



Published in final edited form as:

J Educ Psychol. 2010 May 1; 102(2): 327–340. doi:10.1037/a0018448.

Selecting At-Risk First-Grade Readers for Early Intervention: Eliminating False Positives and Exploring the Promise of a Two- Stage Gated Screening Process

Donald L. Compton, Douglas Fuchs, Lynn S. Fuchs, Bobette Bouton, Jennifer K. Gilbert,
Laura A. Barquero, Eunsoo Cho, and Robert C. Crouch

Vanderbilt University

Abstract

The purposes of this study were (a) to identify measures that when added to a base 1st-grade screening battery help eliminate false positives and (b) to investigate gains in efficiency associated with a 2-stage gated screening procedure. We tested 355 children in the fall of 1st grade, and assessed for reading difficulty at the end of 2nd grade. The base screening model, included measures of phonemic awareness, rapid naming skill, oral vocabulary, and initial word identification fluency (WIF). Short-term WIF progress monitoring (intercept and slope), dynamic assessment, running records, and oral reading fluency were each considered as an additional screening measure in contrasting models. Results indicated that the addition of WIF progress monitoring and dynamic assessment, but not running records or oral reading fluency, significantly decreased false positives. The 2-stage gated screening process using phonemic decoding efficiency in the first stage significantly reduced the number of children requiring the full screening battery.

Key Terms

Screening; Prediction; Response-to-Intervention; Reading; Reading Disability; Replication; Short-term Progress Monitoring; Dynamic Assessment; Oral Reading Fluency; Running Record

The success of prevention models, such as Responsiveness-To-Intervention, hinges on an accurate determination of which children are at risk for future difficulty (e.g., Compton, Fuchs, Fuchs, & Bryant, 2006; Fuchs & Fuchs, 2007; Good, Simmons, & Kame'enui, 2001; MaCardle, Scarborough, & Catts, 2001; VanDerHeyden, Witt, & Gilbertson, 2007). Correct identification of children at risk for reading difficulty (RD) in kindergarten and first grade can trigger early intervention prior to the onset of significant problems, which, in turn, can place children on the path of normal reading development. Universal screening is a principle means of identifying at-risk children (see Glover & Albers, 2007). In both research and practice, it usually involves precursor measures of literacy (e.g., phonemic awareness, letter naming fluency, concepts about print, word reading, oral language ability) and the use of a cut-point to demarcate risk and non-risk (for a review, see Jenkins, Hudson, & Johnson, 2007).

By definition, a diagnostic screening measure is a brief assessment that provides predictive information about a child's development in a specific academic area. Its purpose is to identify any children who are at-risk so that these children can receive extra support through early intervention. The screening measure is given to all children and used to identify an initial risk pool of children suspected of being at risk of developing RD. Screening information must be dichotomized into a yes-no decision of risk for each child screened. Typically, risk decisions are made by selecting a critical cut-point along a continuum of

scores on a single or group of screening measures. A child scoring below the cut-point is considered at risk of developing RD, whereas a child scoring above the cut-point is not. The cut-point can be adjusted up or down to produce more or fewer positive decisions.

For early intervention programs to work effectively screening procedures for determining RD risk must yield a high percentage of true positives (approaching 100%). Adjusting the cut-point to be more lenient will give the desired effect of increasing the probability that a greater percentage of true-positives will be identified as at risk for RD. Unfortunately, more lenient cut-points will also result in a greater number of false-positives, or children who score below a cut-point but eventually become competent readers. False positives undermine prevention efforts by burdening schools with the obligation to provide early intervention to an unnecessarily large percentage of the population (Fletcher et al., 2002; Jenkins & O'Connor, 2002). Alternatively, if the cut-point is made stricter to decrease the probability of selecting false-positives, then the number of true-positives selected will necessarily go down. Children who score above a cut-point on the screener but later exhibit serious reading problems are known as false negatives. False negatives diminish prevention efforts by depriving at-risk children of the intervention they require (Jenkins et al., 2007; Torgesen, 2002a). Thus, when setting cut-points a balance must be established between true-positives and false-positives. This balance should be determined by the negative ramifications of misdiagnosing true positives as not at risk for RD versus the cost of providing intervention to children who are false-positives and will develop normally in reading without the intervention.

The accuracy of a screener to correctly distinguish true positives from true negatives is often characterized in terms of “sensitivity” and “specificity.” Sensitivity, refers to the degree a measure correctly identifies children as at high risk for RD (i.e., true positives). It is calculated by dividing the number of true positives by the sum of true positives and false negatives. Sensitivity increases as false negatives decrease. Specificity, on the other hand, refers to how well a measure correctly identifies children at low risk for RD (i.e., true negatives). It is derived by dividing the number of true negatives by the sum of true negatives and false positives. Specificity increases as false positives decrease. For early identification to be accurate, screening must yield a high percentage of *true* positives (e.g., sensitivity rates above .90 [Jenkins, 2003]), while limiting *false* positives, and thereby producing a manageable (economical) risk pool.

Challenges to the Development of a Useful Universal Screen

Universal screening batteries in reading typically suffer from two persistent problems: Classification accuracy rates are generally too low to be used in early prevention models, and screening batteries with adequate classification accuracy often use multiple measures that require too much administration time per child and, therefore, are inefficient. Below, we expand on each of these shortcomings and describe how the present study addresses them.

The majority of screening studies used a one-stage approach, and reported classification accuracy well beyond the acceptable range, with false positives ranging from 20% to 60% (see Jenkins & O'Connor, 2002; Torgesen, 2002b); false negatives running from 10% to 50% (Catts, 1991; Scarborough, 1998; Torgesen, 2002b). We were able to identify only two early screening studies that demonstrated acceptable classification accuracy with sensitivities above .90 and specificities above .80. O'Connor and Jenkins (1999) used a multiple-measure screening battery involving letter naming fluency, phoneme segmentation, and sound repetition administered in the fall of kindergarten to correctly classify 100% true positives and 88% true negatives with respect to word reading performance at the end of first grade. Compton et al. (2006) reported that, in fall of first grade, a screening battery comprising word identification fluency (WIF), sound matching, rapid digit naming, and oral

vocabulary, when combined with 5 weeks of WIF progress monitoring, predicted RD on a composite reading measure at the end of second grade with sensitivity and specificity estimates of .90 and .83, respectively.

Whereas the Compton et al. (2006) screening battery predicted future RD risk with precision, it was too long and inefficient for use as a universal screen with all first-grade children. Recognition of this fact prompted the notion of a two-stage screening process. In the first stage, a single efficient measure would be administered to all children in hopes of eliminating from the risk pool those considered at low risk for developing RD (i.e., true negatives). Only those children with scores in the risk range would then be administered a battery of tests in the second stage.

