Published in final edited form as:

Learn Individ Differ. 2008 ; 18(3): 316–328. doi:10.1016/j.lindif.2008.03.002.

Use of a Measure of Reading Comprehension to Enhance Prediction on the State High Stakes Assessment

Edward S. Shapiro, Emily Solari, and Yaacov Petscher

Lehigh University, Center for Promoting Research to Practice; University of Texas Health Science Center Houston, Children's Learning Institute; Florida State University, Florida Center for Reading Research

Abstract

The current study examined the diagnostic accuracy of two screening measures of risk for future difficulties in reading comprehension, as well as the degree to which adding a screening measure of reading comprehension enhanced the prediction of Oral Reading Fluency to outcomes of student reading performance on the state high stakes assessment for grades 3 through 5. Data from fall and winter assessments of the DIBELS Oral Reading Fluency (DORF) and 4Sight Benchmark Assessment (4Sight) measures along with outcomes on the Pennsylvania System of School Assessment (PSSA) across a total of 1000 students from 6 schools were examined using indices of diagnostic efficiency, ROC curve, and logistic regression analyses. Results showed that the addition of a measure of reading comprehension (4Sight) to DORF enhanced the decision making process for identifying students at risk for reading difficulties, especially for those students at higher elementary grades and those who achieved benchmark levels on the DORF. Although DORF alone showed a good level of prediction to the statewide assessment, the combination of the DORF plus 4Sight measures resulted consistently in the best predictive outcomes. Suggestions are made to consider alternative cut points for the DORF and 4Sight measures.

Keywords

RTI; Reading comprehension; Oral reading fluency

Results of statewide assessment, often called high-stakes testing, have important implications for districts, teachers, and individual students that impact district level decision making and policy. Usually administered once during the school year, statewide assessments do not provide information about student growth over shorter time periods and provide even less information to guide effective instruction for students who are at-risk (Shephard, 2000). The value and potential impact placed on high-stakes testing has encouraged school districts to put into place more frequent monitoring of student academic outcomes during the school year to identify those students potentially at risk for not reaching proficiency levels prior to the administration of the high-stakes test. Such identification could allow districts to put in place programs to focus instruction for such students in areas that would improve student outcomes on the high-stakes test.

Response to Intervention (RTI) is becoming a more commonly adopted methodology for identifying and intervening with students at-risk for reading difficulties (Burns, Appleton, & Stehouwer, 2005; Gersten & Dimino, 2006; Haager, Klingner, & Vaughn, 2007), as well as serve as the process for identifying students with learning disabilities (Danielson, Doolittle, & Bradley, 2005; Speece, Case, & Molloy, 2003; Vaughn, Linan-Thompson, & Hickman, 2003). Most RTI models for identifying and intervening with students at risk for reading difficulties include three essential elements: a) universal screening of all students to determine which students are at-risk for reading failure, b) early intervention programs for students who are at high risk for reading failure as determined by universal screening, and c) frequent monitoring of students' progress. Assessing response to instruction or the effectiveness of the intervention requires implementing an effective treatment, measuring the student's response to the treatment, and creating a criteria at which children who score lower and move slower than a specified rate of progress are identified as non-responders to intervention (Fuchs, Fuchs, Mathes, Lipsey, & Roberts, 2002).

An essential component of RTI models is the use of valid and reliable data sources to enable decision making around student achievement. Oral Reading Fluency (ORF) is a commonly used measurement to assess students reading progress and predict later reading outcomes in grades two through six. Oral reading fluency is a multi-component process that develops over time and has been operationally defined as the oral production of text with both speed and accuracy (Adams, 1990; Berninger, Abbott, Billingsley, & Nagy, 2001) and has been commonly used as a skill to measure progress in reading for a few reasons. First, it has been demonstrated in the literature that ORF is an important predictor of later reading outcomes (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Shinn, 1998). Research has also shown that levels of oral reading fluency provide important information related to reading comprehension performance (Adams, 1990; National Reading Panel, 2000; Fuchs et al., 2001). These studies have been driven by an assumption that efficient decoding (i.e., reading fluently), frees up cognitive load and allows for more attention to be directed toward higher level processes such as comprehension of text (Fuchs et al., 2001).

Although research related to oral reading fluency and its relationship to reading comprehension suggest that reading fluency is an important predictor of reading comprehension for most students (Nathan & Stanovich, 1991; Thurlow & van den broek, 1997), limited research has been conducted to determine if this is true for students across different grade levels. Because reading fluency is a multi-component process that taps a variety of cognitive and language processes such as phonological and orthographic processes, morphology, semantic and syntactic relationships and efficient access to lexicon (Adams, 1990; Berninger et al., 2001; Kame'enui & Simmons, 2001), individual differences in all the subcomponent skills cause varied reading fluency rates across students. Given that the developmental paths of components of reading fluency are differential across individuals, the relationship between oral reading fluency and comprehension may change as students develop these subcomponents (Wood, 2006).

The second reason reading fluency rates have been used as an effective progress monitoring tool is because oral reading fluency has been shown to be sensitive to student growth over short periods of time (Fuchs, Tindal, & Deno, 1981; Marston & Magnusson, 1985; Tindal &

Deno, 1981). Assessment of growth over a short period of time is an essential component of a tiered model of reading instruction. When fluency measurement tools are utilized, reading instruction can adapt to student response to intervention, allowing for more efficient use of instructional time and school resources.

The third reason reading fluency has been commonly used as a risk indicator for reading failure is the ease with which ORF can be measured. In general, ORF is assessed through curriculum based measures and scored by the number of words read correctly in 1 min. The short length of assessment makes ORF an easy measurement tool to gauge student performance on reading quickly and efficiently. Many school districts have set benchmark goals for ORF to help make instructional decisions for students who may be at-risk for reading failure and failure on standardized assessments (Good, Simmons, & Kame'enui, 2001; McGlinchey & Hixson, 2004; Stage & Jacobsen, 2001).

Relationship between oral reading fluency and state standardized reading assessments

The effective use of oral reading fluency scores as predictors of reading proficiency and achievement requires the establishment of concurrent and predictive validity with statewide reading assessments (Wood, 2006). A number of studies have investigated the correlational relationship between ORF and state standardized assessments (Buck & Torgesen, 2003; Crawford, Tindal, & Stieber, 2001; Good et al., 2001; McGlinchey & Hixson, 2004; Schilling, Carlisle, Scott, & Zeng, 2007; Shapiro, Keller, Lutz, Edwards, & Hintze, 2006; Roehrig, Petscher, Nettles, Hudson, & Torgesen, in press; Stage & Jacobsen, 2001; Wood, 2006). For example, Crawford et al. (2001) examined the relationship between second and third grade fluency rates and third grade state-wide assessments. In their sample, there was a correlation of .60 in second grade and .66 in third grade between ORF and third-grade reading scores on the end of year statewide assessment. Good et al. (2001) reported a .67 correlation between a spring measurement of ORF and the statewide reading assessment in Oregon. Fuchs et al. (2001) found a correlation of .80 between ORF and reading comprehension portion of the statewide assessment in Iowa. In Michigan, a study using several cohorts of students reported correlations from .49 and .81 between ORF and the reading portion of the statewide assessment (McGlinchey & Hixson, 2004). In a study investigating the relationship between ORF and Washington statewide reading assessment, Stage and Jacobsen (2001) reported a .44 correlation between ORF and the reading section of the standardized state assessment. In general, correlational studies have found moderate to strong correlations between oral reading fluency rates and state standardized reading assessments.

