2011

The Importance of Predictive Power in Early Screening Assessments: Implications for Placement in the Response to Intervention Framework

Yaacov M. Petscher, Young-Suk Kim, and Barbara R. Foorman

The Importance of Predictive Power in Early Screening Assessments: Implications for Placement in the RTI Framework

Yaacov Petscher

Young-Suk Kim

Barbara R. Foorman

Florida Center for Reading Research

Florida State University

Abstract

As schools implement Response-to-Intervention (RTI) to identify and serve students with learning difficulties, it is critical for educators to know how to evaluate screening measures. In the present study, DIBELS oral reading fluency was used to compare the differential decisions that might occur in screening accuracy when predicting two reading comprehension measures (i.e., *Stanford Achievement Test – 10th Edition* [SAT10] & *Gates-McGinitie Reading Test -4th Edition* [GMRT]) at the end of second grade. The results showed that the DIBELS oral reading fluency tended to have higher sensitivity and negative predictive power for SAT10, and higher specificity and positive predictive power for GMRT. Furthermore, attempting to achieve a criterion of positive predictive power for a given reading comprehension outcome (SAT10 in this study) appears to render a favorable balance compared to other indices of diagnostic accuracy. These results are discussed in light of trade-offs and a need for considering specific contexts of schools and districts.

Key words: Screening accuracy, Predictive Power, RTI, DIBELS oral reading fluency, Reading Comprehension,

The Importance of Predictive Power in Early Screening Assessments: Implications for Placement in the RTI Framework

Assessment is at the center of a Response to Intervention (RTI) framework. As a preventive service delivery model, the early identification of students who are at risk for future reading failure is the key to appropriately placing students into interventions. Thus, universal screening is a critical first step in most RTI models (Jenkins, Hudson, & Johnson, 2007). Effective screening measures are typified by brevity and ease of use, and should demonstrate high accuracy in predicting whether students will succeed or on a criterion outcome of interest (e.g., standardized reading assessment, state achievement test). When scores from screening assessments are validated, they are typically designed to maximize a particular statistical outcome (Streiner, 2003; e.g., correct classification according to a gold standard outcome, reducing the number of under-identified students). Thus, it is imperative that educators and researchers are informed about the trade-offs of maximizing different outcomes. This paper discusses the important statistical and methodological components that should be considered when choosing a screening assessment, and highlights such considerations with an illustration using the Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency test [DIBELS ORF; Good et al., 2002]) and two widely used reading comprehension tests to draw attention to the differences in at-risk identification and to provide general guidelines for comparing assessments for early screening.

Background and Context

An ideal screening assessment requires several features. First, practical utility is an important criterion such that it should be inexpensive, brief, easy to administer, score,

and interpret, and ideally easily linked to instruction (Schatschneider, Petscher, & Williams, 2008). Another critical, and often assumed, quality of a screener is high screening accuracy and discrimination so that it distinguishes with precision students who will develop difficulties in a target area (e.g., reading) from those who will not (Glover & Albers, 2007). Ultimately, all of these features serve students by identifying them accurately with respect to risk status and enabling appropriate allocation of resources for effective intervention.

When evaluating new screening assessments, two considerations must be made that impact the screening accuracy; namely, what measure will be used as the "gold standard" outcome and what psychometric properties of screening accuracy are maximized. Both decisions are equally important to the evaluation process as the former will inform the assessor to what the screener is predicting, and the latter will be informative about the goal of the screener. These two components interact to provide information about the extent to which scores from a screener are valid and meaningful for identifying potentially at-risk students of interest (i.e., difficulties in reading).

The outcome to which the developer wishes to predict is a critical decision because it serves as the criterion for how risk is operationally defined, and is the foundation on which all the statistical indices are based. For example, the screening accuracy of the Phonological Awareness Literacy Screening in kindergarten (PALS; Invernizzi, Meier, Swank, & Juel, 1999) was evaluated by using students' performance on the PALS in the fall to predict spring risk on the PALS . The cut-points for the fall kindergarten DIBELS Letter Naming Fluency and Initial Sound Fluency tasks were chosen based upon students' spring of kindergarten and fall of first grade performances

on several of the DIBELS measures. Conversely, the authors of the Test of Silent Word Reading Fluency (TOSWRF; Mather, Hammill, Allen, & Roberts, 2004) were most interested in evaluating what percent of students who performed poorly on the screener also performed poorly on a series of standardized, norm-referenced measures such as the Letter Word Identification and Passage Comprehension tasks on the Woodcock-Johnson Revised (Woodcock & Johnson, 1990).