The purpose of this study was to develop and validate a two-stage screening procedure with both sensitivity and specificity above .90. In the first stage, we examine measures that, when added to the base model of Compton et al. (2006) further decreased the number of false positives and therefore increased specificity of the screening battery. In the second stage, we identify a single measure showing promise to eliminate true negatives, thereby limiting the number of students requiring the more time intensive full screening battery.

A Two-Stage Approach to Screening

In selecting stage-one measures with potential to identify a large percentage of true negatives while leaving true positives in the risk sample, we decided to use standardized word-level measures because they are quick to administer, represent a large and diverse corpus of words/nonwords, and correlate highly with future reading skill (e.g., Wagner et al., 1997). Although these measures exhibit significant positive skew at the start of first grade, this is of little consequence when identifying children at the upper range, who are true negatives for RD. Thus, we pit various standardized word-level reading measures against each other to identify which show the most promise in eliminating true negatives in the first step of a two-step screening process.

In stage two, children who fail the initial screen are assessed with a multivariate screening battery to discriminate true positives from false positives. In selecting additional measures to improve specificity of the base model, we sampled tests from two broad categories. Measures in category #1 quantify children's potential to benefit from early reading instruction. Specifically, we considered (a) two types of short term (i.e., 5 weeks) progress-monitoring measures based on WIF and (b) a measure of dynamic assessment (DA). Measures in category #2 represent oral reading in connected text because it better approximates demands of the reading process than reading isolated words.

Stage-two measures of responsiveness-to-instruction—Our two WIF progress-monitoring measures were used to distinguish children scoring below criteria on the screening battery but showing adequate response to classroom instruction, as revealed by slope and/or level of performance. The two WIF measures are identical in form, but differ in the range of words selected for the sampling corpus: one samples words narrowly (WIF_N; sampling from the 100 most frequent words); the other, more broad (WIF_B; sampling from the 500 most frequent words). The combination of WIF_N slope and intercept has been shown to significantly improve the classification accuracy of the base screening battery Compton et al. (2006).

Our other strategy for indexing a student's potential to benefit from early reading instruction was a single-session DA, which indexes the degree of scaffolding, or amount of assistance, a child needs on tasks that tap key learning processes (see Grigorenko & Sternberg, 1998). In this study, we operationalized DA in terms of graduated prompts (moving from initially

implicit hints that gradually become more explicit) to represent a titration process for estimating the minimum amount of help a child needs to learn. Campione, Brown, Ferrara, Jones, and Steinberg (1985) showed that students of lower ability (vs. higher ability) required more help to reach criterion and show transfer. Studies have reported that DA scores make a statistically significant and unique contribution in predicting achievement above and beyond one-point-in-time scores (Campione, 1989; Day, Engelhardt, Maxwell, & Bolig, 1997; Resing, 1993; Swanson, 1995; Swanson & Howard, 2005), specifically with respect to reading performance (Jenkins & O'Connor, 2002; Murray et al., 2000; Spector, 1992).

Stage-two measures of reading in connected text—We also considered measures of reading in connected text for use in the multivariate screening battery. Reading connected text better resembles the actual demands of reading than the kind of measures typically used as universal screeners and is considered a more authentic assessment of school and home reading (Fawson, Ludlow, Reutzel, Sudweeks, & Smith, 2006). One text-reading measure was Running Records (RR); the other was curriculum-based measurement passage oral reading fluency (ORF).

Originally introduced by Clay (1993) to identify the needs of struggling first graders enrolled in Reading Recovery, RRs are now used widely to gauge the instructional reading level of developing readers and identify children at risk for RD (Bean, Cassidy, Grumet, Shelton, & Wallis, 2002). RRs, in which students read leveled passages, is a test of contextual reading accuracy and strategy use (Clay; Fountas, & Pinnell, 1996; Rathvon, 2004). We found no published studies using RRs at first grade, alone or in combination with other measures, to predict later RD.

As indicated, ORF is an example of curriculum based measurement. There is substantial evidence to support ORF's validity and reliability (e.g., Deno, Marston, Shinn, & Tindal, 1983; Fuchs & Fuchs, 1992; Fuchs et al., 2001), and it is used extensively for screening/ benchmarking and progress monitoring (see Fuchs & Fuchs, 1998; 2007; Hosp & Fuchs, 2005; Kaminski & Good, 1996). With ORF, students read a graded passage orally for a brief, fixed time (usually 1 to 3 min). The score is the number of words read correctly. At second and third grade, ORF correlates strongly with high-stakes reading tests (e.g., Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008; Schilling, Carlisle, Scott, & Zeng, 2007), but distributional problems limit its effectiveness as a first-grade predictor of future reading risk (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009). It is possible, however, that when used in combination with other first-grade screeners, ORF can reduce false positives. Even so, we located no published studies using ORF in combination with other measures in first grade to predict later RD.

Study Purpose

We focused on improving first-grade screening by trying to decrease false positives and exploring a 2-stage gated screening process. We posed two major questions: (a) Do additional screening measures (two alternatives of WIF intercept and slope, DA, RR, and ORF) improve classification accuracy of a base model comprising phonemic awareness, rapid naming skill, oral vocabulary, and initial WIF; and (b) Can a gated screening procedure make universal screening more efficient? To answer these questions, we recruited a sample of 355 first-grade children, over-sampling for children who exhibit low initial reading skills.

Regarding the first question, we explored how well the classification model developed by Compton et al. (2006) transferred to a new population. This distinguished our effort from that of O'Connor and Jenkins (1999), the only related study we could identify that provided

cross validation data. Instead of cross validating on an independent portion of the original sample, we collected a new sample separated by over five years, to replicate the classification model. We contrasted measures to gain insight into which might provide the best added value in reducing false positives. We asked whether additional screening measures (two types of WIF intercept and slope, DA, ORF, and RR) improve classification accuracy of a base model comprising phonemic awareness, rapid naming skill, oral vocabulary, and initial WIF for use in the second part of a 2-stage screening procedure. To address the second research question we contrasted various standardized word-level measures against each other to identify those that showed the most promise in eliminating true negatives for use in the first state of a 2-stage screening procedure to limit the number of children requiring the full screening battery.

Method

Participants

Participants were selected from 56 first-grade classrooms in 14 schools in urban and suburban districts located in Middle Tennessee. Seven study schools were Title I. We assessed every formally-consented child ($n=712$) with three 1-min study identification measures: WIF_N-screen, rapid letter naming (RLN), and rapid sound naming (RSN). With WIF_N-screen, children are presented with a single page of 50 high-frequency words randomly sampled from 100 high-frequency words from the Dolch pre-primer, primer, and first-grade level lists (L. Fuchs et al., 2004). They have 1 min to read words. If they hesitate on an item for 4 sec, the examiner prompts them to proceed. Test/retest reliability exceeds .90. With RLN, the speed at which children name an array of the 26 letters is measured. The score is the number of letters correctly identified in 1 min. With RSN, the speed at which children produce the sounds associated with an array of the 26 letters is measured. The score is the number of sounds correctly identified in 1 min. Test/retest reliability of RLN and RSN exceeds .85. For all three measures, scores were prorated if a child named all items in less than 1 min.

