questions: (1) Is the target student a good or poor reader? (2) Is the target student in a reading intervention? (3) Is the target student classified as an EL, if so what is the students CELDT level? and (4) What is the target student's Ethnicity?

## Results

### Preliminary Data Analysis

An examination of the percent of missing data for the study sample revealed that no missing data were present for the Error Recognition and DORF measures; however, 1.8% of the data for the Daze measure was missing. The prevalence of missing data was low. As such, listwise deletion was used in analyses involving the Daze measure. Descriptive statistics of the sample and classroom peers are reported in Table 2. Means and standard deviation values for all reading measures are reported in Table 3. An independent samples t-test was computed to identify if there was a difference in student achievement between participating and non-participating students on the DORF, Daze and ER measures. The results of these analyses indicated statistically significant

differences among participating and non-participating students for the DORF and Daze measures only (see Table 4). Non-participating students on average had lower scores on both the DORF and Daze measures (see Table 3).

In terms of types of errors identified, on average EO students identified a greater amount of errors across all five error types as compared to their EL counterparts (see Table 5). However, statistically significant differences in types of errors identified by EO students and ELs were present only for verb errors (see Table 6). Similarly, students not at risk for reading failure on average identified more errors across error types as compared to students at risk for reading failure. Statistically significant differences in errors identified were present across all five error types for students at risk as compared to students not at risk for reading failure (see Table 7).

**Research Question 1: Accuracy of Error Recognition Measure**

Figures 1-3 include several classification metrics for the experimental, DORF, and, Daze measures respectively. The classification metrics for the experimental measure were calculated using the cutoff values of 13, which yielded the highest discriminative ability. The classification metrics for the DIBELS measures were calculated using the cutoff values associated with the appropriate benchmark goals (i.e., not at risk = at benchmark or above and at risk = below benchmark goal). The classification accuracy of the error recognition experimental measure was 0.859 indicating that the experimental measure is approximately 86% accurate in identifying students at risk for reading failure. In comparison, the DORF and Daze measures were approximately 81% and 70% accurate respectively. Sensitivity and specificity data indicate that the experimental,

DORF, and Daze measures are 83.33% 79.415, and 82.35% effective respectively at identifying individuals at risk for reading failure and 85.29%, 82.61%, and 50% effective respectively at identifying individuals not at risk for reading failure respectively.

Figure 4 portrays the empirical non-parametric ROC curves for the error recognition, DORF and Daze reading screeners. The AUC values and 95% confidence intervals are included in Table 8. Analysis reveals a statistically significant AUC value of 0.833, 95% Confidence Interval (CI) = [0.712, 0.953], p-value = < 0.001 for the experimental error recognition reading screener. As such, the utility of this measure falls in the moderate range of AUC classifications. The utility of the DORF, AUC = 0.88, 95% CI = [0.795, 0.973], p-value < 0.001, and Daze, AUC = 0.79, 95% CI = [0.673, 0.910], p-value < 0.001 screeners fall within the moderate and low range respectively.

**Research question 2: Accuracy of Error Recognition compared to DORF and Daze**

Statistical comparisons of the AUC values were computed using the DeLong et al. (1988) non-parametric method (see Table 9). Although the AUC value associated with the error recognition task was smaller when compared to the AUC value associated with the DORF measure, this difference was not statistically significant Z= 0.872, 95% CI = [-0.061, 0.160], p-value = 0.38. Similarly, the error recognition task AUC value was larger than the Daze AUC value, however, this difference was also not statistically significant Z= 0.546, 95% CI = [-0.133, 0.199], p-value = 0.585.

**Research question 3: Error Recognition accuracy for ELs**

Figure 5 portrays the empirical non-parametric ROC curve for the error recognition screener and DIBELS screeners using only the results obtained for those

students classified as ELs. The AUC values and 95% confidence intervals for these analyses are also included in Table 8. Results differed for students classified as ELs; however, these analyses were based on only a small sample of 23 students. An AUC value of 0.738, 95% CI = [0.460, 1.00], p-value = 0.07 for the experimental error recognition reading screener falls within the moderate range when using only the data obtained for ELs; however, these results were not statistically significant. The AUC value obtained for the DORF measure was statistically significant at 0.925, 95% CI = [0.814, 1.00], p-value = 0.001 and fell within the high accuracy range. The AUC value obtained for the Daze measure, however, was not statistically significant at 0.738, 95% CI = [0.512, 0.963], p-value = 0.066 and fell within the moderate accuracy range.

## Discussion

In response to the growing concern about student reading failure, an RTI system has been integrated into several school districts. A primary aspect of an RTI system is the periodic screening of student reading skills. The purpose of this paper was to identify if an experimental error recognition task would demonstrate a greater degree of accuracy for identifying 3rd grade students at risk for reading failure than commonly used measures. The classification accuracy for the experimental error recognition task was moderate at approximately 86%. A screener's classification accuracy depends on both the ability of the screener to correctly identify students at risk and correctly identify students not at risk. The sensitivity and specificity data indicates that the error recognition measure is more effective than the DORF and Daze measures in both identifying students at risk for reading failure and identifying students not at risk (see Figures 1 – 3).