Wood (2006) used hierarchical linear modeling to investigate the relationship between ORF and the Colorado statewide assessments. Classroom-level and grade-level variation were examined to determine if ORF predicted performance on the statewide testing above and beyond scores on the previous year's state assessment. Similar to previous studies, Wood reported moderate to strong correlational relationships between ORF and the statewide assessments ($r=.70$ for grade three, $r=.67$ for grade four, and $r=.75$ for grade five). Additionally, cutoff scores were determined using standard error of students who scored

proficient on the Colorado reading assessment with the lower 95% confidence interval used as the cutoff point for passing the ORF measure. When determining cutoff scores in this way, Wood found high levels of sensitivity (i.e., those predicted to pass and actually did pass) across grade levels (between .86 and .95) with lower levels of specificity (i.e., those predicted to fail and actually did fail) (between .58 to .67). The prediction of which students would pass or fail the Colorado assessments was significantly above the base rate, indicating that the ORF measure did a reasonably good job in predicting student outcomes on the end of year exam across third, fourth and fifth grade, however, the level of diagnostic efficiency above chance increased between grades three and five, from 19% for grade three, 33% for grade four and 41% for grade five.

Keller and Shapiro (2005) reported longitudinal correlations between ORF and Pennsylvania state reading assessment in second grade between .69 and .71, and fourth grade between .67 and .69 when there was 1 year between the ORF assessment and the state reading achievement test. When there were 2 years between the measurement of ORF and the state achievement test, correlational levels remained consistent, between .72 and .74 for third grade students. Additionally, ROC curve analyses were used to determine optimal cut scores for diagnostic accuracy for pass/fail rates on the Pennsylvania achievement tests. This analysis demonstrated that when using the spring ORF scores, standardized reading scores could be predicted correctly for approximately 75% of the students. Likewise, Roehrig et al. (in press) examined the predictability of DIBELS ORF to the Florida Comprehensive Assessment Test for third grade students. Compared to when the recommended DIBELS cut scores were used for identifying risk levels, sensitivity and specificity of the DIBELS ORF could be substantially improved using a recalibration of cut scores.

Although several studies have demonstrated moderate to high correlations between ORF and statewide achievement tests, many questions still remain related to diagnostic efficacy of using ORF to predict state standardized assessments. The relationship between ORF and state standardized tests at multiple time points in an effort to determine the most efficient use of oral reading fluency has been understudied. Determining cut-points that best predict performance on statewide assessments at multiple points throughout the year would be advantageous in a tiered model of instruction to guide instructional grouping and content.

Is ORF enough? Do we need something else to predict pass/fail rates?

Several questions still remain concerning the validity of using oral fluency rates as a predictor of standardized test scores and perhaps more importantly, the use of ORF as the main indicator of risk status for tiered instruction within an RTI model. Word reading fluency has been demonstrated to be a necessary, but not sufficient, foundation for development of skilled reading comprehension (Catts & Hogan, 2003). According to Duke, Pressley, and Hilden (2004) 20–25% of students who are identified with comprehension deficits also have word decoding problems. However, there is another group of students who never have decoding and fluency problems but demonstrate comprehension deficits later in elementary school (Catts & Hogan, 2003). The Simple View of Reading (Gough & Tunmer, 1986) can potentially explain performance of students who are able to decode text efficiently, but still have comprehension difficulties. The Simple View of Reading argues

that both linguistic comprehension and decoding are essential for successful reading comprehension. Students who are actively reading text are simultaneously engaging in linguistic comprehension, therefore if an individual has weak linguistic comprehension skills, they would in turn have poor comprehension performance. This low linguistic comprehension group of students may not be identified as at-risk for reading comprehension failure if progress is monitored solely with reading fluency assessments. Although policy and practice seem to be moving toward a RTI model of identification and early prevention for reading disabilities, an important challenge to the field is to establish the measures that should be used to determine responsiveness and what skills are necessary for successful reading comprehension beyond reading fluency.

In general, previous studies have concentrated on whole grade analyses and have not accounted for individual differences in student's reading fluency rates that may affect predictability outcomes of ORF on reading comprehension assessments and state reading assessments. Comparisons of high and low fluency readers to determine if reading fluency equally predicts performance on comprehension assessments or state-wide standardized assessments for all students have been minimal. Given the high level of comprehension skill needed to successfully complete and pass statewide assessments, questions remain related to assessing higher level reading components (i.e., comprehension) throughout the year and using these data in conjunction with ORF to predict outcomes on state reading assessments.

A second area that has been understudied is the predictive validity of ORF in students who are older than third grade. In general, a qualitative change in reading development and academic expectancy takes place in third grade; instruction moves away from lower level processing (decoding and fluency) and concentrates on higher level skills (comprehension) that are required to successfully pass statewide reading assessments. Previous research has suggested that growth in oral reading fluency slows as children reach intermediate grades (Fuchs, Fuchs, Hamlett, Waltz, & Germann, 1993). Fuchs et al. (2001) suggested that the relationship between oral reading fluency and comprehension is stronger when students are younger and as reading ability moves to more complex processes such as text analysis and comprehension, the relationship weakens. Shinn, Good, Knutson, Tilly and Collins (1992) likewise found that in comparing a confirmatory factor analysis of reading performance of students at third and fifth grade, that reading at third grade was best explained by a unitary factor combining fluency and comprehension. However, reading at fifth grade required a two-factor solution with both fluency and comprehension independently needed to predict outcomes in reading performance. In a cross-sectional sample, Wood (2006) found a significant difference between reading fluency rates of third, fourth and fifth grade students, suggesting that students in this age range are still developing reading fluency skills. From a developmental standpoint, it makes sense that normal achieving students will reach a peak performance in reading fluency rate, at which time more processing abilities can concrete on comprehending text. Few studies, however, have been conducted to determine the influence of the shift in processing ability on the predictive validity of oral reading fluency on standardized reading assessments.

In an effort to improve prediction of student outcomes on statewide high stakes reading assessments and to provide early identification of students with reading comprehension

difficulties, the Success for All Foundation has developed measures specifically linked to state assessments that can be given repeatedly across the school year for students from grade 3 to 11 (Successforall.org). Entitled 4Sight Benchmark Assessment, these measures are customized to link to the standards and anchors for each state for which the measures are developed. In particular, the 4Sight measures focus on aspects of reading comprehension and are formed to mimic the statewide assessment. The measures are designed to be repeated up to five times (equivalent alternate forms) during the course of a school year, with one or two measures that follow baseline assessments typically given prior to the statewide assessment. Outcomes of the 4Sight measures are highly correlated to state assessments, allowing schools to potentially identify and target students who may be at risk for not reaching proficiency on the state assessment. As such, the measures can potentially be used to enhance the decision making process within an RTI model for reading where ORF may be the primary measure being used for such decisions. Over 20 states have adopted the 4Sight Benchmark Assessments as a mechanism to provide early prediction of outcomes on their respective state assessments (<http://www.successforall.net/ayp/benchmarks.htm>).

The 4Sight Benchmark Assessments are different in some respects from more commonly used standardized measures of reading comprehension in that the test was built specifically to resemble the style, content, and response requirements of the state reading assessment. However, like other standardized reading measures, the test assesses subskills commonly found in most reading measures.

Purpose of study/research questions

The purpose of the current study will be to add to the literature on the relationship between ORF, reading comprehension and a statewide standardized reading achievement test, the Pennsylvania System of School Assessment (PSSA) across one school year in students in third, fourth, and fifth grade. The study will examine the predictive value of ORF at two time points (fall and winter) prior to the administration of the reading section of the PSSA (administered during late winter). An additional comprehension based assessment, 4Sight Benchmark Assessment (assessed in both fall and winter) will be used to determine if there is value added by using both ORF and 4Sight to predict student performance on the PSSA.

In this study we address two research questions. First, what is the relationship between ORF, 4Sight at two time points during the school year (fall and winter) and PSSA at third, fourth, and fifth grade? Although previous studies have investigated the relationship between ORF and state achievement tests (Crawford et al., 2001; Fuchs et al., 2001; Good et al., 2001; McGlinchey & Hixson, 2004; Wood, 2006) our study adds a comprehension based assessment used to monitor student progress on Pennsylvania state reading standards as part of the predictive decision model.