We used these data to divide the 712 children into high, average, and low performing groups with the use of latent class analysis and then randomly selected study children from each group. We over-sampled low-performing children to increase the number of struggling readers in the prediction models. A total of 485 children were included: 310 low-study-entry (LSE), 83 average-study-entry (ASE), and 92 high-study-entry (HSE). Participant selection occurred in late September and early October of first grade. At follow-up in spring of second grade, 130 of the original 485 children (27% of the sample) had moved from the district and were unavailable for assessment. Table 1 displays descriptive and inferential data as well as effect sizes on demographic variables and first-grade screening measures for those who moved versus completed the study. There were no significant differences between movers and stayers on sex, race, and subsidized lunch status. However, a significantly higher percentage of LSE students moved. In addition, stayers were associated with significantly greater scores on RLN, RSN, and WIF_N, presumably due to the greater percentage of LSE children who moved away. In this study we report results from only participants who were present for assessment at the end of second grade. Table 2 presents means, standard deviations, F tests, and effect sizes on the three study identification measures for the 355 children constituting the LSE, ASE, and HSE groups. As anticipated, large differences existed among the three groups, with the most pronounced difference on WIF_N.

Procedure

Following subject selection, participants were assessed individually by trained examiners (each of whom had demonstrated at least 95% accuracy during practice assessments) at two

additional assessment waves: a prediction battery in the fall of first grade to designate RD risk and an outcome assessment at the end of second grade to classify actual RD and NRD. The fall of first grade prediction battery comprised phonemic awareness, rapid naming, oral vocabulary, WIF_N screen, DA, and ORF. At the same time we administered word identification, word attack, sight word efficiency, and phonemic decoding efficiency to evaluate the univariate measures as the first step in the gated screening procedure. Also in the fall of first grade, we administered WIF_N and WIF_B short-term progress monitoring for 5 consecutive weeks, each time with an alternate form. Additionally, RRs were collected from children's first-grade teachers during fall semester. The outcome battery, administered by trained examiners in April of second grade, comprised standardized reading measures: untimed word identification and word attack, timed sight word reading and decoding, and reading comprehension.

Measures

First-Grade Prediction Battery for Designating Risk for RD

Rapid digit naming (RDN): The Comprehensive Test of Phonological Processing: Rapid Digit Naming (Wagner, Torgesen, & Rashotte, 1999) measures the speed at which an individual can name an array of 36 digits. The array includes 2, 3, 4, 5, 7, 8 arranged in random order in four rows with nine digits per row. The child names the digits as quickly as possible. The score is the number of sec required to complete the task. Test/retest reliability exceeded .85 for the first-grade children. RDN, as opposed to rapid letter naming, was used as a predictor because of the superior distributional properties of RDN in at-risk children.

Phonemic awareness: The Comprehensive Test of Phonological Processing: Sound Matching (Wagner et al., 1999) assesses matching of first and last sounds in words, presented along with drawings depicting the words. To assess first sound matching, children are presented with a word and asked to determine which of three different words (depicted as pictures) start with the same sound (e.g., "Which word starts with the same sound as 'pan'? pig, hat, or cone?"). A parallel procedure assesses last sound matching. The test begins with three practice items and consists of 20 items. Split-half reliability exceeded .90 for the first-grade sample.

Oral vocabulary: Woodcock-Johnson Psychoeducational Battery – Revised: Oral Vocabulary (Woodcock, McGrew, & Mather, 2001) assesses the ability to provide synonyms and antonyms in response to stimulus words presented orally. Split-half reliability exceeded .90 for the first-grade sample.

Progress monitoring: WIF N-level, WIF N-slope, WIF B-level, WIF B-slope: Children were administered two alternate forms of WIF_N and WIF_B each week for 5 weeks (see Participants section for description of WIF). Alternate forms were constructed by randomly sampling 50 words per form from the 100 most frequent words for WIF_N and the 500 most frequent words for WIF_B. Words were drawn from the Educator's Word Frequency Guide (Zeno, Ivens, Millard, & Duvvuri, 1995). At each assessment wave, two alternate forms of WIF_N and of WIF_B were administered and the average of the alternate forms was used to represent each type of WIF. Each child's performance on narrow and broad WIF over the 5 weeks was fit to a line using an ordinary least squares regression procedure. Slope was expressed as the estimated number of words gained per week; level as the estimated number of words read at week five. Alternate test-form/stability from 2 consecutive weeks exceeded .92. Initial screening performance on WIF (WIF_N-Screen) was the raw score on the WIF_N measure from the sample identification assessment.

Dynamic Assessment (DA): Pseudowords were used in the measure and instruction of DA. Three decoding skills were taught: CVC (taught as linguistic word families), CVCe, and CVC(C)ing. For each decoding skill, (a) instruction occurs only with the *o* vowel with consonants controlled across levels, and (b) five levels of instructional scaffolding are available, which gradually increase the explicitness of instruction. Between each scaffolding level, 6 items (not used for instruction, but paralleling instructional items) are presented. If the student read 5 of 6 nonwords correctly, the skill is deemed mastered, and he/she moved on to the next DA skill. If a student fails to read all 6 words correctly after each of 5 scaffolding levels for a given skill, the DA is terminated. Scores reflect the scaffolding level needed to correctly decode the 6 items across the 3 levels (where 3 = read 6 nonwords correctly after 1st scaffold on each of 3 skills; 15 = did not reach mastery after 5th scaffold on all 3 skills). We piloted DA on 100 1st-grade children with results suggesting that DA may be an important predictor of reading skill growth: the partial correlation (controlling for initial phonemic awareness and RAN) between DA and concurrent word identification was .68, growth in word identification over a 4-week period .40, and word attack after 4-weeks .65 (see Caffrey, 2005). In addition, the DA measure was stable over the 4-week period with a pretest to posttest correlation of .72.

Running Record (RR): Participating teachers collected RR data at the fall of first grade using a commercial assessment kit aligned with the district reading curriculum (Scott Foresman, 2006). Teachers were trained by the district in administration using a commercial set of leveled books designed to take RRs and an accompanying RR form. Instructional reading level (ranging from level 1 to 24) was defined as the highest level book in which a child achieves 90% or higher word reading accuracy and answers at least 80% of comprehension questions correctly. RRs were not required of first-grade teachers; therefore, we have RRs on approximately 90% of the sample. No psychometric data were available on the RRs.