Screeners with high sensitivity and specificity allow for appropriate allocation of resources. That is, by correctly identifying students at risk and students not at risk over and under identification of students requiring Tier 2 and Tier 3 services can be minimized. However, Provost and Fawcett (1997) indicated that simple classification accuracy alone is often a poor metric for measure performance. Therefore, AUC values were also used to determine the accuracy of the experimental error recognition task for identifying students at risk for reading failure. The AUC obtained for the experimental error recognition task (AUC = 0.833) was within the moderate range, indicating that of two randomly chosen students, there is an 83.3% chance that the student at risk for reading failure will be ranked with a higher degree of suspicion than the student who is not at risk for reading failure. The 95% confidence interval for the AUC obtained for the experimental error recognition task was within the fair to excellent range. As such, the results of these analysis indicate that at minimum the error recognition screener is somewhat useful and at best very useful in distinguishing between 3rd grade students not at risk for reading failure and 3rd grade students at risk for reading failure.

When comparing the AUC values across all three screening measures the DORF screener has the greatest AUC value followed by the error recognition measure and the Daze measure respectively. The results of these analyses indicate that all three screeners fall within the moderate accuracy range of AUC value classification. Statistical comparisons of the AUC values are shown in Table 9. Analysis reveals differences between the three screening measures' ability to distinguish between two diagnostic groups were not significant. Given these insignificant differences in distinguishing

between diagnostic groups the use of the error recognition measure for screening purposes is essentially equal to that of current screeners. In addition, the error recognition measure is a multi-component measure of reading that combines phonemic awareness, phonological decoding, and phonics in a single group administered task. As such, the error recognition measure is likely an appropriate screening measure given the combination of various reading features, high technical adequacy, and efficiency of the measure for this sample.

The AUC obtained for the experimental error recognition task (AUC = 0.738), when including only students classified as ELs, was within the moderate range; however, it was not statistically significant. In addition, the standard error associated with this AUC value is quite high resulting in a 95% confidence interval ranging from 0.460 (not an informative test) to 1.000 (an excellent test). The Daze measure had nearly identical results as the error recognition measure and thus there were no advantages for using the error recognition measure over the Daze measure for the sample of ELs. The DORF measure, however, was within the high accuracy range. Thus these results indicate that the use of the DORF measure would be advantageous over the error recognition and Daze measure for ELs. However, the sample of ELs used in these analyses was extremely small at only 23 students. Thus, the results of these analyses should not be extended to the greater population of 3rd grade ELs.

In addition to classification accuracy and AUC values, information regarding the types of errors identified by students can also be derived using the error recognition measure. Preliminarily analyzes revealed ELs, as compared to EO students, identified a

significantly lower amount of verb errors. When comparing students at risk versus students not at risk significant differences were apparent across all error types. Essentially, students at risk for reading failure demonstrated greater difficulties in recognizing errors across all five error types. These results indicate that in general students at risk for reading failure need support across all error types. Individual student differences, however, may demonstrate more specific student needs.

## Implications

The results of these analyses indicate that the error recognition task is only slightly less accurate than the commonly used DORF measures and is slightly more accurate than the commonly used Daze measure; however, these differences were not statistically significant. Traditionally, measure effectiveness is used to determine the use of a measure in practice. However, measure efficiency and cost can affect its sustainability in practice (Lemons, Fuchs, Gilbert, & Fuchs, 2014). That is, if measure effectiveness is comparable and not statistically different, but also more efficient or cost effective, then the measure is presumably more sustainable in practice. The DIBELS measures are widely used in practice, no doubt largely due to the low cost of the measures and efficiency in both administration and result analysis interpretation. However, administering screening probes individually to students can often represent a challenging task for teachers, given the various tasks teachers are also responsible for. Thus, minimizing the amount of time spent administering screening measures can be beneficial for teachers. The error recognition measure demonstrates comparable accuracy to DORF measures in identifying students at risk for reading failure with a greater degree

of efficiency. However, given that the DORF measure in the study demonstrated a slightly greater degree of accuracy and is currently in place in several school districts, recommendations are primarily geared towards researchers. That is, researchers should attempt to replicate and improve upon the error recognition task described in this paper before practitioners make use of these findings.

## Limitations/ Future Directions

The results of the current study are subject to limitations. One such limiting factor is the small sample size, especially for ELs, and the sampling technique used. A ROC curve generated from a finite number of instances is actually a step function that approaches a true curve as the number of instances approach infinity (Fawcett, 2006). As such a larger sample would better approximate a true curve. In addition, the sample included in the study differed from non-participating classroom peers on both the DORF and Daze measures. On average non-participating students had significantly lower DORF and Daze scores. As such, the generalizability of the study results are limited.