Second, what is the predictive accuracy of ORF and 4Sight on reading PSSA outcomes for students who are considered at risk for reading difficulties at two different time points (fall and winter)? We want to determine if the relationship between ORF, 4Sight, and PSSA differs for students who are considered at-risk for reading difficulties as determined by their

oral reading fluency rate. Research has shown that students with lower rates of oral reading fluency tend to perform worse on comprehension based assessments and in particular state reading achievement tests (Wood, 2006). However, studies that specifically investigate students who are identified as below oral reading fluency benchmarks and considered some risk and at-risk are minimal. Specifically, we aimed to determine if the addition of a comprehension based assessment (4Sight) to a prediction model, in conjunction with ORF will increase diagnostic accuracy for at-risk students and more accurately identify students who are at-risk for failing the Pennsylvania reading achievement test.

METHOD

Participants

Data from a total of 1000 students in grades 3 ($n=401$), 4 ($n=394$) and 5 ($n=205$) were obtained for this study. The students came from a total of 6 elementary schools in Pennsylvania across a total of 3 districts (4 schools were from one district). The mean level of free and reduced lunch across schools was 38.9% (range=22.2% to 61.9%). The four schools from the same district were fairly similar in the average ethnic makeup (White=48.0%, Black=34.2%, and Hispanic=11.7%). One of the other two schools had 16.5% Black, 8.9% Hispanic, and 73.8% White while the other school was 98.6% White. The four schools from the same district contained grades 3, 4, 5 while the two schools from the other districts contained grades 3 and 4 only.

Measures

Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency Passages (DORF)—The fall and winter benchmark reading probes for grades 3, 4, and 5 developed by DIBELS (Kaminski & Good, 1998) were used to assess Oral Reading Fluency. Students were asked to read each passage aloud for 1 min. A total of 3 passages were administered and the median score for students across those passages was recorded as the dependent measure to reflect reading performance in words read correct per minute. Typically, the DORF measures are administered three times per year (fall, winter, spring). Given that the statewide assessment (described below) was administered in March, we only examined the outcomes of fall and winter assessments in relationship to the statewide assessment.

The reliability and validity of CBM reading passages has been established for over two decades (c.f., Good, Simmons, & Kame'enui, 2001; Good & Jefferson, 1998; Marston, 1989; Shinn, 1989; Wayman et al., 2007). Outcomes of student performance on the DORF are assigned to categories of low risk (at or above benchmark), some risk, or at risk. Specific scores related to each level are provided on the DIBELS website (<http://dibels.uoregon.edu/benchmarkgoals.pdf>).

4Sight Benchmark Assessment (successforall.net)—This reading measure was a one hour, group administered test consisting of 30 items, similar in format to the statewide assessment, including multiple choice items and items scored as open-ended. The test used in this study was specifically developed based on two major standards and five Pennsylvania

assessment anchors related to reading performance with particular emphasis on reading comprehension and was designed to be predictive of outcomes on the statewide assessment. The skills assessed as part of the measure were: Comprehension and Reading Skills, Interpretation and Analysis of Fictional and Non-Fictional Text, Learning to Read Independently, Reading Critically in All Content Areas, and Reading, Analyzing, and Interpreting Literature.

Over the course of the year, the measure was designed to be given multiple times (potential of 5 administrations: fall, mid-fall, winter, late winter, spring), with the initial administration at the beginning of the year representing the baseline score for the student. Two equivalent forms of the test were available with form A designated as the fall, winter, and spring measure and form B to be given at mid-fall and late winter. For this study, data from form A administered at baseline (fall) and winter was used. As per the instructions for using the 4Sight Benchmark Assessments, the winter administration of form A consisted of the identical items as the baseline test but placed in a different random order. Students were assigned a total raw reading score on the test that also generated a predicted scaled score on the state assessment corresponding to the categories of the Pennsylvania System of School Assessment (PSSA) statewide assessment (Advanced, Proficient, Basic, Below Basic). The total raw score attained on the test as well as the assigned category was used in this study to reflect student outcomes.

Reliability for the 2006–2007 year of the 4Sight measure for Pennsylvania was reported based on the inter-form correlations as ranging from .69 to .74 across grades 3, 4 and 5 based on samples of 21,500, 21,700, and 22,500 students across the state of Pennsylvania. Concurrent validity for the form used in this study with the PSSA for reading in the year 2006–2007 was reported to be .78 for grades 3 and 4, and .79 for grade 5 (Success for All Foundation, 2007).

Pennsylvania System of School Assessment (PSSA)—The PSSA is the measure designed for educational accountability purposes in Pennsylvania (DRC, 2007a,b). The reading portion of the PSSA covered two broad skill areas that were based on the Pennsylvania Assessment standards: (1) Comprehension and Reading Skills and (2) Analysis of Fiction and Nonfiction text (CTB McGraw-Hill, 2006; DRC, 2007a,b). The PSSA generated a standard score that classified student performance into four levels: Advanced, Proficient, Basic, Below Basic. The student's attained score as well as their classification category was used in this study as a measure of student outcome.

The technical information available about the PSSA reading indicated that all of the reading forms of the PSSA had high reliability coefficients (Cronbach's alpha) being .90 for all three grades. A modified bookmark procedure was used to establish the performance cut points for PSSA (DRC, 2007a,b). The performance levels set for the third grade reading PSSA test were as follows; Advanced/Proficient=1180 or above, Basic=1044 to 1179; Below Basic=b1044. At fourth grade, performance levels were Advanced/Proficient=1246 or above, Basic=1156 to 1245; Below Basic=b1156. At fifth grade, performance levels were Advanced/Proficient=1312 or above, Basic=1158 to 1311; Below Basic=b1158 (DRC, 2007a,b). Successful performance on the PSSA is considered as Proficient or above.

The reading PSSA included both multiple choice and performance tasks as well as open ended tasks. Extensive evaluation of content validity, construct validity, item fit and calibration were described in the technical analysis manuals and showed the PSSA to have strong psychometric characteristics consistent with statewide assessments (DRC, 2007a,b).

Procedure

During the second and third weeks of September (fall) and again during the second week of January (winter), the DORF assessment was administered in all schools. The 4Sight Benchmark assessment was administered during the last week of September (fall) and again during the second week of December (winter). DORF was administered to individual students by teams of education professionals who had received at least 2 h of training on the administration and scoring of the Oral Reading Fluency measures. The teams had been administering these measures for the past several years. The 4Sight measures were administered by teachers and reading specialists during a period of language arts. All individuals administering these measures also attended a training session provided by district personnel who had previously attended a series of professional development workshops on the administration and interpretation process of the 4Sight measures. Consistent with state assigned dates for PSSA assessment, the PSSA was administered during the second week of March (late winter).

Data analysis

Diagnostic validity—The validity of the scores on the DORF and 4sight measures in the prediction of PSSA risk was tested by generating receiver-operating characteristic (ROC) curves. ROC curves were used to examine the probability that a student is correctly classified as at some level of risk on both PSSA and the screening measure (i.e., sensitivity), and the probability that a student correctly classified as not at some level of risk on the PSSA and screening measure (i.e., specificity). Though a variety of methodologies exist for evaluating the relationship between a screening measure and a test (e.g., discriminant analysis, equipercentile equating, logistic regression), ROC curve analysis has been shown to provide greater flexibility with regard to estimated diagnostic accuracy and predictive power (Hintze & Silbergitt, 2005), as well as in determining the balance between Type I and Type II errors. In conjunction with the ROC curves, we calculated conditional probability indices in order to provide multiple perspectives on the efficiency of the screening scores: sensitivity (SE), specificity (SP), positive predictive power (PPP), negative predictive power (NPP), and total predictive value (TPV). The area under the curve (AUC) was also generated as part of the ROC analysis, and is generally considered to be sufficient as an effect size indicator (Swets, 1988).