Oral Reading Fluency (ORF): ORF was measured using first-grade passages (400-word folktales; Jenkins, Heliotis, Haynes, Stein, & Beck, 1986) from the Comprehensive Reading Assessment Battery (Fuchs, Fuchs, & Hamlett, 1989). Students read aloud 2 passages, each for 3 min; the score is the number of words read correctly. Test/retest reliability exceeded .90 for the first-grade children. These ORF passages can be considered “context-free” in that they were not drawn from the classroom curriculum.

First-Grade Univariate Screen and Second-Grade Battery to Determine RD Status

Untimed decoding skill: The Woodcock Reading Mastery Test – R/NU: Word Attack (WRMT-R: WAT, Woodcock, 1998), a norm-referenced test, evaluates children’s ability to pronounce pseudowords presented in list form. It contains 45 nonsense words, ordered from most easy to most difficult. Split-half reliability exceeded .90 and .94 for the first-grade and second-grade sample, respectively.

Untimed word identification skill: The Woodcock Reading Mastery Test – R/NU: Word Identification (WRMT-R: WID, Woodcock, 1998), a norm-referenced test, asks children to read single words in list form. It consists of 100 words ordered in difficulty. Split-half reliability exceeded .90 and .96 for the first-grade and second-grade sample, respectively.

Sight word reading efficiency: The Test of Sight Word Reading Efficiency (TOWRE: SWE, Torgesen, Wagner, & Rashotte, 1997) is a norm-referenced measure of sight word reading accuracy and fluency, assessing the number of real words accurately read in 45 sec.

It consists of 104 words ordered in difficulty. Split-half reliability exceeded .91 and .95 for the first-grade and second-grade sample, respectively..

Phonemic decoding efficiency: The Test of Phonemic Decoding Efficiency (TOWRE: DE, Torgesen et al., 1997) is a norm-referenced measure of decoding accuracy and fluency, measuring the number of nonsense words accurately decoded in 45 sec. It consists of 63 words ordered in difficulty. Split-half reliability exceeded .90 and .95 for the first-grade and second-grade sample, respectively..

Reading comprehension: Woodcock Reading Mastery Test – R/NU: Passage Comprehension (WRMT-R: PC) (Woodcock, 1998) is a norm-referenced, modified cloze procedure. For the first set of items, the tester presents a symbol, or rebus, and asks the child to point to the picture corresponding to the rebus. Next, the child points to the picture representing words printed on the page. In later items, the child reads a passage silently and identifies the missing word in the passage. Split-half reliability exceeded .90 for the second-grade sample.

Classification of Second-Grade RD Status—As with Compton et al. (2006), we classified children as RD and NRD based on a composite that summed weighted standardized scores for untimed word identification and word attack, timed sight word reading and decoding, and reading comprehension. The weighting factor for each word identification and decoding measure was .167; for comprehension, .333. This allowed the composite measure of reading to be equally weighted across untimed word-level reading and decoding, timed word-level reading and decoding, and reading comprehension. Children with scores below 85 on the composite were classified RD. The composite was used to provide a balanced representation of reading ability by limiting the effects of a single reading skill on the classification. In this way, 54 children were identified as RD at the end of second grade.

Data Analysis—We began by replicating Compton et al. (2006), a necessary step for the present extension. To obtain parameter estimates for predicting the probability of RD, a logistic regression model was run with original variables (WIF_N-screen, RDN, SM, OV, WIF_N-level, WIF_N-slope) and 206 students from Compton et al. To classify the 355 children in the present sample into RD and NRD classes, we derived each child's predicted RD status by applying the derived logistic regression model (with associated parameters). We then reported resulting indices of classification accuracy: sensitivity, specificity, and area under the ROC curve (AUC).

Then, to extend the results presented in Compton et al. (2006), we used only the present sample of 355 children. Given that each child was a member of a first- and second-grade classroom, we used a 2-level analysis procedure. We used multi-level analyses for two reasons. This allowed us to estimate the variance associated with first- and second-grade classroom membership on RD designation. Although we were not interested in modeling variance at the classroom level, we thought it would be of value to provide estimates of intraclass correlations (ICCs) for those interested in classroom effects on RD prediction. (Note: Partitioning variance into child and classroom levels has little effect on estimated fixed effect parameters and therefore does not influence classification accuracy.) Also, a 2-level model allowed for more accurate estimations of the standard errors for the child-level variables and therefore allowed for more accurate estimation of *p*-values of model predictors.

Because students from the same first-grade classroom did not all enter the same second-grade classroom, students were cross-classified in analyses (Raudenbush & Bryk, 2002).

Therefore 2-level cross-classified binary logistic regression models were used to predict membership in the second-grade RD and NRD groups for each classification approach. We initially tested for independence in the data using an unconditional multilevel model. The 2-level unconditional cross-classified model was created to predict students' log-odds of being RD (level 1) controlling for the variance accounted for by first- and second-grade classrooms (level 2). In the logistic regression equations first-grade classrooms are represented by the subscript j and second-grade classrooms by the subscript k .

$$\eta_{ijk} = \beta_{0jk} \quad \text{Level 1}$$

where: $\eta_{ijk} = \log\left(\frac{\phi_{ij}}{1 - \phi_{ij}}\right)$

$$\beta_{0jk} = \gamma_{00} + b_{00j} + c_{00k} \quad \text{Level 2}$$

where: η_{ijk} is the log of the odds of success. Success in our case is "being RD".

β_{0jk} is the log-odds of students in 1st grade classroom j and in 2nd grade classroom k "being RD".

γ_{00} is the grand-mean log-odds of "being RD".

b_{00j} is the random effect of 1st grade classroom j .

c_{00k} is the random effect of 2nd grade classroom k .

In the cross-classified models, the random main effect of each first- and second-grade classroom was included in the model. The average first-grade classroom effect, b_{00j} , and the average second-grade classroom effect, c_{00k} , were assumed to be normally distributed with a mean of 0 and variances of τ_{b00} and τ_{c00} , respectively. This model assumes that sources of variability in RD status are associated with particular characteristics of first- and second-grade classrooms.