Future research in this area should attempt to improve the error recognition task described in this study and replicate the results across a variety of samples. The measure itself can be improved in several ways. First and foremost, with the increased use of technology in the classroom it would be beneficial if this measure was integrated into a computer based program thus allowing for efficacy in both administration and scoring procedures. Second, the reading speed of the audio recording can be varied in order to identify whether speed impacts the accuracy of the measure. Third, with a larger more representative sample of 3rd grade students, further analyses can be computed from the

data. That is, given that the error recognition measure can control for the types of errors included in the measure, further analyses based on the types of errors made by at risk versus not at risk students can be derived.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. Amer Educational Research Assn.

Ardoin, S. P., & Christ, T. J. (2008). Evaluating curriculum-based measurement slope estimates using data from triannual universal screenings. *School Psychology Review*, *37*(1), 109.

Brown, A. L. (1980). Metacognitive development and reading. In R. J. Spiro, B. C.

Bruce, & W. F. Brewer (Ed.) *Theoretical issues in reading comprehension*. Hillsdale, NJ: Erlbaum.

Brown, A. L. (1982). Learning how to learn from reading. *Reader meets author: Bridging the gap*, 26-54.

Brown, A. L., Palincsar, A. S., & Purcell, L. (l986). Poor readers: Teach, don't label. In U. Neisser (Ed.) *The School Achievement of Minority Children: New Perspectives*. (pp. 105-143) NJ: Lawrence Erlbaum Associates.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845.

Deno, S. L., Reschly, A. L., Lembke, E. S., Magnusson, D., Callender, S. A., Windram, H., & Stachel, N. (2009). Developing a school-wide progress-monitoring system. *Psychology in the Schools*, *46*(1), 44-55.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

Fischer, J. E., Bachmann, L. M., & Jaeschke, R. (2003). A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive care medicine*, *29*(7), 1043-1051.

Fletcher, JM., Lyon, GR., Fuchs, LS., Barnes, MA. Learning disabilities: From identification to intervention. Guilford; New York, NY: 2007.

Garner, R., & Taylor, N. (1982). Monitoring of understanding: An investigation of attentional assistance needs at different grade and reading proficiency levels. *Reading Psychology: An International Quarterly*, *3*(1), 1-6.

Glover, T. A., & Albers C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology,* 45(2), 117-135.

Good, R. H., Kaminski, R. A., Dewey, E. N., Wallin, J., Powell--Smith, K. A., & Latimer, R. J. (2013). *DIBELS Next Technical Manual*.

Goffreda, C. T., Diperna, J. C., & Pedersen, J. A. (2009). Preventive screening for early readers: Predictive validity of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). *Psychology in the Schools*, *46*(6), 539-552.

Hajian-Tilaki, K. O., Hanley, J. A., Joseph, L., & Collet, J. P. (1997). A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. *Medical Decision Making*, *17*(1), 94-102.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

Hasbrouck, J., & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher*, *59*(7), 636-644.

Jefferson, R. E., Grant, C. E., & Sander, J. B. (2016). Effects of Tier I Differentiation and Reading Intervention on Reading Fluency, Comprehension, and High Stakes Measures. *Reading Psychology*, 1-28.

Jenkins, J. R., & Johnson, E. S. (2008). *Universal screening for reading problems: Why and how should we do this*? Retrieved December 22, 2015 from: http://www.rtinetwork.org/essential/assessment/screening/readingproblems

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice*, *24*(4), 174-185.

Kuhn, M. (2004). Helping students become accurate, expressive readers: Fluency instruction for small groups. *The Reading Teacher*, *58*(4), 338-344.

Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, *43*(5), 242-252.

Linan-Thompson, S., Vaughn, S., Hickman-Davis, P., & Kouzekanani, K. (2003). Effectiveness of supplemental reading instruction for second-grade English language learners with reading difficulties. *The Elementary School Journal*, *103*(3), 221-238.

Mathes, P. G., & Denton, C. A. (2002, September). The prevention and identification of reading disability. In *Seminars in Pediatric Neurology* (Vol. 9, No. 3, pp. 185-191). WB Saunders.

Moon, H. (2016). The Relationship of Curriculum-Based Measurement Oral Reading as an Indicator of Reading Achievement. *Culminating Projects in Special Education*, Paper 27.

National Assessment of Educational Progress – 2015 Mathematics & Reading Assessments. (2015, October 28). Retrieved November 12, 2015, from http://www.nationsreportcard.gove/reading_math_2015/#?grade=4

National Reading Panel. (2000). *Report of the National Reading Panel: Reports of the subgroups*. Washington, DC: National Institute of Child Health and Human Development Clearinghouse.

O'Connor, R.E., Bocian, K.M., Sanchez, V., Beach, K.D., & Flynn, L.J. (2013). Special education in a four-year response to intervention (RtI) environment: Characteristics of students with learning disability and grade of identification. *Learning Disabilities Research & Practice, 28*, 98-112.

O'Connor, R. E., & Jenkins, J. R. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading*, *3*(2), 159-197.

Oka, E., & Paris, S. (1986). Patterns of motivation and reading skills in underachieving children. In S. Ceci (Ed.), *Handbook of cognitive, social, and neuropsychological aspects of learning disabilities* (Vol. 2). Hillsdale, NJ: Erlbaum.