Positive predictive power (PPP) is the probability that a student who is identified as being at risk according to DORF or 4Sight scores is truly at risk. Similarly, Negative predictive power (NPP) is the chances that a student who is identified as being not at risk according to DORF or 4Sight scores is truly not at risk. The Total predictive value (TPV) reflects a combination of the SE and SP, and gives the probability of correctly identifying students for both at-risk and not at-risk test results. Lastly, the AUC is a probability index ranging from 0.5 (no better than chance) to 1.0 (perfect prediction), and provides the probability of the

independent variable correctly classifying a pair of individuals, where one student is at risk and the other is not. Positive and negative predictive power is limited in their generalizability as they are affected by the base rate of the presented condition in the sample. These six indices were selected for reporting as they represent the most commonly reported efficiency statistics (Streiner, 2003), and all but the positive and negative predictive power are not affected by base rates, which is an important property when examining diagnostic accuracy in a high or low prevalence population (Meehl & Rosen, 1955; Streiner, 2003).

Diagnostic validity analyses were conducted for the fall and winter administrations of the screening tests (DORF and 4Sight) in 3rd, 4th, and 5th grades. Since the spring assessment of both DORF and 4Sight occurred after the PSSA was given, it would have been inappropriate to make a prediction about the PSSA from a screen administered after the state achievement test. For purposes of this analysis, risk categories were dichotomized into those at or above benchmark (low risk) and those at some or at-risk (at-risk). A third measure representing the additive relationship between the two screens at each time point was included in order to evaluate the value-added nature of 4Sight to DORF in predicting risk on the PSSA. In light of the fact that multiple ROC curves were generated for each time point (i.e., fall and winter) within grades, it was important to evaluate the dependency of the results on the sample (Hanley & McNeil, 1983). To this end, analyses for within-group AUC differences were conducted to measure the differences across screeners at each time point.

Differential probability analysis—Testing if the scores on DORF and 4Sight differentially predicted performance on the PSSA for students who were identified as at-risk or not at-risk on the screeners involved a series of logistic regression analyses. Odds ratios were reported to enhance interpretation and provide a communicable effect size. An important question was the degree to which the addition of a measure of reading comprehension such as 4Sight enhanced the predicted outcomes on PSSA as determined by ORF alone. To examine this question, a series of logistic regression analyses were conducted to test if the scores on ORF, 4Sight, and combination of the measures differentially predicted performance on the PSSA for students who were identified as at-risk or not at-risk on the screeners.

RESULTS

Reading comprehension and fluency

Descriptive statistics and correlations for students' performance on DORF, 4sight, and PSSA measures are reported in Table 1 for 3rd–5th grade samples. A very strong relationship was observed between fall and winter DORF across all grades (range=0.92–0.94), while a relatively weaker but strong correlation existed on the 4Sight measure between both times (range=0.72–0.80). Moderate correlations between DORF and 4Sight indicated that reasonable concurrent validity existed within each time period; similarly, the correlations between measures across-times (i.e., DORF fall to 4Sight winter, 4Sight fall to DORF winter) were generally of the same magnitude as the concurrent measure estimates.

Analysis of the relationship between DORF and 4Sight and the state wide achievement test (i.e., PSSA) indicated that 4Sight was more strongly correlated with PSSA than DORF across all grade levels. A correlation contrast comparison (Meng, Rosenthal, & Rubin, 1992) was used to test if the correlation between each screening measure and the PSSA significantly differed from each other, for each time point students were assessed. Results indicated that the relationship between 4Sight and PSSA was stronger than the association observed for DORF and PSSA during both fall ($z=2.91$) and winter ($z=3.78$) assessment in third grade. This trend was also seen for 4th grade students' scores in fall ($z=2.24$) and winter ($z=2.09$); however, no significant differences between correlations were seen in 5th grade.

Diagnostic accuracy

Using the current benchmark scores to define risk levels for DORF and proficient or above on 4Sight, as well as for the PSSA, 2×2 summary tables were constructed in order to calculate the sensitivity, specificity, predictive power, and total predictive value. The results from these summary tables are reported in Table 2. As can be seen from the indices, it appeared that for 3rd grade students both DORF and 4Sight were equally sensitive to the PSSA scores in fall (.96 and .96, respectively) and winter (.95 and .94, respectively). Specificity estimates showed that a stronger probability of correct classification of non-risk was observed for fall DORF and winter 4Sight. Due to the nature of the index calculations, the stronger specificity probabilities led to a higher level of PPP for the assessments. That is, high fall DORF specificity led to stronger PPP (.33) when compared to fall 4Sight PPP (.29). Similarly, the larger specificity for winter 4Sight resulted in a stronger PPP (.40) than winter DORF (.36). Both DORF and 4Sight in fall and winter were equally accurate in predicting who was likely to not to be at risk on the PSSA (NPP=.98), while the TPV was the strongest for DORF in the fall and 4Sight in the winter.

Fourth grade results showed a different pattern of predictive utility than observed for the 3rd graders. Sensitivity was substantially stronger for 4Sight in both fall and winter compared to DORF (Table 2), with a larger specificity probability for fall DORF and winter 4Sight. The PPP indicated more accurate predictions of risk were made based on fall DORF and winter 4Sight, while NPP provided greater precision for the 4Sight measure across both time points. Interestingly, while the TPV for DORF was stronger in the fall than 4Sight (.57 and .53), and stronger in the winter for 4Sight than DORF (.69 and .63), the consistency between 3rd and 4th grade TPV was higher for 4Sight. While 3rd grade fall DORF may have a higher TPV in the fall, this also represented a 6% decrease from 3rd to 4th grade, while only a 1% drop was observed for fall 4Sight. Similar findings were seen for the Winter measures whereby DORF TPV was lower by 3% and 4Sight by only 1%.

Diagnostic efficiency in 5th grade indicated that the 4Sight measure was a better screener for risk than DORF across both fall and winter administrations. The sensitivity and specificity was stronger for 4Sight in the Fall (1.00, .63) compared to DORF (.97, .58). Similarly, though correct classification of risk was better for winter DORF (.93) than 4Sight (.87), the specificity was greater for 4Sight than DORF (.74 and .61), as were the PPP and TPV estimates.

Another means of examining the process of decision making between Oral Reading Fluency and Comprehension as they relate to outcomes on the PSSA can be determined by looking at the cross-tabulation of students scoring at each of the risk categories defined by DORF, those who then scored as Advanced/Proficient on the 4Sight benchmark measure, and then similarly scoring Advanced/Proficient on the PSSA. Table 3 shows that across grades, those students who scored at or above benchmark on the winter DORF (Low Risk) who then scored Advanced/Proficient on the winter 4Sight, were found to have scored Advanced/Proficient between 88.1% and 96.5% of the time, with the strongest outcomes at grade 3. Those students scoring in the some risk category on the winter DORF but scored as Advanced/Proficient on the winter 4Sight were found to have scored at Advanced/Proficient on the PSSA between 60.9% and 87.1% across grades. In particular, at grade 3, only 12.9% of students scoring in the some risk category who scored Advanced/Proficient on 4Sight scored below Advanced/Proficient on the PSSA. Finally, examining those who scored At Risk on the DORF and Advanced/Proficient on the 4Sight were found to have been at Advanced/Proficient on the PSSA between 28.6% and 57.1% across grades.