After running the unconditional models, we completed a series of conditional models using child-level variables to predict RD status within the cross-classified framework. Ignoring dependency in the data would have increased the likelihood of inflated standard errors and Type I errors (Raudenbush & Bryk, 2002). The general form of the conditional 2-level cross-classified model is:

$$\eta_{ijk} = \beta_{0jk} + \beta_{1jk}X_{ijk} + \beta_{2jk}X_{ijk} \dots \beta_{njk}X_{ijk} \quad \text{Level 1}$$

where: $\eta_{ijk} = \log\left(\frac{\phi_{ij}}{1 - \phi_{ij}}\right)$

$$\begin{aligned}\beta_{0jk} &= \gamma_{00} + b_{00j} + c_{00k} \\ \beta_{1jk} &= \gamma_{10} \\ \beta_{2jk} &= \gamma_{20} \\ &\vdots \\ \beta_{njk} &= \gamma_{n0}\end{aligned}$$

Level 2

where: $\gamma_{10} - \gamma_{n0}$ represent the increase in the log-odds of students “being RD” associated with child-level predictors $X_1 - X_n$.

The effects of the various screening measures on the log-odds of RD were assumed to be fixed across either first- or second-grade classrooms. In other words, the effect of student screening score on the log-odds of RD was assumed to be the same across classrooms.

Differences in classification accuracy across models were assessed using AUCs. The trade-off between sensitivity and specificity is characterized using ROCs and more specifically AUC. A ROC curve is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for the different possible cut-points of a diagnostic test. AUC is a measure of test discrimination, or the ability of the measure to correctly classify RD versus NRD. Consider the situation in which second-grade children are already correctly classified into the RD and NRD groups. If one were to randomly pick one child from the RD group and one from the NRD group and test both, the child scoring lower on the first-grade prediction battery should be the one from the RD group. The AUC represents the percentage of randomly drawn pairs for which this is true (that is, the test correctly classifies the two children in the random pair) and ranges from 0.5 (i.e., chance performance) to 1.0 (i.e., perfect performance) (Swets, 1992). An AUC greater than .90 is considered excellent; .80 to .90, good; .70 to .80, fair; and below .70, poor. In this study, differences in predictive accuracy across models, as measured by AUC differences, were indicated by calculating a critical ratio z value between two AUC, with values greater than 1.96 designated as significant. Critical ratio values were corrected for the correlation introduced between the two AUC from using the same sample of participants for each model (see Hanley & McNeil, 1983). To help with comparisons of sensitivity and specificity across models we set a cut-point for each model in which sensitivity that was as close to 0.90 as possible allowing the reader to focus on the associated specificity across models.

For the 2-step gated procedure, we identified cut-points on various first-grade standardized measures of word-level reading (word identification, word attack, sight word efficiency, and phonemic decoding efficiency) such that all true positives for RD in second grade had a 99% probability of scoring below the cut-point. This effectively removed a portion of true negatives who would not need to be administered the larger screening battery. We then compared the relative proportion of children who would require administration of the larger screening battery as a function of initial screening device and used a z test for proportions with dependent samples to test for statistical significance.

Results

Means, standard deviations, and correlations among the first-grade screening measures and the second-grade composite reading outcome (used to define RD) are presented in Table 3. All correlations except those involving WIF_N-slope were statistically significant. WIF_N-slope was not significantly correlated with oral vocabulary, WIF_N-screen, WIF_B-intercept, ORF, or RR. The correlations between the first-grade predictor variables and the composite measure of second-grade reading were all significant, ranging from .21 to .83.

The correlation between DA and other measures were negative because lower performance on DA indicated less scaffolding was necessary and therefore superior performance.

Replication

Two prediction models based on simple logistic regression are presented in Table 4. Model 1 presents the original logistic regression model based on the 206 participants in Compton et al. (2006). These parameters do not exactly match those reported in the original study, because Compton et al. used standard WIF scores based on local norming data. Because we did not have local norms for the present study, we based the original and replication models on raw score estimates of WIF_N-screen, WIF_N-level, and WIF_N-slope. The overall classification accuracy was acceptable within an early identification framework with an AUC over .90 and specificity of .83 associated with a sensitivity of .90. The model parameters developed in the original model were then applied to the present sample of 355 children to predict membership in the RD and NRD groups. Results of the replication are shown in Model 2, which replicated well with a specificity of .84 associated with sensitivity of .91 and an AUC of .925. However, results were likely inflated by the composition of the replication sample. The original sample comprised an at-risk population, in which the upper end of distribution is truncated. In the replication sample, we purposefully sampled from the middle and upper portions as well as the lower end of the distribution. The model replicated well with children in the middle and upper portion of the distribution because most scored well above the original risk population on the screening measures and had a low probability of developing RD. To illustrate, we calculated AUCs for each of the LSE, ASE, and HSE groups. For the HSE, the model was perfect, correctly identifying all HSE children as true negatives (i.e., AUC = 1.0). The AUC decreased considerably as group risk increased, with an AUC of .930 for the ASE group and .841 for the LSE group. So although the overall replication results were positive, they were less impressive when considering LSE children, the group that most resembled those in the original model-building sample.

Model Extension: Measures to Reduce False Positives

In the extension models, we examined increases in prediction accuracy above the base model (WIF_N-screen, phonemic awareness, rapid naming skill, and oral vocabulary) by adding measures of WIF progress monitoring, DA, RR, or ORF to the model. In examining the effects of WIF progress monitoring and DA, we used the full sample of 355 children. Whereas in examining the effects of using RR and ORF, we relied on a subsample of 320 children who were administered both the RR and measure of ORF. Prior to running these models, we tested for dependency in the data using an unconditional multilevel model. Two-level unconditional cross-classified model were created to predict students' log-odds of being RD (level 1) controlling for the variance accounted for by first- and second-grade classrooms (level 2). The partitioning of variance into first- and second-grade classrooms for two base models (Model 3 and 7) and five extensions models (Models 4, 5, 6, 8, and 9) are shown in Table 5. Variance associated with classroom was not statistically significant for either first- or second-grade classrooms. However, the percentage of variance associated with first grade classroom membership ranged from 7.8 to 11.6 (i.e., Intraunit correlation coefficient [IUCC]; a cross-classified model ICC). The variance associated with second-grade classrooms was considerably smaller, ranging from 0.03% to 0.1%. The first-grade IUCCs were in the typical range of 0.05 to 0.20 in educational research (e.g., Raudenbush, Liu, & Congdon, 2004; Snijders & Boskers, 1999). Given the relatively high IUCCs associated with first-grade classrooms, we retained the 2-level cross-classified binary logistic regression models to better estimate the standard errors associated with the child-level predictors.