The three different approaches described above represent commonly used techniques for selecting the gold-standard outcome. One may choose to use a later time point of the same measure/battery, such as in the case of PALS, or one may choose to use a different measure within the same test battery, such as with DIBELS. Alternatively, a different outcome may be selected (e.g., in the case of the TOSWRF), so that practitioners and researchers may evaluate how well it predicts risk on an external measure such as a state achievement test.

Although the psychometrics of the scores may appear to be strong when using a particular outcome, it is important that the practitioner understands for which outcome that risk is defined. In the above examples, the early identification of students at risk will only be specific to that particular assessment, regardless of what outcome each screener predicts. Thus, the outcome selected operationally defines what risk status truly is. It behooves researchers and practitioners, therefore, to select screening and outcome measures that result in reliable data and valid decisions, *and* are conceptually-defensible measure of reading success. Educators should be aware that even when the outcome measures meet all the psychometric requirements, outcome measures vary in random and systematic measurement errors, and the extent or emphasis of areas in the target construct.

This is particularly true for reading comprehension because reading comprehension is not a unitary construct, but draws on multiple processes (Cain, Oakhill, & Bryant, 2004; Davis, 1994). Thus, the extent to which different reading comprehension tests tap into different subskills of reading such as word recognition, working memory, and language comprehension varies (Andreassen & Braten, 2010; Keenan, Betjeman, & Olson, 2008). For example, the *Stanford Achievement Test – 10th edition* (SAT-10; Harcourt Brace, 2004)SAT10 and *Gates-MacGinitie Reading Test – 4th edition* (GMRT; *MacGinitie & MacGinitie* , 2006), which are two frequently used reading comprehension measures, are purported to measure the construct of reading comprehension, and have excellent psychometric properties. However, these two tests may vary in the extent to which they measure different subprocesses of reading (e.g., word reading and language comprehension; inference making) and various types of texts (e.g., expository vs. narrative texts).

When a group of individuals are administered a screener and a gold standard outcome, a resulting contingency matrix may be generated (Table 1). From this matrix, there are four types of classifications which may occur (Schatschneider et al., 2008): students who were identified as at-risk on the screen and either failed the outcome (Cell A, true positive) or passed the outcome (Cell B, false positive), and students who were identified as not at-risk on the screen and either failed the outcome (Cell C, false negative) or passed the outcome (Cell D, true negative). In general, most screening and diagnostic measures try to maximize what are considered to be either population-based or sample-based indices. Population-based indices are statistical proportions that describe the population level of risk according to the gold standard outcome that is chosen, and

describe the sensitivity and specificity of the scores. The sensitivity of a screener is the proportion of individuals who failed the outcome and were identified as at-risk on the screener; from Table 1, sensitivity may be calculated with A/(A+C). Specificity [D/(D+B)] is the proportion of individuals who pass the outcome test in the population who are not at-risk on the screening assessment. Sensitivity has been an important index in the RTI framework because it is the percentage of children correctly identified by a screener as needing further assessments and/or intervention. Several recommendations have been provided about appropriate thresholds for sensitivity and specificity; however, many researchers attempt to have levels of at least 0.80, with some recommending minimum values of 0.90 (Compton, Fuchs, Fuchs & Bryant, 2006; Jenkins 2003). A useful summative measure may be used to describe the proportion of students who were correctly identified as either at-risk or not at-risk. The overall correct classification index (OCC) may be calculated with [(A+D)/(A+B+C+D)].

Positive predictive power and negative predictive power are the two primary sample-based indices. Predictive power describes the proportion of students screened who ultimately perform successfully or poorly on the gold-standard outcome. Positive predictive power is the percentage of students identified as at-risk on the screen who fail the outcome test [calculated with A/(A+B)], while negative predictive power [D/(C+D)] is the percentage of students identified as not at-risk on the screen who pass the outcome test. These sample-based indices differ from the population-based indices as they are considered to be based on the makeup of the sample. Whereas sensitivity and specificity are properties of the test itself (Streiner, 2003), sample-based indices are dependent on the proportion of students in the sample that are at risk (i.e., base rate). Thus, if a screener

was used in two separate samples where one was higher achieving than the other, similar estimates of sensitivity and specificity could be obtained while different values for the positive and negative predictive power would be calculated.