ROC curve analyses

In order to model the diagnostic accuracy of DORF and 4Sight over the entire range of scores, Receiver Operating Characteristic (ROC) curves were generated to more fully examine the diagnostic accuracy of the screeners. In addition to modeling the individual measures at each time point, PSSA was regressed on a combined DORF/4Sight index, which represented the additive value of 4Sight to DORF. The evaluation of a ROC curve entails the examination of the AUC index (closer to 1=better correct classification), as well as the trade-off between the sensitivity (along the y-axis) and the false-positive (i.e., 1 specificity, x-axis) changes across the curve. A well discriminating measure will be most likely to show that progressive increases in sensitivity result in very small increases in false-positives (Tatano-Beck & Gable, 2001), and tend to be further away from the diagonal line (i.e., 50% chance of correct classification).

Fall DORF (AUC=.874) generally demonstrated better diagnostic accuracy than 4Sight (AUC=.866); however, the combined measure had a better classification rate than either of the two screeners individually (AUC=.885). Since the data from the curves came from the same sample of students, it was important to test if the curves were independent of each other (Hanley & McNeil, 1983). A z-test of the AUC was conducted using methodology outlined by Hanley and McNeil (1983). Results indicated that none of the curves were reliably differentiated from each other ($z=1.96$). Data from winter indicated that 4Sight (AUC=.914) was a better classification measure of PSSA risk than DORF (AUC=.870), and that the value-added measure (AUC=.894) provided improved classification when compared to DORF by itself. Similar to the 3rd grade Fall findings, none of the curves were significantly different from each other.

Results from the analysis of fourth grade fall data indicated that 4Sight (AUC=.828) provided better diagnostic accuracy than DORF (AUC=.785), with the value-added measure performing better than DORF alone (AUC=.801). Winter results showed that 4Sight

(AUC=.878) was statistically superior to DORF (AUC=.792) in risk classification ($z=2.81$), and the combined index (AUC=.818) performed better than DORF alone.

Results from the ROC curve analysis for fifth grade were dissimilar from the previous grades and found all three measures at both time points provided similar classification results. During the fall assessment, the combined measure performed the best (AUC=.933); however, both 4Sight (AUC=.920) and DORF (AUC=.918) were at comparable levels of classification. Similarly, during the winter assessment the combined measure (AUC=.931) had a higher predictive accuracy rate, though it was not statistically better than either the individual 4Sight (AUC=.906) or DORF (AUC=.916) measures.

Logistic regression

In order to ascertain the predictive probability of risk on the PSSA (i.e., PPP) based on levels of risk on DORF and 4Sight, the raw data was recoded into three categories consistent with DORF recommended benchmarks of risk: low risk, some risk, at-risk. The DORF data were analyzed by risk level in a logistic regression to identify both the predictive power of each screener's scores to future risk on the PSSA. Results indicated that while the predictive power was generally low across all measures and risk types, it was substantially larger for the students identified as at-risk on the screener than the other two groups (Table 4). Compared to the PPP values that were obtained from the 2x2 contingency tables, this present finding indicated that the predictive power was stronger for the at-risk students by themselves than when risk was simply designated as a combined measure of some risk and at-risk.

Compared to the PPP results in Table 2, it is worthy to note that the students who scored at some level of risk on the screener contributed to the relatively low overall PPP of the screener's scores. When the data was broken out by the risk types, it was observed that in third grade, at-risk students on the 4Sight measure had on average predictive probability of risk on the PSSA 2.8 times greater than the some risk students; while at-risk students' probability on DORF was 3.9 times greater than some risk students. Fourth graders' probability of PSSA risk was 3.1 times greater for DORF at-risk than DORF some risk students, and 2.6 times greater for 4Sight at-risk students than some risk. Similarly, fifth graders' probability of PSSA risk was 3.6 times greater for DORF at-risk than DORF some risk students, and 5.3 times greater for 4Sight at-risk than some risk students.

DISCUSSION

The purpose of this study was to examine the relationship of a measure of oral reading fluency and reading comprehension to the statewide reading assessment in Pennsylvania. Specifically, we wanted to examine the degree to which a measure of reading comprehension, designed to be closely aligned with the statewide assessment, added value to the predictive relationship of oral reading fluency alone.

Correlations between fall and winter ORF, 4Sight benchmark assessment, and the PSSA for 3rd and 4th grade were consistent with the extant literature that demonstrated significant correlational relationships for oral reading fluency measures both in Pennsylvania (Shapiro

et al., 2006), as well as in many other states (e.g. Buck & Torgesen, 2003; Crawford et al., 2001; Good et al., 2001; McGlinchey & Hixson, 2004; Schilling et al., 2007; Roehrig et al., in press; Stage & Jacobsen, 2001; Wood, 2006). Correlations between the 4Sight benchmark measures and PSSA were somewhat higher and consistent across grades and were similar to those reported by the developers of the 4Sight benchmarks (Success for All Foundation, 2007). Within measure correlations for DORF were very high fall to winter across, while somewhat lower for 4Sight. These outcomes were expected since the fall 4Sight measure is designed as a baseline measure consisting of skill competencies not yet instructed in the grade in which they are administered, but are designed to have strong relationships to the statewide, high stakes test.

Diagnostic validity of DORF and 4Sight

Using the benchmark levels as defined by DIBELS, excellent levels of sensitivity were found across grades and for both the DIBELS and 4Sight measures. At the same time, levels of specificity were much lower and would be considered to be unacceptable for almost all DORF and 4Sight time points, with perhaps the exception of the winter 4Sight specificity for fifth grade. These results suggest that DORF and 4Sight do a reasonable job of predicting PSSA outcomes for students who score at benchmark or above on both of the measures and reach a proficient level on the PSSA. However, those students who score below proficient on the PSSA are much less likely to be predicted accurately based on their DORF or 4Sight scores. Relative to each other, 4Sight benchmarks have higher levels of winter sensitivity compared to winter DORF. The difficulties of accurate prediction for the DORF and 4Sight for students who fell below proficient on the PSSA were especially evident when examining the low levels of PPP across grades and measures.

This suggests that when students score at the some risk or at risk categories of the DORF, or basic/below basic levels of the 4Sight, the measures did not successfully predict students above or below proficiency on the PSSA.

The low specificity and PPP outcomes were particularly salient in Table 3. This table shows the percentage of students at each grade who scored at the some risk category of the DORF, scored advanced/proficient on the 4Sight, and then proceeded to score as advanced/proficient on the PSSA. As evident from Table 3, between 60% and 87% of students across grades who scored below the winter DORF benchmark scored advanced/proficient on the winter 4Sight and on the PSSA. These outcomes suggest that the additional component of an assessment of reading comprehension (i.e., 4Sight benchmarks) substantially assists the diagnostic decision making for students who do not reach the DORF benchmark. The data also suggest that the set cut points for DORF and potentially the 4Sight benchmarks may need to be adjusted if one wants to increase the diagnostic predictability of the measures. Similar findings were suggested by Roehrig et al. (in press) using data from the Florida Comprehensive Achievement Test.

Further investigation of the diagnostic validity of the DORF and 4Sight measures was provided by the ROC analysis which looked at the diagnostic predictions across the entire range of scores of each measure. Outcomes of the analysis examined and compared the prediction to the PSSA scores by the DORF scores alone, 4Sight scores alone, and the

combination of the two scores. Across grades, the combination of the measures (4Sight and DORF) always resulted in better classification rates (AUC) compared to DORF alone. At grade 3, no significant differences were found in the predictive ability of the two assessments to the reading portion of PSSA. At grade four, 4Sight predicted better than DORF alone, and all measures including the combination (4Sight and DORF) showing equal strength in prediction at grade 5.