Parker, D. C., Zaslofsky, A. F., Burns, M. K., Kanive, R., Hodgson, J., Scholin, S. E., & Klingbeil, D. A. (2015). A Brief Report of the Diagnostic Accuracy of Oral Reading Fluency and Reading Inventory Levels for Reading Failure Risk Among Second-and Third-Grade Students. *Reading & Writing Quarterly*, *31*(1), 56-67.

Provost, F. J., & Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *KDD* (Vol. 97, pp. 43-48).

Salvia, J., Ysseldyke, J., & Bolt, S. (2012). *Assessment: In special and inclusive education*. Cengage Learning.

Schilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *The Elementary School Journal*, *107*(5), 429-448.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240* (4857), 1285.

Torgesen, J. K. (2009). The Response to Intervention Instructional Model: Some Outcomes From a Large-Scale Implementation in Reading First Schools. *Child Development Perspectives*, *3*(1), 38-40.

Vaughan-Jensen, J., Adame, C., McLean, L., & Gámez, B. (2011). Test Review: D. Wechsler Wechsler Individual Achievement Test. San Antonio, TX: Pearson, 2009. *Journal of Psychoeducational Assessment*, *29*(3), 286-291.

Wanzek, J., & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *The promise of response to intervention: Evaluating current science and practice*, *39*(2), 113-142.

Yeo, S. (2009). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education,* 1- 12.

# Tables

Table 1
*Correlations for Reading Measures*

| | ER1 | ER2 | WIAT-III | DORF | DAZE |
|---|---|---|---|---|---|
| ER1 | 1 | | | | |
| ER2 | .75 | 1 | | | |
| WIAT-III | .65 | .88 | 1 | | |
| DORF | .72 | .73 | .75 | 1 | |
| DAZE | .46 | .57 | .59 | .67 | 1 |

*Note*: ER1 = Error Recognition Test 1; ER2 = Error Recognition Test 2 Test re-test; WIAT-III = Wechsler Individual Achievement Test- Third Edition; DORF = DIBELS Oral Reading Fluency.

Table 2
*Descriptive Statistics of Sample vs. Peers*

| | Sample | | | Peers | | |
|---|---|---|---|---|---|---|
| | Male | Female | Total | Male | Female | Total |
| N | 22 | 35 | 57 | 23 | 17 | 40 |
| Age Range | 7:11–8:1 | 7:11–8:11 | 7:11–9:1 | 8:0-9:1 | 8:0-9:1 | 8:0-9:1 |
| EL | 9 | 14 | 23 | 5 | 4 | 9 |
| CELDT Range | 1– 4 | 2 - 5 | 1 - 5 | 1 - 3 | 3 | 1 - 3 |

*Note*: EL = English Learner; CELDT = California English Language Development Test.

Table 3
*Mean and Standard Deviations for Sample vs. Peers*

| | Total Sample | | | English Learners | | | Peers | | |
|---|---|---|---|---|---|---|---|---|---|
| | Male | Female | Total | Male | Female | Total | Male | Female | Total |
| ER1 | 11.95 | 14.09 | 13.26 | 10.89 | 11.71 | 11.39 | 11.7 | 14.92 | 13.45 |
| | (6.09) | (5.59) | (5.83) | (6.95) | (6.59) | (6.59) | (5.72) | (3.50) | (4.81) |
| ER2 | 12.18 | 11.91 | 12.02 | 11.22 | 9.50 | 10.17 | 11.3 | 14 | 12.77 |
| | (5.49) | (4.70) | (4.97) | (5.78) | (4.42) | (4.94) | (5.98) | (4.02) | (5.07) |
| DORF | 85.77 | 96.14 | 92.14 | 81.11 | 77.29 | 78.78 | 53.7 | 95.17 | 76.32 |
| | (33.29) | (35.20) | (34.55) | (33.86) | (29.06) | (30.32) | (26.73) | (42.19) | (41.03) |
| Daze | 13.71 | 14.34 | 14.11 | 11.11 | 11 | 11.04 | 7.1 | 14.08 | 10.91 |
| | (6.93) | (6.77) | (6.78) | (6.39) | (6.42) | (6.26) | (4.56) | (8.32) | (7.60) |

*Note*: ER1 = Error Recognition Test 1; ER2 = Error Recognition Test 2 Test re-test; Standard deviation values shown in brackets.

Table 4
*Sample vs. Peer Independent Samples Test*

| | t | df | p-value | Mean Difference | Std. Error Difference | 95% Confidence Interval | |
|---|---|---|---|---|---|---|---|
| | | | | | | Lower Bound | Upper Bound |
| DORF | 2.055 | 91 | 0.043 | 14.446 | 7.029 | 0.482 | 28.409 |
| DAZE | 2.736 | 90 | 0.007 | 3.801 | 1.389 | 1.105 | 6.499 |
| ER1 | -.137 | 91 | 0.891 | -0.191 | 1.397 | -2.974 | 2.591 |
| ER2 | -.602 | 91 | 0.549 | -0.755 | 1.255 | -3.254 | 1.744 |

*Note*: ER1 = Error Recognition Test 1; ER2 = Error Recognition Test 2 Test re-test; df = degrees of freedom; DORF = DIBELS Oral Reading Fluency.