Consider an example where a screener is used in two schools, each with 2,000 total students. In School A, 50% of the students were "at risk" based on the state achievement test, while 15% were "at risk" on the same test in School B, and for the screener selected, the reported sensitivity was 0.95 and the specificity was 0.90. Using this information, the contingency tables in Table 2 were constructed. In Schools A and B the sensitivity was 0.95 [i.e., 950/(950+50) in School A; 285/(285+15) in School B] and the specificity was 0.90 [i.e., 900/(900+100) in School A; 1,530/(1530+170) in School B]. As expected, these population-based indices are identical in both schools. However, when the sample-based indices are calculated, very different findings are observed. The positive predictive power in School A is 0.90 [i.e., 950/(950+100)] compared to 0.63 in School B [i.e., 285/(285+170)]; while the negative predictive power in School A is 0.95 [i.e., 900/(900+50)] compared to 0.99 in School B [i.e., 1530/(1530+15)]. This illustration demonstrates the critical importance of understanding and attending to the sample-based statistics, which should be given greater credence than the population based indices when evaluating screening accuracy and when screens are used to predict to distal outcome performance.

Though perfect screening (i.e., 100% screening accuracy) is desirable, it is elusive due to both the inherent measurement error associated with assessments, as well as the difficulties in measuring developing skills in children (Jenkins, Hudson, & Johnson, 2007). In practice, educators and researchers need to identify their needs and consider

trade-offs of the statistical outcomes described above to determine the screener that best fits the needs of a school, district, or research project. For instance, if a school uses a screener to identify and provide interventions to as many students who may potentially fail the outcome despite demands on resources, a screen with high sensitivity may be more appropriate. In contrast, if a school uses a screen to identify children who may need further monitoring, a screening device with a high negative predictive power may be better suited because such a screener would do a better job in identifying students with a low chance of developing a problem and thus not needing intervention (Schatschneider, et al., 2008).

In summary, when either the gold standard outcome changes or varying psychometric properties are maximized for screening accuracy, the identification of individuals who are likely to fail the outcome will vary. The following research questions were tested in the current investigation:

1) To what extent do indices of sensitivity, specificity, positive and negative predictive power, and the overall rate of correct classification change when using a screener to predict failure on two gold standard measures of reading comprehension (i.e., SAT-10 and GMRT)?
2) To what extent do indices of sensitivity, specificity, positive and negative predictive power, and the overall rate of correct classification change when manipulating cut-points to achieve .80 sensitivity (Method 1) or positive predictive power (Method 2)when predicting failure on one gold standard measure of reading comprehension (i.e., SAT-10)?

Method

Participants and Data Source

The participants were 17,778 second grade students who attended a Reading First school during the 2005-2006 school year. According to school records, this cohort of participants reflected the diversity found in Florida: 50% were female, 40% were identified as White, 32% as Black, 22% as Latino, 4% as Multiracial, and <1% as either Native American or Asian. Across the sample, 71% were eligible for free or reduced-price lunch, and 11% were served on an Individual Education Plan for a disability. Programs for limited English proficiency served 15% of students. A summary of the student demographics and the demographics for all students in the state are provided in Table 3.

Measures

Oral reading fluency. DIBELS ORF (Good, Kaminski, Smith, Laimon, & Dill, 2001) is a measure that assesses oral reading rate in grade-level connected text. Students are asked to read three passages out loud consecutively, for 1 minute per passage, and are given the prompt to "be sure to do your best reading" (Good et al., 2001, p. 30). Words omitted, substituted, and hesitations of more than 3 seconds are scored as errors, although errors that are self-corrected within 3 seconds are scored as correct. Errors are noted by the assessor, and the score produced is the number of words correctly read per minute. The median score of the three passages is the score type used for decision making about level of risk and level of intervention needed. Information about how the risk levels for ORF benchmarks were developed and what ranges of scores correspond to various levels of risk are available from several technical reports by the DIBELS authors (e.g., Good et al., 2002). Speece and Case (2001) reported parallel form reliability of .94, and strong

interrater reliability (.96) has been observed in Florida (Progress Monitoring and Reporting Network, 2005). Research has demonstrated adequate to strong predictive validity of DIBELS ORF for reading comprehension outcomes (.65 to .80; Barger, 2003; Good et al., 2001; Roehrig et al., 2008; Shapiro, Solari, & Petscher, 2008; Wilson, 2005). Part of the guiding principles Good et al. (2002) utilized in developing the original cut scores was to retain intervals for *low risk* levels that resulted in at least 80% of students meeting the end of year goal. Additionally, they wanted to set an interval for *high risk* whereby 20% or fewer of students met the third grade goal. Good et al. also outlined that the *some risk* students should have a 50% probability of meeting the end of year goal. Data for the current study consisted of the number of words read correctly per minute.