Predictive validity of DORF for some risk and at-risk students

Data analysis consistently showed that while predictions to PSSA from DORF for those students reaching benchmark (on the DORF) were strong, predictions for those below benchmark were more problematic and mixed. To further investigate the trend within the below benchmark group, a series of logistic regressions were conducted with the three DORF defined categories of low risk, some risk, and at-risk. The outcomes confirmed the findings of the previous analyses showing strong predictions for those scoring at benchmark on the DORF. However, students scoring in the at-risk category were more accurately predicted compared to those in the some risk category. This finding has potential impact for schools using RTI as a model for identifying students in need of supplemental instruction. In particular, the data indicate that DORF does a reasonable job at identifying pass/fail rates for students who are at each end of the spectrum (at-risk or low risk), but a poor job with students identified as some risk by the DORF measure. For these students, it may be important to use further measures (comprehension) to determine if they are in need of supplemental instruction and more accurately predict pass/fail rates on end of year testing. It is important to recognize that the design of the DIBELS benchmarks were such that while Low and At risk benchmarks predict meeting or not meeting future goals (such as scores on high stakes assessments), the “some risk” designation is supposed to indicate that an accurate prediction is not possible one way or the other. Therefore one wouldn’t expect DIBELS scores in the some risk category to have either good sensitivity or specificity in predicting the PSSA.

The findings of this study have substantial implications for the decision making processes, especially within an RTI framework. When data analysis teams are analyzing DORF benchmark data to make decisions about which students are in need of instructional interventions, the use of the existing DORF benchmarks alone for students in grades 3 through 5 may result in far more students being identified for strategic (i.e., tier 2) intervention when in actuality they may not need such services. In other words, there is potential for more students to be identified as at some risk based on DORF measures alone than are truly at risk creating a high number of false-positives for intervention. As a result, some students would be placed in supplemental interventions (Tier 2) when in reality they do not need supplemental instruction. From an instructional perspective, erring on the side of providing more students with intervention than really need such intervention is far less of a concern than errors made in the opposite direction, not providing instruction to those truly in need. The results of this study show that students who do reach the DORF defined benchmarks have a high probability of success on the high stakes test and are not in need of supplemental instruction. At the same time, when a school finds itself in times of tight

resource allocation, making sure that the available resources are given to those most certainly in need of intervention becomes a priority.

Contribution of 4Sight for some risk and at-risk students

The addition of a measure of reading comprehension into the prediction model with oral reading fluency generally improved the accuracy of the prediction of outcomes. This finding was particularly true for students who fell below benchmark on the oral reading fluency measure. Although the 4Sight measure appeared to do as good or better a job as DORF in all grades, at no point did the predictions from DORF exceed those made by the 4Sight alone or combination of the measures. Given the increasing shifts toward comprehension as one approaches the fifth grade (e.g., Shinn et al., 1992), such findings were not surprising. In particular, fluent readers who are challenged by the content of the material may be adjusting their fluency rates in order to fully process the information. As such, students who sacrifice speed for accuracy show up as in the “some risk” category according to DORF benchmarks, but are really successfully meeting the requirements of comprehension. This is important to note because the statewide assessments rely heavily on a student’s ability to comprehend text, therefore a student who may be reading at a slower rate to enable his/her comprehension may fall into the some risk category on DIBELS, but perform adequately on the end of the year reading assessment. As evident in Table 3, the percentage of students who achieved benchmark on the oral reading fluency measure and succeeded at the advanced/proficient level of the 4Sight but still fell below proficient on the PSSA increased from 3.5% at grade 3 to 11.9% and 9.5% at grade 4 and 5, respectively.

The contribution of the comprehension measure for decision making can be examined further by looking at those students scoring at or above benchmark on DORF but who subsequently score below proficient on both the 4Sight and the PSSA. Across grades, students in this category made up 46.5% of third graders, 50.0% of fourth graders, and 82% of fifth graders. These data suggest that the addition of the comprehension measure for those above benchmark, especially for the higher elementary grades, may be an important measure to uncover those students who while seemingly have no difficulty in their fluency but struggle with comprehension. Previous studies have demonstrated that while fluency is essential for reading comprehension, it may not be sufficient for successful reading comprehension for all students (Catts & Hogan, 2003; Duke et al., 2004; Gough & Tunmer, 1986). In particular, fluent readers who are challenged by the content of the material may be adjusting their fluency rates in order to fully process the information. It is possible that these students will potentially display late developing reading comprehension difficulties as identified by Catts and Hogan (2003). Future longitudinal follow-up of these students into adolescence is clearly needed to address this question.

A particular problem uncovered in this study was the potential of using the recommended DORF benchmarks for grades 4 and 5. It is important to note that the benchmarks for grades 4 and 5 on DORF were taken from the normative reports of Fuchs et al. (1993) and Hasbrouck and Tindal (1992), and were not derived based on the conditional probabilities of grades K through 3. In fact, the Hasbrouck and Tindal (1992) data has been updated (Hasbrouck & Tindal, 2005), although the updated normative data were similar to the

original 1992 norms. As such, it should not be surprising that the benchmarks provided for grades 4 and 5 especially could be problematic.

The outcomes of this study strongly suggest that one needs to look carefully at the recommended DORF benchmarks for the upper elementary grades, especially in regard to predicting outcomes on the statewide, high stakes test. Although ORF measures were never developed to be predictive of statewide assessment tests, the statewide tests are very often the measure against which schools judge the impact of their instructional content and processes. As students develop and advance through elementary grades, the predictability of the ORF changes, with the combination of measures that include both oral reading fluency and reading comprehension resulting in the better predictability to the statewide assessment outcomes.

It is important to note that the oral reading fluency measure provided equal levels of predictability to the 4Sight measure to PSSA outcomes for 5th graders, for those who scored at or above the benchmark levels. Given the nature of the time it takes to administer these two measures (1 min per reading probe versus an hour), this study clearly reinforces the strength of oral reading fluency as a strong screening measure for overall reading outcomes throughout the upper elementary grades. However, the relatively poor prediction rates for students who score below the oral reading fluency benchmarks are highly problematic. Research needs to examine how cut points need to be adjusted to maximize sensitivity and specificity for fluency measures. The level of PPP found in this study can mislead schools when they are examining the impact of their supplemental or Tier 2 instruction. Both schools and educational researchers that are evaluating RTI models are at least partially looking at reductions of risk levels for reading failure as keys to instructional improvement and efficacy. The potential over-identification of students in the some risk category may lead some schools to finding fault with their RTI process when in truth, the process is working well toward outcomes on the high stakes test of reading.

Limitations

As with any study, the current investigation has several limitations that must be considered when examining the findings. First, the population used for this study consisted of data from a total of 6 schools, 4 from one district and 2 from different school districts. Although the 4 schools from the same district had similar demographic characteristics, the 2 schools from different districts had some demographics that were a bit different in ethnicity. However, both schools had similar levels of poverty when compared to the other 4 schools from the same district. While we looked at whether there were substantial differences in mean performance across these schools on our measures and did not find substantive differences, future replications of these findings across districts in Pennsylvania or other states are needed to provide generalizability of these findings.

Another limitation of this study is the use of the 4Sight benchmark assessments in Pennsylvania. Clearly, not all districts nationally are using a measure like 4Sight in addition to the DORF measures in RTI models. The nature of a relatively brief (one hour, group administered) measure of reading comprehension that is predictive of the statewide, high stakes assessment that could be repeated at key intervals prior to the administration of the

statewide test, presented an opportunity to enhance the potential decision making of RTI teams. As the 4Sight measures are also being used in many other states, it would be important to replicate these findings across those states. Likewise, there may be individual differences between districts and sites using different measures to map comprehension over time, and unless such analyses similar to those presented in this study are completed, it would not be possible to know that the findings here are generalizable to those sites.

CONCLUSIONS

Any study of the classification and diagnostic accuracy of measure includes tradeoffs between specificity and sensitivity.