Table 5
*Mean and Standard Deviations for Types of Errors Identified on ER Measure*

| | EL Sample | Non-EL Sample | At Risk Sample | Not At Risk Sample |
|---|---|---|---|---|
| Preposition | 2.4 | 2.9 | 2 | 3.2 |
| | (1.6) | (1.4) | (1.1) | (1.5) |
| Verb | 1.5 | 2.1 | 1.1 | 2.7 |
| | (1.3) | (1.0) | (0.9) | (1.0) |
| Noun | 2.4 | 3.1 | 2 | 3.4 |
| | (1.7) | (1.2) | (1.4) | (1.3) |
| Pro-Noun | 2.3 | 3.1 | 2 | 3.3 |
| | (1.9) | (1.4) | (1.8) | (1.2) |
| Adjective | 3.0 | 3.8 | 2.7 | 4.2 |
| | (1.7) | (1.2) | (1.4) | (1.2) |

*Note*: ER = Error Recognition; EL = English Learner; Standard Deviation values shown in brackets.

Table 6
*EL vs. Non-EL Independent Samples Test*

| | t | df | p-value | Mean Difference | Std. Error Difference | 95% Confidence Interval | |
|---|---|---|---|---|---|---|---|
| | | | | | | Lower Bound | Upper Bound |
| Preposition | -1.20 | 55 | 0.234 | -0.477 | 0.396 | -1.271 | 0.317 |
| Verb | -2.20 | 55 | 0.032 | -0.669 | 0.305 | -1.279 | -0.584 |
| Noun | -1.86 | 55 | 0.068 | -0.726 | 0.390 | -1.509 | 0.056 |
| Pro-Noun | -1.22 | 55 | 0.226 | -0.536 | 0.438 | -1.413 | 0.342 |
| Adjective | -1.96 | 55 | 0.055 | -0.751 | 0.383 | -1.518 | 0.016 |

*Note*: EL = English Learner.

Table 7
*At Risk vs. Not at Risk Independent Samples Test*

| | t | df | p-value | Mean Difference | Std. Error Difference | 95% Confidence Interval | |
|---|---|---|---|---|---|---|---|
| | | | | | | Lower Bound | Upper Bound |
| Preposition | -3.29 | 55 | 0.002 | -1.235 | 0.375 | -1.987 | -0.484 |
| Verb | -5.04 | 55 | <0.001 | -1.325 | 0.263 | -1.851 | -0.798 |
| Noun | -3.88 | 55 | <0.001 | -1.382 | 0.357 | -2.097 | -0.667 |
| Pro-Noun | -3.30 | 55 | 0.002 | -1.338 | 0.405 | -2.150 | -0.525 |
| Adjective | -4.05 | 55 | <0.001 | -1.407 | 0.347 | -2.103 | -0.710 |

Table 8
*Area Under the Curve*

| | Area | Standard Error | p-value | 95% Confidence Interval | |
|---|---|---|---|---|---|
| | | | | Lower Bound | Upper Bound |
| ER1 | 0.831 | 0.064 | 0.000 | 0.706 | 0.957 |
| DORF | 0.880 | 0.047 | 0.000 | 0.789 | 0.972 |
| Daze | 0.788 | 0.062 | 0.000 | 0.667 | 0.909 |
| ER1EL | 0.738 | 0.142 | 0.066 | 0.460 | 1.000 |
| DORFEL | 0.925 | 0.038 | 0.001 | 0.814 | 1.000 |
| DazeEL | 0.738 | 0.115 | 0.066 | 0.512 | 0.963 |

*Note*: ER = Error Recognition; WIAT-III = Wechsler Individual Achievement Test- Third Edition; DORF = DIBELS Oral Reading Fluency; EL = English Learners.

Table 9
*Area Under the Curve Comparisons*

| | ER1 vs. DORF | ER1 vs. DAZE | DORF vs. DAZE |
|---|---|---|---|
| Difference | 0.049 | 0.043 | 0.093 |
| Standard Error | 0.056 | 0.080 | 0.057 |
| 95% CI | (-0.061, 0.160) | (-0.113, 0.199) | (-0.019, 0.205) |
| Z statistic | 0.872 | 0.546 | 1.620 |
| p-value | 0.383 | 0.585 | 0.105 |

*Note*: ER = Error Recognition; WIAT-III = Wechsler Individual Achievement Test- Third Edition; DORF = DIBELS Oral Reading Fluency.