Table 6 shows results of the base model (Model 3) and model extensions including WIF_N, WIF_B, and DA as predictors, respectively (Models 4–6). The base model including WIF_N-screen, phonemic awareness, rapid naming skill, and oral vocabulary provided adequate classification accuracy, with sensitivity of .90 (designated), specificity of .85, and AUC of .948. In addition, all predictors made a unique contribution to predicting RD class membership in the presence of the other predictors. Adding WIF_N progress monitoring, WIF_B progress monitoring, or DA significantly improved classification accuracy by decreasing the number of false positives in the base model by 8, 17, and 9, respectively, and by increasing AUC estimates to .953–.963. WIF_N-level, WIF_B-level, WIF_B-slope, and DA each uniquely added to the prediction accuracy of the base model. In contrast, although Compton et al. (2006) reported the combination WIF_N-level and WIF_N-slope significantly improved classification accuracy over the base model, neither was a unique predictor of RD in the presence of the other predictors. The addition of WIF_N progress monitoring resulted in WIF_N-screen and RDN no longer being unique predictors of RD. Adding WIF_B progress monitoring rendered WIF_N-screen, RDN, and sound matching predictors nonsignificant. In contrast, all screening variables remained significant predictors in the model, including DA. Finally, differences in classification accuracy among the three extension models (4–6) were not statistically significant. Thus, adding any of these three additional measures to the base model improved classification accuracy and resulted in specificity rates approaching .90.

Table 7 displays results of the base model (Model 7) and model extensions adding RR (Model 8) and ORF (Models 9). On this somewhat smaller sample, the base model again provided adequate classification accuracy with sensitivity of .90 (designated), specificity of .87, and AUC of .950. However, adding RR or ORF to the base model decreased the number of false positives by 0 and 2 respectively, and neither was statistically significant. ORF, but not RR, was a significant predictor of RD status; however, this was not adequate to help the overall model classify RD children better than the base model. Thus adding passage level reading measures did not increase the predictive accuracy of the base model and therefore do not warrant inclusion within screening. The effects for ORF over the base model were similar when using the entire sample of 355 children.

Two-Step Gated Procedure: Measures to Identify True Negatives

In the 2-step gated procedure, we attempted to identify single word-level reading measures that would remove true negatives from the Step 2 screening activities. To accomplish this, for sight word efficiency, phonemic decoding efficiency, word identification, and word attack, we identified the standardized cut-point such that all true positives would complete the Step 2 screening activities. We then calculated a 99% confidence interval around that score and used the upper score of the interval to set the cut-point. This cut-point ensured a 99% chance that true positives would score below this value. Table 8 presents the cut-points, number of true negatives scoring above the cut-point, and the proportion of children eliminated from the sample. The *z* test for proportions was significant for each measure, signifying that each measure significantly decreased the proportion of children needing the full screening battery. In addition, when phonemic decoding efficiency was used as the Step 1 screener, it significantly reduced the proportion of children requiring the full screening battery over sight word efficiency ($z = 3.20$), word identification ($z = 7.20$), and word attack ($z = 4.25$). Results suggest that using a standard score cut of 108 on phonemic decoding efficiency to eliminate true negatives from Step 2 screening activities reduced the overall number of children who require the full screening battery by 43.4%, without changing the overall classification accuracy of the prediction model. It is important to note that within a 2-step gating process the overall classification model should be developed first for the entire sample; then cut-points to eliminate true negatives should be determined. Developing the

models in the opposite order results in a less accurate overall classification model with elevated false positive rates.

Discussion

The major purposes of the present study were to identify measures that, when added to a base first-grade screening model, would remove false positives and to also examine the tenability of a 2-step screening procedure to increase screening efficiency. Toward that end, we began by replicating the Compton et al. (2006) classification model with a new population of first-grade children, which included students with a broader range of initial reading performance and students who would be the recipient of reading instruction influenced by the No Child Left Behind reforms. In these ways, the present study differs from and is more ambitious than the only other screening cross-validation effort, that of O'Connor and Jenkins (1999). Moreover, by sampling more broadly in the replication sample, we gained insight into how sensitivity and specificity differ at varying points on the achievement continuum. In general, the original model, which relied on first-grade screeners of WIF_N-screen, RDN, SM, OV, WIF_N-level, WIF_N-slope to predict end of second-grade RD status, replicated well. The number of false negatives across the original and replication models was low; thus, we conclude that false negative rates are of little consequence when predicting early RD, at least when a broad screening battery is used. At the same time, other work (e.g., Compton, Fuchs, Fuchs, Elleman, & Gilbert, 2008; Leach, Scarborough, & Rescorla, 2003; Lipka, Lesaux, & Siegel, 2006) suggests the presence of children who exhibit late-emerging RD who are missed by early screening efforts and therefore who constitute false negatives. This presumably would become more evident with longer follow-up.

The AUC was slightly higher for the replication sample, which is unusual in a replication. Differences in the populations sampled in the original and present studies may account for the overall improvement in model performance. The present model accurately classified HSE and ASE children, who were omitted from the original study that was used to develop the prediction model. In contrast, the LSE children in the present sample were not classified as adequately, that is, specificity fell below .80 (when sensitivity was held at .90). Results indicate that false positives in the population of LSE children represent the primary issue. We are unsure why the LSE sample failed to replicate well. Changes in the population that affect the base rate of RD, screening scores, or the covariance between RD occurrence and screening performance can disrupt the models. Perhaps the instructional initiatives associated with No Child Left Behind and Reading First created a shift in the relative importance of the screening measures. It could also be that demographic changes across the 5-year period between studies affected the prediction model. In any case, these results indicate that schools will need to periodically adjust their screening model coefficients by recalibrating the models using the most current school or district data.

To extend the focus of the present study toward a 2-step screening procedure, we then explored whether measures added to the base screening battery would help decrease false positives. We employed multilevel modeling to partition variance associated with RD prediction between child- and classroom-levels. This was done to estimate and control for the influence of classroom membership on RD prediction through the estimation of IUCCs and to provide better estimates of the child-level predictor standard errors for estimating *p*-values. First-grade IUCCs ranged from .078 to .116; second-grade ICCs were smaller (.0003 to .001). The estimated first-grade IUCCs are similar to those previously reported (Raudenbush et al., 2004; Snijders & Boskers, 1999), but second-grade estimates fall well below the expected range. We offer two explanations for the discrepancies between first- and second-grade IUCCs. It may be that the dispersion of children from 56 first-grade

classrooms into 105 second-grade classrooms led to few children within any given second-grade classroom, thereby depressing estimation of second-grade influence on RD designation. Or it may be that first-grade classroom membership, with its primary focus on reading instruction, accounts for more classroom influence on later RD. In either case, the IUCCs associated with first-grade classroom membership are large enough to justify further exploration into classroom-level factors that account for variance in RD prediction.