SAT-10. The SAT-10 is a group-administered, untimed, standardized measure of reading comprehension. Students answer a total of 54 multiple-choice items that assess their initial understanding, interpretation, critical analysis, and awareness and usage of various reading strategies. The internal consistency for the SAT-10 on a nationally representative sample of students was .88. Validity was established with other standardized assessments of reading comprehension, providing strong evidence of content, criterion, and construct validity (coefficients > .70; Harcourt Brace, 2004). For the present analyses, the percentile rank associated with the scale score on the total reading comprehension domain was used.

GMRT. The Reading Comprehension subtest of the GMRT consists of 40 single and short three to four sentence passages of narrative and expository text, followed by several multiple choice questions. The questions are purported to tap understanding of

details and ability to make inferences and integrate information in the passages. Internal consistency estimates of .96 and test-restest reliability of .85 to .90 were reported for the 2006 standardization sample (MacGinitie & MacGinitie, 2006), with construct validity estimates of .79 to .81 also reported. For the present analyses, the percentile rank associated with the scale score on the reading comprehension subtest was used.

Procedures

Data used in this study were drawn from the Progress Monitoring and Reporting Network (PMRN), an archival data source that houses student performance data on reading measures. Both the data and students used in this study were obtained from the PMRN, which is maintained by the Florida Center for Reading Research as part of its role in providing support for schools and districts throughout the state. The PMRN is a centralized data collection and reporting system through which schools in Florida report reading data and receive reports of the data for instructional decision making. The participants were administered the DIBELS assessments according to the state of Florida's assessment plan in the fall, winter, and spring. In the present study, we use data from the fall assessment period. The SAT-10 and GMRT were administered at the end of the school year.

Data Analysis

To answer our first research question, the screening accuracy of the screen at the fall was tested by creating a series of 2 x 2 contingency tables, similar to those presented in Table 1, which describe the number of students who were identified as at-risk or not at-risk on the screen (i.e., DIBELS ORF) and the two gold standard outcome variables (i.e., SAT-10 and GMRT). Scores on the measures were recoded into dichotomous

variables according to the cut-points on each measure which corresponded to risk. Across the measures, scores at or above the respective cut-points for "low risk" were coded as "1" (i.e., success), and scores corresponding to either "moderate" or "high" risk were coded as "0" to indicate that students did not meet the threshold for low risk.

At the fall assessment period, the University of Oregon Center on Teaching and Learning (2006) reported that ORF scores at or above 44 are considered to be low risk. Students with scores less than 44 may be identified as either moderate or high risk. Pertaining to the outcome tests, performance at or above the 40th percentile is often used to denote students who are low risk on state achievement tests (American Institutes for Research, 2007), while scores below this value are reflective of moderate or high risk performance. No universally agreed upon threshold exists for risk designation in education sciences, thus, it is important to consider commonly used practices for such score transformations. Although the 40th percentile cut-point is often utilized for state achievement outcomes (American Institutes for Research, 2007), we opted to use a more conservative value of the 50th percentile to identify students as at-risk on the SAT-10 and GMRT for illustrative purposes.

Using the reported cut-points for DIBELS ORF, the sensitivity, specificity, positive and negative predictive power, and the overall percentage of correctly classified students were calculated. Although other aspects of diagnostic efficiency may be tested (e.g., likelihood ratio, odds ratio), these five indices are more commonly found in technical reports and research papers to describe classification accuracy (Streiner, 1993).

The second research question was addressed by using receiver-operating characteristic (ROC) curve analysis in order to determine the cut-points of DIBELS ORF

which corresponded to a maximized screener property. Although several methods exist to evaluate the appropriateness of developed cut-scores (e.g., equipercentile equating and discriminant analysis), ROC curve analysis has been demonstrated as having greater flexibility with regard to estimated screening accuracy and determining the balance between Type I and II errors (Silberglitt & Hintze, 2005). Optimal cut-scores for differentially maximizing sensitivity and positive predictive power in the sample was determined by an examination of the values in the ROC curve, and subsequently using selected values in a 2 x 2 contingency table to evaluate the indices previously described. Using previously discussed recommendations for index thresholds, we sought to use two methods to establish cut-points which achieved either .80 sensitivity (Method 1) or .80 positive predictive power (Method 2).