Figures

| | | **Predicted Condition** | | | |
|---|---|---|---|---|---|
| | **Total Population**<br>57 | **Positive**<br>24 | **Negative**<br>33 | **Prevalence**<br>$\frac{\Sigma Condition\ Positve}{\Sigma Total\ Population}$<br>$\frac{24}{57} = 42.11\%$ | |
| **True Condition** | **Condition Positive**<br>24 | **True Positive**<br>20 | **False Negative (Type II Error)**<br>4 | **Sensitivity**<br>$\frac{\Sigma True\ Positve}{\Sigma Condition\ Positve}$<br>$\frac{20}{24} = 83.33\%$ | **False Negative Rate**<br>$\frac{\Sigma False\ Negative}{\Sigma Condition\ Positve}$<br>$\frac{4}{24} = 16.67\%$ |
| | **Condition Negative**<br>33 | **False Positive (Type I Error)**<br>4 | **True Negative**<br>29 | **False Positive Rate**<br>$\frac{\Sigma False\ Positve}{\Sigma Condition\ Negative}$<br>$\frac{4}{34} = 11.76\%$ | **Specificity**<br>$\frac{\Sigma True\ Negative}{\Sigma Condition\ Negative}$<br>$\frac{29}{34} = 85.29\%$ |
| | **Accuracy =**<br>$\frac{\Sigma TruePositve + \Sigma TrueNegative}{\Sigma Total\ Population}$<br>$\frac{20+29}{57} = \frac{49}{57}$<br>$= 85.96\%$ | **Positive Predictive Value (Precision)**<br>$\frac{\Sigma TruePositve}{\Sigma Test\ outcomepositive}$<br>$\frac{20}{24} = 83.33\%$ | **False Omission Rate**<br>$\frac{\Sigma False\ Negative}{\Sigma Test\ outcomenegative}$<br>$\frac{4}{33} = 12.12\%$ | **Positive Likelihood Ratio**<br>$\frac{Sensitivity}{FPR} = \frac{0.833}{0.118}$<br>$= 7.06$ | **Diagnostic Odds Ratio**<br>$\frac{Positive\ LR}{Negative\ LR} = \frac{7.06}{0.20}$<br>$= 35.3$ |
| | | **Negative Predictive Value**<br>$\frac{\Sigma False\ Positive}{\Sigma Test\ outcomepositive}$<br>$\frac{4}{24} = 16.67\%$ | **Negative Predictive Value**<br>$\frac{\Sigma TrueNegative}{\Sigma Test\ outcomenegative}$<br>$\frac{29}{33} = 87.88\%$ | **Negative Likelihood Ratio**<br>$\frac{FNR}{Specificity} = \frac{0.174}{0.856}$<br>$= 0.20$ | |

*Figure 1*. Error Recognition Classification Accuracy Metrics

| | | Predicted Condition | | | |
|---|---|---|---|---|---|
| | **Total Population**<br>57 | **Positive**<br>26 | **Negative**<br>31 | **Prevalence**<br>$\frac{\Sigma Condition\ Positve}{\Sigma Total\ Population}$<br>$\frac{23}{57} = 40.35\%$ | |
| **True Condition** | **Condition Positive**<br>23 | **True Positive**<br>19 | **False Negative (Type II Error)**<br>4 | **Sensitivity**<br>$\frac{\Sigma True\ Positve}{\Sigma Condition\ Positve}$<br>$\frac{19}{23} = 82.61\%$ | **False Negative Rate**<br>$\frac{\Sigma False\ Negative}{\Sigma Condition\ Positve}$<br>$\frac{4}{23} = 17.39\%$ |
| | **Condition Negative**<br>34 | **False Positive (Type I Error)**<br>7 | **True Negative**<br>27 | **False Positive Rate**<br>$\frac{\Sigma False\ Positve}{\Sigma Condition\ Negative}$<br>$\frac{7}{34} = 20.59\%$ | **Specificity**<br>$\frac{\Sigma True\ Negative}{\Sigma Condition\ Negative}$<br>$\frac{27}{34} = 79.41\%$ |
| | **Accuracy =**<br>$\frac{\Sigma TruePositve + \Sigma TrueNegative}{\Sigma Total\ Population}$<br>$\frac{19+27}{57} = \frac{46}{57}$<br>$= 80.70\%$ | **Positive Predictive Value (Precision)**<br>$\frac{\Sigma TruePositve}{\Sigma Test\ outcomepositive}$<br>$\frac{19}{26} = 73.08\%$ | **False Omission Rate**<br>$\frac{\Sigma False\ Negative}{\Sigma Test\ outcomenegative}$<br>$\frac{4}{31} = 12.90\%$ | **Positive Likelihood Ratio**<br>$\frac{Sensitivity}{FPR} = \frac{0.826}{0.205}$<br>$= 4.03$ | **Diagnostic Odds Ratio**<br>$\frac{Positive\ LR}{Negative\ LR} = \frac{4.03}{0.22}$<br>$= 18.32$ |
| | | **Negative Predictive Value**<br>$\frac{\Sigma False\ Positive}{\Sigma Test\ outcomepositive}$<br>$\frac{7}{26} = 26.92\%$ | **Negative Predictive Value**<br>$\frac{\Sigma True Negative}{\Sigma Test\ outcomenegative}$<br>$\frac{27}{31} = 87.10\%$ | **Negative Likelihood Ratio**<br>$\frac{FNR}{Specificity} = \frac{0.174}{0.794}$<br>$= 0.22$ | |