The question is on which side of the decision does one want to err. In the area of students potentially identified as at risk who may truly not be at risk (that is low PPP, low specificity), an RTI model using DORF alone would likely decide that such students should be included in tiered intervention groups. In truth, those students may not really need that level of remediation to be successful on the high stakes test. Providing remediation to those who do not really need it, is much less of a problem than the opposite (i.e., not providing remediation to those who do). Thus, the high level of NPP found in this study reflects that the measures used here could all do an excellent job in finding those students who are truly in trouble in reading. Of course, providing intervention to students who may not really need intervention may also tax the available resources. In schools where resource allocation may be limited, making more accurate determinations of potential outcomes based on screening data becomes increasingly important. As such, a careful examination of the calibration of cut points for predicting outcomes needs to be analyzed.

The outcomes of the current study showed that the addition of a measure of reading comprehension to the use of an oral reading fluency measure enhanced the decision making process for identifying students who do or do not have difficulties in comprehension. In particular, using the outcomes of the statewide high stakes measures, the 4Sight benchmark measure used for students at the higher elementary grades was capable of separating those students who did not achieve the defined DORF benchmark but were successful on the outcomes of the statewide assessment. Although DORF and 4Sight alone certainly showed a good level of prediction to the statewide assessment, the combination of the two measures provided the best opportunities for accurate prediction especially at higher elementary grades. Despite the enhancement of the prediction, cut points for the screening measures appeared to produce high levels of false positives, those students who fell below the benchmark on oral reading fluency, yet achieved success on measures of the state reading assessment. Such outcomes are somewhat counter-intuitive to the beliefs of teachers that the DORF measure primarily assesses skills in fluency and decoding, but miss students who read quickly without comprehension (Hamilton & Shinn, 2003). The present study confirmed the strong relationship between ORF and overall reading outcomes that has been consistently substantiated in multiple studies over the past decade (Buck & Torgesen, 2003; Crawford et al., 2001; Good et al., 2001; McGlinchey & Hixson, 2004; Schilling et al., 2007; Shapiro et al., 2006; Roehrig et al., in press; Stage & Jacobsen, 2001; Wood, 2006). At the same time, the study may have also found that there are students for who high levels

of fluency in reading is not a critical skill to achieve full understanding of what they have read.

Future research is needed to both replicate and confirm the findings of this particular investigation in other states, as well as to improve the diagnostic accuracy of screening measures for reading. Continuing examination of the issues regarding reading performance of students in the upper elementary grades is needed to facilitate the processes of implementation of RTI models.

References

- Adams, MJ. Beginning to read: Thinking and learning about print. Cambridge, MA: MIT Press; 1990.
- Berninger, VW.; Abbott, RD.; Billingsley, F.; Nagy, W. Processes underlying timing and fluency of reading: Efficiency, automaticity, coordination, and morphological awareness. In: Wolf, M., editor. Dyslexia, fluency, and the brain. MD: York: Parkton; 2001. p. 383-414.
- Buck, J.; Torgesen, J. Technical Report 1. Tallahassee, FL: Florida Center for Reading Research; 2003. The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test.
- Burns MK, Appleton JJ, Stehouwer JD. Meta-analytic review of responsiveness-to-intervention research: Examining field-based and research-implemented models. *Journal of Psychoeducational Assessment*. 2005; 23(4):381–394.
- Catts HW, Hogan TP. Language basis of reading disabilities and implications for early identification and remediation. *Reading Psychology*. 2003; 24:223–246.
- Crawford L, Tindal G, Stieber S. Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment*. 2001; 7:303–323.
- CTB-McGraw Hill. Pennsylvania Grade 3 Assessment Mathematics and Reading Technical Report. CTB McGraw Hill; 2006. (available http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2006_Gr3_Tech_Report.pdf)
- Danielson L, Doolittle J, Bradley R. Past accomplishments and future challenges. *Learning Disabilities Quarterly*. 2005; 28:137–139.
- Data Recognition Corporation. Technical report for the Pennsylvania System of School Assessment, 2006 Reading and Mathematic Grades. 2007; 4:6–7. (available at http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2006_ReadingMathGr4_6_7_Tech_Report.pdf).
- Data Recognition Corporation. Technical report for the Pennsylvania System of School Assessment, 2006 Reading and Mathematic Grades. 2007; 5:8–11. (available at http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2006_ReadingMathGr4_6_7_Tech_Report.pdf).
- Duke, NK.; Pressley, M.; Hilden, K. Difficulties with reading comprehension. In: Stone, CA.; Stillman, R.; Ehren, BJ.; Apel, K., editors. *Handbook of language and literacy*. 2004.
- Fuchs, D.; Fuchs, LS.; Mathes, PG.; Lipsey, MW.; Roberts, HP. Is “learning disabilities” just a fancy term for low achievement? A meta-analysis of reading differences between low achievers with and without the label. In: Bradley, R.; Danielson, L.; Hallahan, D., editors. *Identification of learning disabilities: Research to practice. The LEA series on special education and disability*. Mahwah, NJ: Lawrence Erlbaum Associates; 2002. p. 737-762.
- Fuchs LS, Fuchs D, Hamlett CL, Waltz L, Germann G. Formative evaluation of academic progress: How much growth should we expect? *School Psychology Review*. 1993; 22:27–48.
- Fuchs LS, Fuchs D, Hosp MK, Jenkins JR. Oral Reading Fluency as an Indicator of Reading Competence: A Theoretical, Empirical, and Historical Analysis. *Scientific Studies of Reading*. 2001;5.
- Fuchs, LS.; Tindal, G.; Deno, S. Effects of varying item domain and sample duration on technical characteristics of daily measures in reading. University of Minnesota: Institute for Research on Learning Disabilities; 1981.

- Gersten R, Dimino JA. RTI (response to intervention): Rethinking special education for students with reading difficulties (yet again). *Reading Research Quarterly*. 2006; 41(1):99–108.
- Good, RH.; Jefferson, G. Contemporary perspectives on curriculum-based measurement validity. In: Shinn, MR., editor. *Advanced applications of curriculum-based measurement*. New York: Guilford; 1998. p. 61-88.
- Good RH, Simmons D, Kame'enui E. The importance of decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*. 2001; 5:257–288.
- Gough PB, Tunmer WE. Decoding, reading, and reading disability. *Remedial & Special Education*. 1986; 7:6–10.
- Haager, D.; Klingner, J.; Vaughn, S., editors. *Evidence-based reading practices for response to intervention*. Baltimore, MD: Paul H. Brookes; 2007.
- Hamilton C, Shinn MR. Characteristics of word callers: An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills. *School Psychology Review*. 2003; 32:228–240.
- Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristics curves derived from the same cases. *Radiology*. 1983; 148:839–843. [PubMed: 6878708]
- Hasbrouck JE, Tindal G. Curriculum-based oral reading fluency norms for students in grades 2 through 5. *Teaching Exceptional Children*. 1992 Spring;:41–44. [PubMed: 1396950]
- Hasbrouck, JE.; Tindal, G. Technical report #33. Eugene, OR: Behavioral Research and Training; 2005. Oral reading fluency: 90 years of measurement.
- Hintze JM, Silbergitt B. A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review*. 2005; 34:372–386.
- Kame'enui E, Simmons D. Introduction to this special issue: The DNA of reading fluency. *Scientific Studies of Reading*. 2001; 5:203–210.
- Kaminski, RA.; Good, RH. Assessing early literacy skills in a problem-solving model: Dynamic indicators of basic early literacy skills. In: Shinn, MR., editor. *Advanced applications of curriculum-based measurement*. New York: Guilford; 1998. p. 113-142.
- Keller, MA.; Shapiro, ES. General outcome measures and performance on standardized tests: An examination of long-term predictive validity. Paper presented at the meeting of the National Association of School Psychologists Convention; Atlanta, GA. 2005.
- Marston, DB. A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In: Shinn, MR., editor. *Curriculum-based measurement: Assessing special children*. New York: Guilford Press; 1989. p. 18-78.
- Marston D, Magnusson D. Implementing curriculum-based measurement in special education and regular education settings. *Exceptional Children*. 1985; 52:266–276. [PubMed: 4076300]
- McGlinchey MT, Hixson MD. Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review*. 2004; 33:193–203.
- Meehl PE, Rosen A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*. 1955; 3:195–216.
- Meng X, Rosenthal R, Rubin DB. Comparing correlated correlation coefficients. *Psychological Bulletin*. 1992; 111:172–175.
- Nathan RG, Stanovich KE. The causes and consequences of differences in reading fluency. *Theory Into Practice*. 1991; 30:176–184.
- National Reading Panel. Reports of the subgroups. Bethesda, MD: National Institute of Child Health and Human Development; 2000. *Teaching children to read: An evidence based assessment of the scientific research literature on reading and its implications for reading instruction*.
- Roehrig AD, Petscher Y, Nettles SM, Hudson RF, Torgesen JK. Accuracy of the DIBELS Oral Reading Fluency Measure for Predicting Third Grade Reading Comprehension Outcomes. *Journal of School Psychology*. in press.
- Schilling SG, Carlisle JF, Scott S, Zeng J. Are fluency measures accurate predictors of reading achievement? *Elementary School Journal*. 2007; 107:429–448.