Results

Missing Data Analysis

According to the descriptive data analyses 6% unique data points were missing across all studied variables and time points. The demographic makeup of students with missing data was examined to determine if the data were missing at random, or if demographics constituted a systematic error. However, frequency distributions suggested that the data were not missing in any discernable pattern (Table 3). Moreover, students with missing data approximated the students with complete data with regard to demographic frequencies. Although the prevalence of missingness was low, Little's test of data missing completely at random (Little, 1988) indicated that the data were not missing completely at random χ^2 (4) = 54.11, $p < .001$. In order to correct for an unbalanced design and potential biases in parameter estimation, multiple imputation was

conducted in SAS PROC MI analysis, with the free or reduced price lunch, minority status, and item scores variables using Markov Chain Monte Carlo estimation with 10 imputations.

Descriptive and Correlation Data

The descriptive statistics for students' performance on the selected measures are reported in Table 4. On average, students correctly read 56 words correct per minute (wcpm; *SD* = 31.66), and had developmental scale scores of 599 on the SAT-10 and 442 on the GMRT. A better contextualization of the developmental scale scores is to provide the associated percentile rank of each score; a score of 599 on the SAT-10 corresponded to the 50th percentile according to its norming sample, while a score of 442 on the GMRT was associated with performance at the 48th percentile of its norming sample. Moderate to strong correlations were observed across all measures, ranging from .64 between ORF and SAT-10 to .73 between SAT-10 and GMRT. Forty-four percent of students were identified as failing the SAT-10, compared with 56% on the GMRT.

Research Question 1: Screening Accuracy with Varying Outcomes

Using the identified scores to define risk on the ORF, SAT-10, and GMRT measures, 2 x 2 contingency tables were constructed in order to evaluate the screening accuracy of the screens. The results from calculations are reported in Table 5. As can be observed from the indices in the fall, when ORF was used to predict to two different gold standard reading comprehension outcomes, differential classification occurred relative to risk identification. When ORF predicted failure on the SAT-10, 66% of the students who failed the SAT-10 (i.e., scored <50th%ile) scored below 44 wcpm on DIBELS ORF, compared with a 60% correct classification of risk based on predicting failure on the

GMRT. This six percent observed difference in sensitivity in favor of the SAT-10 was counter-balanced by a six percent difference in specificity for the GMRT. That is, although 81% of students who passed the SAT-10 were fluent at or above 44 WCPM, an identification rate of 87% was estimated for the GMRT. A similar pattern of differential advantages for estimates between the outcomes was observed for positive and negative predictive power. Although the negative predictive power of the ORF-SAT-10 (75%) relationship was greater than the ORF-GMRT relationship (63%), a similar 12% discrepancy was estimated with positive predictive power in favor of the GMRT (86%) compared to the SAT-10 (74%).

Research Question 2: Screening Accuracy with Varying Cut-Scores

Recalibrated cut-scores for DIBELS ORF predicting failure on the SAT-10 were conducted. The resulting ROC curve suggested that an ORF cut-score of 48 would be appropriate to achieve a sensitivity value of .80 (Method 1). Similarly, in order to attain a positive predictive power estimate of .80 (Method 2), a fluency cut-score of 36 was needed. By using the respective points to maximize each screening goal, the resulting accuracy indices were calculated and reported in Table 5. Comparisons between the two methods for each index demonstrate the nature of discrepancies which occur when focusing on specific screen targets. The sensitivity was 28% higher in Method 1, but lower by 26% in specificity and 16% in positive predictive power, while higher by 11% in negative predictive power when compared to Method 2. Thus, Method 1 produces higher levels of sensitivity and negative predictive power while Method 2 results in higher positive predictive power and specificity.

Discussion

With RTI frequently required as a way to identify children at risk for future reading difficulties, schools are expected to implement RTI. One critical element of effective RTI placement is assessment, including screening, to identify students' instructional needs. Thus, it is important that educators understand trade-offs for choosing different gold outcomes and the statistical properties to maximize in order to meet the needs of districts and schools. In the present study, a widely used screen, DIBELS ORF, was examined for screening accuracy when predicting to two standardized measures of reading comprehension. In particular, we focused on delineating population-based indices vs. sample-based indices of screening accuracy. Sensitivity and specificity are examples of frequently reported population-based indices. In contrast, positive predictive power and negative predictive power are sample-based indices because these are influenced by students' performance level in the sample. As schools and/or districts differ in their demographic composition and students' performance levels, these sample-based indices are likely to provide relevant and useful information if schools and districts were to adopt a screen.