A clear message emerged from our attempts to identify additional measures for limiting false positives within the context of a second stage of screening. Measures designed to directly assess (progress monitoring) or forecast (DA) children's response to classroom instruction added significantly to prediction accuracy by reducing false positives. By contrast, measures designed to assess children's ability to read passages, whether focused on accuracy (RRs) or fluency (ORF), added little to the prediction models by way of limiting false positives. The failure of RRs to improve classification accuracy must be evaluated with caution since it was the only measure collected without fidelity data. There is evidence from Allington (1976) suggesting that teacher administration of RRs can be unreliable. DA, WIF_N progress monitoring, and WIF_B progress monitoring were each found to significantly improve classification accuracy, with no statistical advantage found between the three measures, indicating that the addition of any of the three measures in a second screening step increases predictive accuracy. It is up to the user to decide whether two 1-min probes administered weekly for five weeks or a single DA assessment (lasting 20–30 min) is more efficient (see Jenkins et al., 2007). These results replicate claims by Compton et al. (2006) that WIF_N progress monitoring can improve classification accuracy and further extends the study by demonstrating that WIF_B progress monitoring and DA likewise enhance classification.

The present study also provided interesting contrasts between the predictive utility of WIF_N and WIF_B slope estimates. WIF_B-slope was a unique and significant predictor of future RD status; WIF_N-slope was not. The lack of WIF_N-slope to uniquely predict future RD status replicates Compton et al. (2006). The two WIF measures are identical in form, but differ in the range of words sampled: WIF_N samples the 100 most frequent words; WIF_B, the 500 most frequent words. The sampling procedure used to construct WIF_B resulted in lists containing more diverse and difficult words (as evidenced by lower level and slope estimates compared WIF_N). It may be that WIF_B-slope better estimates children's successful encounter and learning of the diverse set of words that constitute early reading instruction. That WIF_B-slope, not WIF_N-slope, was a unique predictor of future RD highlights a disconnect in the literature regarding the utility of slope measures in predicting future reading skill and RD status. For instance, Schatschneider, Wagner, and Crawford (2008) showed that slope estimated using ORF measured in September, December, February, and April of first grade did not predict, above and beyond end-of-first-grade ORF level, performance on a standardized measure of reading comprehension administered at the end of second grade. By contrast, Fuchs et al. (2004) reported that fall of first grade WIF-slope was a unique predictor of end of year standardized reading performance even after controlling for initial WIF level. Differences between the criterion measures, progress-monitoring assessments, and participant sampling techniques make it impossible to isolate the cause of the discrepancy across studies. Thus, the literature does not allow us to formulate conclusions about the predictive utility of slope. With that in mind, what makes the results of this study notable is that WIF_B-slope was a significant predictor of future RD status in the presence not only of WIF_B-level but also of WIF_N-screen, rapid digit naming, sound matching, and oral vocabulary. This finding allows us to speculate that sampling procedures affect the usefulness of slope as a predictor and that sampling more widely from the corpus of words to construct WIF likely results in a better slope estimate and therefore a better predictor of response to classroom instruction and RD status.

Finally, we further extended the screening literature by examining the efficacy of using 2-step screening procedures, hoping that it might yield an effective overall screening system. In such a 2-step procedure, all children are administered a single, brief measure, and only children who score within the risk range on that initial measure complete the longer screening battery. In the present study, we pitted various measures against each other to identify which show the most promise in eliminating true negatives in the first step of the 2-step screening processes, thereby limiting the number of children requiring the full screening battery. Overall, the measure of phonemic decoding efficiency eliminated the greatest number of true negatives (43.4% of the sample) from screening. Phonemic decoding efficiency significantly outperformed measures of sight word efficiency, word identification, and word attack in reducing the sample to be screened further. Moreover, as discussed, measures of students' capacity to learn from classroom reading instruction, as represented by WIF progress monitoring or DA, appear to lend utility at a second screening step. We therefore recommend the use of 2-step gated procedures as a means to increase the efficiency of 1-step universal screening procedures.

Future research should extend this work by exploring the relations between child- and classroom-level predictors that influence RD risk. It is probable that other types of measures that assess or estimate children's response to classroom reading instruction can be used to improve classification accuracy. One source of screening information that has thus far been ignored is teacher judgment, which might be used in the first or second step of a gated screening procedure to increase classification accuracy rates. Teachers have the unique opportunity to directly observe children's response to instruction. Thus, we encourage researchers to broaden the scope of child-level measures for gauging response to classroom instruction with the intent of eliminating false positives and negatives within the context of a 2-step screening process. At the level of the classroom, the effects of individual differences in teacher effectiveness and variations in curricular focus still require exploration as predictors of classroom-level variance in forecasting RD risk. There is evidence to suggest that RD risk can be reduced when systematic, high quality instruction in phonemic awareness and phonemic decoding skills, fluency in word recognition and text processing, construction of meaning, vocabulary, spelling, and writing is provided by the classroom teacher (e.g., Torgesen, 2002b; Vadasy, Sanders, & Peyton, 2006). In addition to curriculum, growing evidence suggests that children who are at risk of developing RD require reading instruction that is more explicit, intense, and comprehensive than the instruction required by the majority of children (see Foorman & Torgesen, 2001), and there are some indicators that teachers vary in the ability to provide focused and explicit instruction to children at risk for developing RD (see Moats, 2009). Thus, the context of the classroom should receive greater attention as a source of variance in explaining RD risk (e.g., Foorman, & Nixon, 2006; Foorman, York, Santi, & Francis, 2008; Skindrud, & Gersten, 2006). Combining both child- and classroom-level effects in future prediction models should greatly extend our understanding of how classroom practices mediate child-level risk for RD.

References

- Allington RL. Teacher test achievement and taped micro-teaching performance. *Improving Human Performance* 1976;5(1):7–14.
- Bean RM, Cassidy J, Grumet JE, Shelton DS, Wallis SR. What do reading specialists do? Results from a national survey. *The Reading Teacher* 2002;55:736–744.
- Caffrey, E. Unpublished doctoral dissertation. Vanderbilt University; Nashville, Tennessee: 2005. A comparison of dynamic assessment and progress monitoring in the prediction of reading achievement for students in kindergarten and first grade.