- Shapiro ES, Keller MA, Edwards L, Lutz G, Hintze JM. General outcome measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment*. 2006; 42(1):19–35.
- Shephard LA. The role of assessment in a learning culture. *Educational Researcher*. 2000; 29:4–14.
- Shinn, M., editor. *Curriculum based measurement: Assessing special children*. New York: Guilford Press; 1989.
- Shinn, M., editor. *Advanced applications of curriculum-based measurement*. New York: Guilford; 1998.
- Shinn M, Good RH, Knutson N, Tilly DW, Collins V. Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review*. 1992; 21:459–479.
- Speece DL, Case LP, Molloy DE. Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities Research & Practice*. 2003; 18(3):147–156.
- Stage S, Jacobsen M. Predicting student success on a state-mandated performance-based assessments using oral reading fluency. *School Psychology Review*. 2001; 30:407–419.
- Streiner DL. Diagnosing tests: Using and misusing diagnostic screening tests. *Journal of Personality Assessment*. 2003; 81:209–219. [PubMed: 14638445]
- Successful for All Foundation. *4Sight reading and math benchmarks 2006–07 technical report for Pennsylvania*. Baltimore: Success for All Foundation; 2007.
- Swets JA. Measuring the diagnostic accuracy of diagnostic systems. *Science*. 1988; 240:1285–1293. [PubMed: 3287615]
- Tatano-Beck C, Gable RK. Further validation of the postpartum depression screening scale. *Nursing Research*. 2001; 50:155–164. [PubMed: 11393637]
- Thurlow R, van den broek. Automaticity and inference generation. *Reading and Writing Quarterly*. 1997; 13:165–184.
- Tindal, G.; Deno, S. *Daily measurement of reading: Effects of varying the size of the item pool*. Minneapolis. University of Minnesota: Institute for Research on Learning Disabilities; 1981.
- Vaughn S, Linan-Thompson S, Hickman P. Response to instruction as a means of identifying students with reading/learning disabilities. *Exceptional Children*. 2003; 69(4):391–409.
- Wayman ME, Wallace T, Tichá R, Espin CA. Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education*. 2007; 41:85–120.
- Wood DE. Modeling the relationship between oral reading fluency and performance on a statewide reading test. *Educational Assessment*. 2006; 11:85–104.

Table 1Descriptive statistics and correlations for 3rd – 5th grade

Grade	Measure	DORF fall	DORF winter	4Sight fall	4Sight winter	PSSA
3 rd	DORF fall	1.00				0.67
	DORF winter	0.94	1.00			0.68
	4Sight fall	0.71	0.71	1.00		0.75
	4Sight winter	0.66	0.67	0.78	1.00	0.78
	<i>Mean</i>	72.95	88.68	15.83	17.62	1295.92
	<i>SD</i>	32.52	36.02	5.79	5.99	136.28
4 th	DORF fall	1.00				0.64
	DORF winter	0.93	1.00			0.68
	4Sight fall	0.67	0.67	1.00		0.71
	4Sight winter	0.66	0.66	0.72	1.00	0.74
	<i>Mean</i>	87.35	106.04	16.14	17.39	1283.01
	<i>SD</i>	32.19	34.52	5.39	5.64	208.58
5 th	DORF fall	1.00				0.73
	DORF winter	0.92	1.00			0.75
	4Sight fall	0.70	0.64	1.00		0.74
	4Sight winter	0.70	0.66	0.80	1.00	0.76
	<i>Mean</i>	101.66	112.78	17.88	19.22	1282.23
	<i>SD</i>	35.92	35.05	5.03	5.74	201.34

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Diagnostic efficiency results

	Sensitivity	Specificity	PPP	NPP	TPV	AUC
<i>3rd grade</i>						
Fall DORF	0.96	0.55	0.33	0.98	0.63	0.87
Fall 4Sight	0.96	0.44	0.29	0.98	0.54	0.89
Winter DORF	0.95	0.59	0.36	0.98	0.66	0.87
Winter 4Sight	0.94	0.63	0.40	0.98	0.70	0.91
<i>4th grade</i>						
Fall DORF	0.88	0.49	0.30	0.94	0.57	0.79
Fall 4Sight	0.96	0.43	0.28	0.98	0.53	0.83
Winter DORF	0.79	0.59	0.33	0.92	0.63	0.79
Winter 4Sight	0.90	0.62	0.43	0.95	0.69	0.88
<i>5th grade</i>						
Fall DORF	0.97	0.58	0.38	0.99	0.66	0.92
Fall 4Sight	1.00	0.63	0.41	1.00	0.70	0.92
Winter DORF	0.93	0.61	0.41	0.97	0.68	0.92
Winter 4Sight	0.87	0.74	0.47	0.96	0.76	0.91

Note. PPP = positive predictive power, NPP = negative predictive power, TPV = total predictive value, AUC = area under curve.

Table 3

Percentage of students scoring as advanced/proficient on winter 4Sight and PSSA by DORF risk level

	DORF Risk Level	4Sight Risk Level	Percentage (number/total) at Advanced/Proficient on PSSA
Grade 3	Low risk	Adv/proficient	96.5% (110/114)
		Basic/below basic	82.7% (19/23)
	Some risk	Adv/proficient	87.1% (27/31)
		Basic/below basic	53.4% (23/43)
	At risk	Adv/proficient	57.1% (8/14)
		Basic/below basic	15.4% (12/78)
Grade 4	Low risk	Adv/proficient	88.1% (96/109)
		Basic/below basic	44.0% (16/36)
	Some risk	Adv/proficient	63.3% (19/30)
		Basic/below basic	31.4% (17/54)
	At risk	Adv/proficient	40.0% (4/10)
		Basic/below basic	6.7% (4/60)
Grade 5	Low risk	Adv/proficient	90.5% (67/74)
		Basic/below basic	47.1% (8/17)
	Some risk	Adv/proficient	60.0% (14/23)
		Basic/below basic	15.2% (4/22)
	At risk	Adv/proficient	28.6% (4/14)
		Basic/below basic	12.9% (4/31)

Table 4

Mean predicted probability results from logistic regression

	Fall DORF	Winter DORF	Fall 4sight	Winter 4sight
<i>3rd grade</i>				
At-risk	0.52	0.53	0.39	0.55
Some risk	0.12	0.15	0.12	0.23
Low risk	0.03	0.02	0.06	0.03
<i>4th grade</i>				
At-risk	0.40	0.58	0.40	0.65
Some risk	0.17	0.15	0.15	0.26
Low risk	0.08	0.09	0.02	0.05
<i>5th grade</i>				
At-risk	0.54	0.59	0.50	0.64
Some risk	0.12	0.22	0.19	0.08
Low risk	0.04	0.04	0.05	0.08