At the heart of the decision an educator must contend with is the question: Which is the greater perceived evil, to identify too many students for Tier II or Tier III intervention, or to miss students who are in need of services? The answer to this problem is not as simple as it seemingly appears, as multiple elements factor into the decision. The amount of funding a particular school or district may have for interventions could preclude a specific desire to maximize the identification of at-risk students. The key that educators should consider in light of resource allocations and priorities is weighing trade-

offs between providing intervention for those who do not need it and providing no intervention for those who do need it.

The results of the present study showed that DIBELS ORF measures had varying levels of diagnostic accuracy depending on the outcome. ORF tended to have higher sensitivity and negative predictive power for SAT10 than for GMRT. In contrast, ORF had higher specificity and positive predictive power for GMRT than for SAT10. These results suggest that the same screening measure serves somewhat different functions in terms of diagnostic accuracy depending on gold standard outcome. Although both SAT10 and GMRT are purported to measure the same construct (i.e., reading comprehension), and it is assumed that a gold standard is free from error, this notion primary stems from the origins of screening analyses which are derived from signal detection theory and medical models, where outcomes tend to be more dichotomous (i.e., the patient has cancer or does not). Thus, in education sciences, this assumption is less tenable in education where all instruments are flawed by random and/or systematic measurement error. These results imply that the choice of gold standard outcomes is an important consideration for schools and districts, especially when a differential proportion of students in the same sample failed the SAT-10 (44%) compared to the GMRT (56%).

When attempting to achieve a criterion of either positive predictive power or sensitivity for a given reading comprehension outcome (SAT10 in this study), focusing on positive predictive power (i.e., Method 2) appears to render a more favorable balance. In other words, when aiming for .80 for positive predictive power, the loss in sensitivity is comparable to the loss of specificity when sensitivity is achieved at .80 (Method 1). However, Method 2 does not lose as much on negative predictive power (9%) compared

to the loss in specificity (26%) and positive predictive power (16%) when using Method 1. There is strong evidence that students who start at a low level in reading rarely catch up in later grades (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Juel, 1988; Torgesen & Burgess, 1998) and remediating students later (e.g., second grade) takes much more time and resources, but is less successful (Foorman, Breier, & Fletcher, 2003; Torgesen, 2000). Thus, ensuring that fewer children are misidentified as not at risk and those identified as at risk receive intervention corresponds to the goal and the axiom of the RTI framework of allocating resources for early identification and prevention.

The findings from the present study should be interpreted with limitations pertaining to the context of the present study – i.e., the specific sample characteristics, the gold-standard outcome chosen, and ongoing interventions. Our sample was slightly over-representative of White students (40% sample, 30% state), and underestimated the proportion of Black students (31% sample, 38% state). Finally, schools in Florida are required to provide appropriate interventions to students based on their performance on the screening measures in the beginning of the year. However, the extent to which the students in this study were receiving Tier II or Tier III interventions, and how it impacted the results of the present study are unknown.

Implications for Research and Practice

Evaluating a screening assessment requires the educator's awareness about multiple factors pertaining to both the psychometric elements of the screen and practical needs of schools and districts. Schatschneider et al. (2008) provided initial guidelines about what to focus on in a screening process at the school or district level: 1) identify what "at-risk" means; 2) establish the goal for the screening process; 3) study how the

screen was developed; 4) determine the base rate in your school, district, or state; 5) attend to the positive and negative predictive power; and 6) collect local data to evaluate how well the screening process is working. The first critical step is defining what "at-risk" means (e.g., what outcome is used). Risk can take on a host of meanings and can describe performance on a concurrently administered standardized reading assessment, benchmark performance on a progress monitoring measure, passing an end of year state assessment test within the present year, or even success on the state test in a future grade. Without first delineating the type of risk to screen, the choice and utility of the assessment will fail.