*Figure 2*. Oral Reading Fluency Classification Accuracy Metrics Using Benchmark Goal

| | | Predicted Condition | | | |
|---|---|---|---|---|---|
| | **Total Population**<br>57<br>(1 missing value in DAZE data)<br>56 | **Positive**<br>17 | **Negative**<br>39 | **Prevalence**<br>$\frac{\Sigma Condition\ Positve}{\Sigma Total\ Population}$<br>$\frac{22}{56} = 39.29\%$ | |
| **True Condition** | **Condition Positive**<br>22 | **True Positive**<br>11 | **False Negative (Type II Error)**<br>11 | **Sensitivity**<br>$\frac{\Sigma True\ Positve}{\Sigma Condition\ Positve}$<br>$\frac{11}{22} = 50.00\%$ | **False Negative Rate**<br>$\frac{\Sigma False\ Negative}{\Sigma Condition\ Positve}$<br>$\frac{11}{22} = 50.00\%$ |
| | **Condition Negative**<br>34 | **False Positive (Type I Error)**<br>6 | **True Negative**<br>28 | **False Positive Rate**<br>$\frac{\Sigma False\ Positve}{\Sigma Condition\ Negative}$<br>$\frac{6}{34} = 17.65\%$ | **Specificity**<br>$\frac{\Sigma True\ Negative}{\Sigma Condition\ Negative}$<br>$\frac{28}{34} = 82.35\%$ |
| | **Accuracy =**<br>$\frac{\Sigma TruePositve + \Sigma TrueNegative}{\Sigma Total\ Population}$<br>$\frac{11+28}{56} = \frac{39}{56}$<br>$= 69.64\%$ | **Positive Predictive Value (Precision)**<br>$\frac{\Sigma TruePositve}{\Sigma Test\ outcomepositive}$<br>$\frac{11}{17} = 64.71\%$ | **False Omission Rate**<br>$\frac{\Sigma False\ Negative}{\Sigma Test\ outcomenegative}$<br>$\frac{11}{39} = 28.21\%$ | **Positive Likelihood Ratio**<br>$\frac{Sensitivity}{FPR} = \frac{0.500}{0.177}$<br>$= 2.82$ | **Diagnostic Odds Ratio**<br>$\frac{Positive\ LR}{Negative\ LR} = \frac{2.82}{0.61}$<br>$= 4.62$ |
| | | **Negative Predictive Value**<br>$\frac{\Sigma False\ Positive}{\Sigma Test\ outcomepositive}$<br>$\frac{6}{17} = 35.29\%$ | **Negative Predictive Value**<br>$\frac{\Sigma TrueNegative}{\Sigma Test\ outcomenegative}$<br>$\frac{28}{39} = 71.79\%$ | **Negative Likelihood Ratio**<br>$\frac{FNR}{Specificity} = \frac{0.500}{0.824}$<br>$= 0.61$ | |