Second, establishing the goal for the screening process is imperative as it will not only narrow down the list of potential screens, but will help determine the amount of time that could be spent assessing, identifying, and ultimately placing students into appropriate interventions. For example, if the goal is to choose an assessment that will identify the students who have a low chance for developing a problem from the screening process (i.e., reduce under-identification errors), then it is important to maximize negative predictive power. Conversely, if it is more important to have a high percentage of all students to be correctly identified as at risk and not at risk, then the eligible screening assessments should have a high percentage of the overall correct classification of students. For example, in the context of the present study, it is possible that by simply adjusting the cut-points for risk designation on the screen, such as Hintze, Ryan, and Stone (2003) and Roehrig et al. (2008) conducted in their respective studies, a different level of classification will be observed to meet the needs of the school or district. Once this goal is specifically outlined, evaluating a screen to measure that goal will assist in determining whether the definition of risk in the screen is the same as yours. Even if the outcome is

similar, it is important to check the specifications of that outcome and how skills on that measure were assessed. Many skills such as reading comprehension are complex and multidimensional, and can be assessed in ways that tap into lower-level or higher-level skills. Even the response format such as multiple-choice compared to a cloze or short answer will have an impact on the screening accuracy of scores (Jenkins, Johnson, & Hileman, 2004).

The fourth and fifth steps are tied together as the base rate of the problem in a school or district will effectively determine the extent to which a screen could reasonably be applied. If the screen selected was normed on a sample with a similar base rate, then it may be used with little apprehension as it provides information about how it will likely work in the selected sample. Lastly, collecting local data within the school or district will provide the best gauge as to how the screening process is working, and to what extent a different definition of risk, goal, or choice of screen is warranted. This will ensure the fit between selected screening measure and the needs of local schools and district. In summary, this research highlights that screening measures vary in psychometric properties of screening accuracy that are maximized, and are not likely to meet the needs and priorities of *all* schools and districts. Thus, it is researchers' and practitioners' responsibility to be aware of these characteristics and utilize screening measures as they were intended in order to best serve students.

References

American Institutes for Research (2007). *Reading First State APR Data*. Author.

Andreassen, R., & Braten, I. (2010). Examining the prediction of reading comprehension on different multiple-choice tests. *Journal of Research in Reading, 33*, 263-283.

Barger, J. (2003). *Comparing the DIBELS oral reading fluency indicator and the North Carolina end of grade reading assessment.* Asheville, NC: North Carolina Teacher Academy.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96*, 31-42.

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.

Davis, F. B. (1944). Fundamental factors of comprehension of reading. *Psychometrika*, *9*, 185–197.

Foorman, B. R., Fletcher, J. M., & Francis, D. J. (2004). *Texas Primary Reading Inventory.* New York: McGraw-Hill.

Foorman, B.R., Breier, J.I., & Fletcher, J.M. (2003). Interventions aimed at improving reading success: An evidence-based approach. *Developmental Neuropsychology*, *24*, 613-639.

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

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

Good, R.H, Kaminski, R.A., Smith, S., Laimon, D., & Dill, S. (2001). *Dynamic Indicators of Basic Early Literacy Skills* (5th Ed.), University of Oregon, Eugene.

Good, R.H., Wallin, J., Simmons, D.C., Kame'euni, E.J., & Kaminski, R.A. (2002). *System-wide percentile ranks for DIBELS benchmark assessment* (Technical Report, No. 9), University of Oregon, Eugene, OR.

Good, R. H., Kaminski, R. A., Shinn, M., Bratten, J., Shinn, M., Laimon, L., Smith, S., & Flindt, N. (2004). Technical Adequacy and Decision Making Utility of DIBELS (Technical Report No. 7). Eugene, OR: University of Oregon.

Harcourt (2004). *Stanford achievement test: Technical data report (10th ed.).* Orlando, FL: Author.

Hintze, J.M., Ryan, A.L., & Stone, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing. *School Psychology Review, 32,* 541-556.

Invernizzi, M., Meier, J. D., Swank, L., & Juel, C. (1999). *Phonological Awareness Literacy Screening.* Charlottesville: University of Virginia.

Jenkins, J. R. (2003, December). *Candidate measures for screening at-risk students*. Paper presented at the NRCLD responsiveness-to-intervention symposium, Kansas City, MO. Retrieved April 3, 2006, from http://www.nrcld.org/symposium2003/jenkins/index.html

Jenkins, J. R., Johnson, E., & Hileman, J. (2004). When is reading also writing: Sources of individual differences on the new reading performance assessments. *Scientific Studies of Reading, 8*, 125-151.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for service delivery in an RTI framework: Candidate measures. *School Psychology Review, 36*, 582–599.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80,* 437-447.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*, 281-300.

MacGinitie, W., & MacGinitie, R. (2006). *Gates-MacGinitie Reading Tests* (4th Ed.), Iowa City, IA: Houghton Mifflin.

Mather, N., Hammill, D. D., Allen, E. A., & Roberts, R. (2004). *Test of Silent Word Reading Fluency*. Austin, TX: Pro-Ed.

Roehrig, A. D., Petscher, Y., Nettles, S.M., Hudson, R.F., & Torgesen, J.K. (2008). Not just speed reading: Accuracy of the DIBELS oral reading fluency measure for predicting high-stakes third grade reading comprehension outcomes. *Journal of School Psychology*, 46, 343-366.

Schatschneider, C., Petscher, Y., Williams, K. M. (2008). How to evaluate a screening process: The vocabulary of screening and what educators need to know. In L. Justice ,& C. Vukelich (Eds.), *Achieving excellence in preschool literacy instruction* (pp. 304-316). New York: Guilford Press.

Shapiro, E., Solari, E., & Petscher, Y. (2008). Use of an assessment of reading comprehension in addition to oral reading fluency on the state high stakes assessment for students in grades 3 through 5. *Journal on Learning and Individual Differences*, *18*, 316-328.

Silberglitt, B., & Hintze, J.M. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23,* 304-325.

Speece, D. L., Mills, C., Ritchey, K. D., & Hillman, E. (2003). Initial evidence that letter fluency tasks are valid indicators of early reading skill. *Journal of Special Education, 36*, 223-233.

Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment, 81,* 209-219.

Torgesen, J. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research and Practice, 15*, 55-64.

Torgesen, J. K., & Burgess, S. R. (1998). Consistency of reading-related phonological processes throughout early childhood: Evidence from longitudinal-correlational and instructional studies. In J. Metsala & L.Ehri (Eds.).*Word recognition in beginning reading.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Woodcock, R.W., & Johnson, M.B. (1990). *Woodcock-Johnson Psycho-educational Battery- Revised, Examiner's manual.* Chicago, IL: Riverside.

Table 1

Sample 2 x 2 Contingency Matrix

	Outcome	
Screen	Fail	Pass
At-Risk	A True Positive	B False Positive
Not At-Risk	C False Negative	D True Negative

Table 2

Base Rate Comparison for Sample-Based Indices

School A

	State Achievement Test		
Screen	Fail	Pass	Total
At-Risk	950	100	1,050
Not At-Risk	50	900	950
Total	1,000	1,000	2,000

School B

	State Achievement Test		
Screen	Fail	Pass	Total
At-Risk	285	170	455
Not At-Risk	15	1,530	1,545
Total	300	1,700	2,000

Table 3

Demographic Characteristics for the Full Sample (Sample), the Population (State), and Students with Missing Data (Missing).

Demographics	Sample	State	Missing
Girl	50%	52%	51%
White	40%	30%	38%
Black	31%	38%	30%
Latino	22%	26%	23%
Asian	1%	1%	1%
Multiracial	4%	4%	4%
Native American	<1%	<1%	<1%
FRL	77%	76%	76%
ELL	15%	17%	14%
Speech Impaired	5%	5%	5%
Language Impaired	2%	3%	2%
Specific Learning Disability	4%	5%	4%
Other	4%	4%	4%

Note. FRL = Free and/or reduced price lunch, ELL = English language learners.

Table 4

Descriptive Statistics and Correlations for Observed Variables

	Fall ORF	SAT-10	GMRT
Fall ORF	1.00		
SAT-10	.64	1.00	
GMRT	.68	.73	1.00
M	56.11	598.98	442.39
SD	31.66	39.48	38.51

Note. ORF = DIBELS Oral Reading Fluency, SAT-10 = Stanford Achievement Test, GMRT = Gates-MacGinitie Reading Test.

Table 5

Diagnostic Efficiency Results

Variable	Sensitivity	Specificity	PPP	NPP	OCC
Research Question #1					
ORF – SAT-10	66%	81%	74%	75%	74%
ORF – GMRT	60%	87%	86%	63%	72%
Research Question #2					
Method 1	80%	64%	64%	81%	72%
Method 2	52%	90%	80%	70%	73%

Note. ORF = Oral Reading Fluency, SAT-10 = Stanford Achievement Test, GMRT = Gates-MacGinitie Reading Test, PPP = Positive Predictive Power, NPP = Negative Predictive Power, OCC = Overall Correct Classification.