*Figure 3*. Daze Classification Accuracy Metrics Using Benchmark Goal

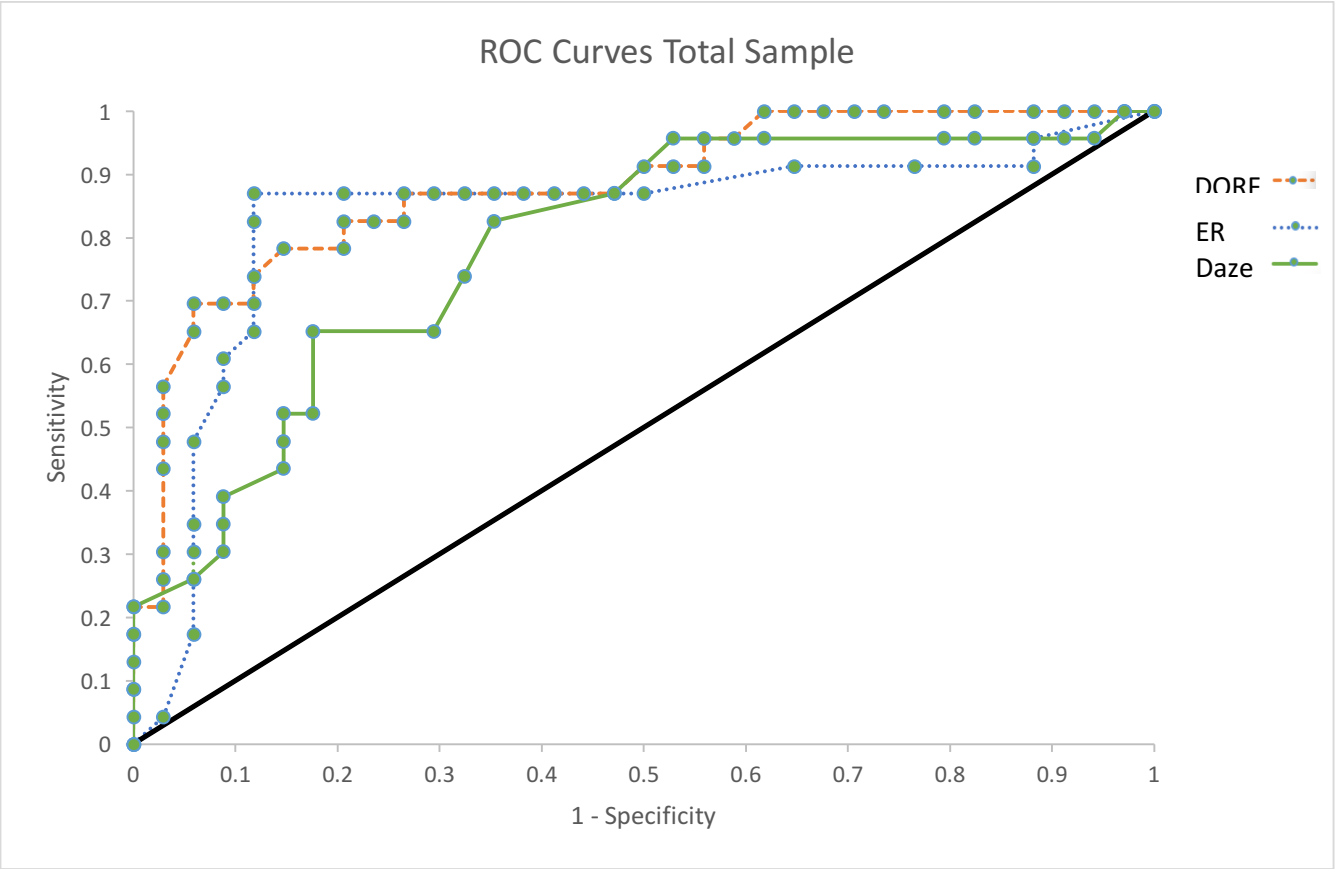


*Figure 4*. Empirical Non-Parametric ROC Curves Total Sample

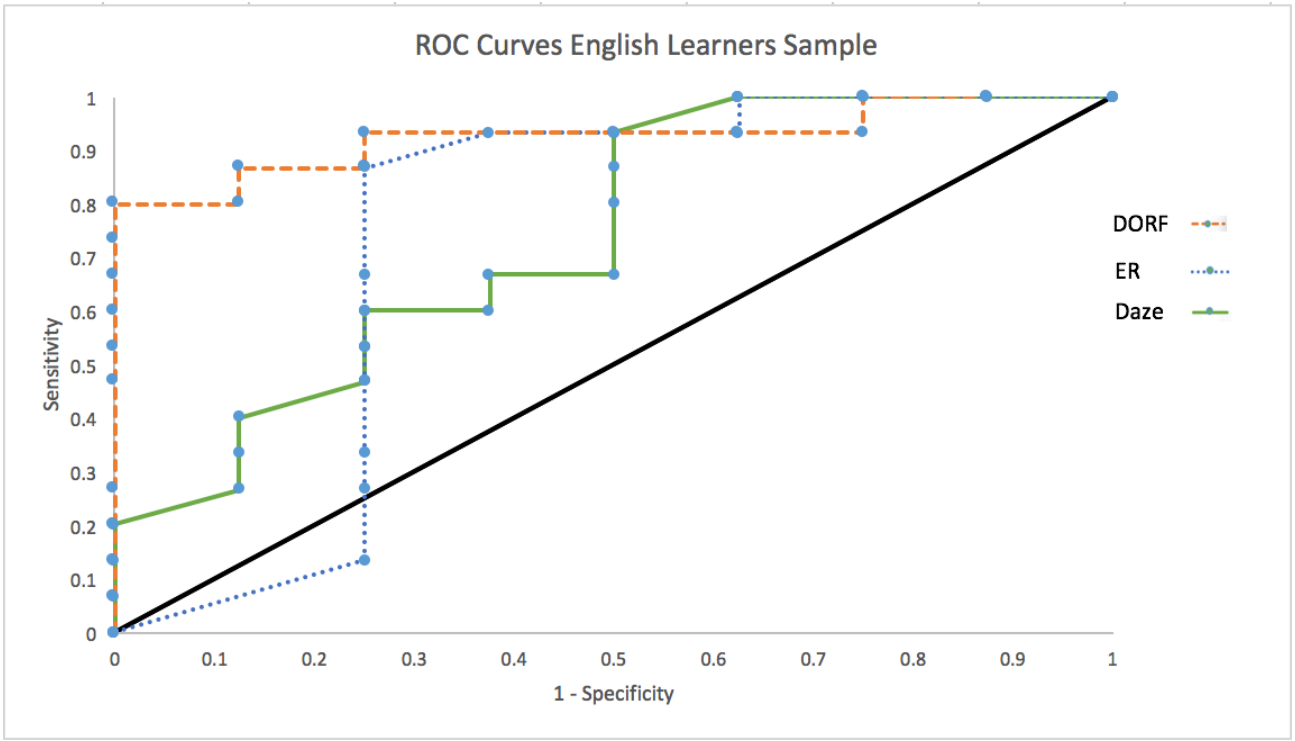


*Figure 5*. Empirical Non-Parametric ROC Curve English Learner Sample