- Campione JC. Assisted testing: A taxonomy of approaches and an outline of strengths and weaknesses. *Journal of Learning Disabilities* 1989;22:151–165. [PubMed: 2708891]
- Campione JC, Brown AL, Ferrara RA, Jones RS, Steinberg E. Differences between retarded and non-retarded children in transfer following equivalent learning performance: Breakdowns in flexible use of information. *Intelligence* 1985;9:297–315.
- Catts H. Early identification of dyslexia: Evidence from a follow-up study of speech-language impaired children. *Annals of Dyslexia* 1991;41:163–177.
- Catts HW, Petscher Y, Schatschneider C, Sittner-Bridges M, Mendoza K. Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities* 2009;42:163–76. [PubMed: 19098274]
- Clay, MM. An observation survey of early literacy achievement. Portsmouth, NH: Heinemann; 1993.
- Compton DL, Fuchs D, Fuchs LS, Bryant JD. Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology* 2006;98:394–409.
- Compton DL, Fuchs D, Fuchs LS, Elleman AM, Gilbert JK. Tracking children who fly below the radar: Latent transition modeling of students with late-emerging reading disability. *Learning and Individual Differences* 2008;18:329–337.
- Day JD, Engelhardt JL, Maxwell SE, Bolig EE. Comparison of static and dynamic assessment procedures and their relation to independent performance. *Journal of Educational Psychology* 1997;89:358–368.
- Deno, SL. Curriculum-based measurement and special education services: A fundamental and direct relationship. In: Shinn, MR., editor. *Curriculum-based measurement: Assessing special children*. The Guildford Press; New York: 1989. p. 1-17.
- Deno SL, Marston D, Shinn MR, Tindal G. Oral reading fluency: A simple datum for scaling reading disability. *Topics in Learning and Learning Disabilities* 1983;2(4):53–59.
- Fawson PC, Ludlow BC, Reutzel DR, Sudweeks RR, Smith JA. Examining the reliability of running records: Attaining generalizable results. *Journal of Educational Research* 2006;100:113–126.
- Foorman BR, Nixon SM. The influence of public policy on reading research and practice. *Topics in Language Disorders* 2006;26(2):157–171.
- Foorman BR, Torgesen J. Critical elements of classroom and small-group instruction promote reading success in all children. *Learning Disabilities Research & Practice. Special Issue: Emergent and Early Literacy: Current Status and Research Directions* 2001;16(4):203–212.
- Foorman BR, York M, Santi KL, Francis D. Contextual effects on predicting risk for reading difficulties in first and second grade. *Reading and Writing* 2008;21(4):371–394.
- Fountas, IC.; Pinnell, GS. *Guided reading: Good first teaching for all children*. Portsmouth, NH: Heinemann; 1996.
- Fuchs LS, Fuchs D. Treatment validity: A unifying concept for the identification of learning disabilities. *Learning Disability Research and Practice* 1998;14:204–219.
- Fuchs LS, Fuchs D. Progress monitoring within a multi-tiered prevention system. *Perspectives* 2007;33(2):43–47.
- Fuchs LS, Fuchs D, Compton DL. Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children* 2004;71:7–21.
- Fuchs LS, Fuchs D, Hamlett CL. Monitoring reading growth using student recalls: Effects of two teacher feedback systems. *Journal of Educational Research* 1989;83:103–111.
- Fuchs LS, Fuchs D, Hosp MD, Jenkins J. Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading* 2001;5:239–259.
- Fletcher JM, Foorman BR, Boudousquie A, Barnes MA, Schatschneider C, Francis DJ. Assessment of reading and learning disabilities a research-based intervention-oriented approach. *Journal of School Psychology* 2002;40:27–63.
- Glover T, Albers C. Considerations for evaluating universal screening assessments. *Journal of School Psychology* 2007;45:117–135.

- Good RH, Simmons D, Kame'enui E, Chard D. The importance and 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.
- Grigorenko EL, Sternberg RJ. Dynamic testing. *Psychological Bulletin* 1998;124:75–111.
- Handley JA, McNeil BJ. A method of comparing the areas under receiver operating curves derived from the same cases. *Radiology* 1983;148:839–843. [PubMed: 6878708]
- Hosp MK, Fuchs LS. Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review* 2005;34:9–26.
- Jenkins, JR. Candidate measures for screening at-risk students; Paper presented at the Conference on Response to Intervention as Learning Disabilities Identification, Sponsored by the National Research Center on Learning Disabilities; Kansas City, MO. 2003 Dec.
- Jenkins JR, Heliotis J, Haynes M, Stein M, Beck K. Does “passive learning” account for disabled readers’ comprehension deficits in ordinary reading situations? *Learning Disability Quarterly* 1986;9:69–76.
- Jenkins JR, Hudson RF, Johnson ES. Screening for service delivery in a response-to-intervention (RTI) framework. *School Psychology Review* 2007;36:582–600.
- Jenkins, JR.; O’Connor, RE. Early identification and intervention for young children with reading/learning disabilities. In: Bradley, R.; Danielson, L.; Hallahan, DP., editors. *Identification of learning disabilities: Research to practice*. Mahwah, NJ: Erlbaum; 2002. p. 99-149.
- Kaminski RA, Good RH III. Toward a technology for assessing basic literacy skills. *School Psychology Review* 1996;25:215–227.
- Leach J, Scarborough H, Rescorla L. Late-emerging reading disabilities. *Journal of Educational Psychology* 2003;95:211–224.
- Lipka O, Lesaux NK, Siegel LS. Retrospective analyses of the reading development of Grade 4 students with reading disabilities: Risk status and profiles over 5 years. *Journal of Learning Disabilities* 2006;39:364–378. [PubMed: 16895160]
- MaCardle P, Scarborough HS, Catts HW. Predicting, explaining, and preventing children’s reading difficulties. *Learning Disability Research and Practice* 2001;16:230–239.
- Moats L. Knowledge foundations for teaching reading and spelling. *Reading and Writing* 2009;22(4): 379–399.
- Murray BA, Smith KA, Murray GG. The test of phoneme identities: Predicting alphabetic insight in prealphabetic readers. *Journal of Literacy Research* 2000;32:421–477.
- O’Connor RE, Jenkins JR. The prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading* 1999;3:159–197.
- Rathvon, N. *Early reading assessment: A practitioner’s handbook*. New York: Guilford Press; 2004.
- Raudenbush, SW.; Bryk, AS. *Hierarchical linear models: Applications and data analysis methods* second edition. London: Sage; 2002.
- Raudenbush, SW.; Liu, X.; Congdon, R. Documentation for the “Optimal Design” software. 2004. *Optimal Design for longitudinal and multilevel research*.
- Resing, WCM. Measuring inductive reasoning skills: The construction of a learning potential test. In: Hamers, JHM.; Sijtsma, K.; Ruijssenaars, AJJM., editors. *Learning potential testing*. Amsterdam: Swets & Zeitlinger; 1993. p. 219-242.
- Roehrig AD, Petscher Y, Nettles S, Hudson RF, Torgesen JK. Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology* 2008;46:343–366. [PubMed: 19083363]
- Scarborough, HS. Early identification of children at risk for reading disabilities: Phonological awareness and some other promising predictors. In: Shapiro, BK.; Accardo, PJ.; Capute, AJ., editors. *Specific reading disability: A view of the spectrum*. Timonium, MD: York Press; 1998. p. 75-119.
- Schatschneider C, Wagner RK, Crawford EC. The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning & Individual Differences* 2008;18(3): 308–315.

- Schilling SG, Carlisle JF, Scott SE, Zeng J. Are fluency measures accurate predictors of reading achievement? *Elementary School Journal* 2007;107:429–448.
- Foresman, Scott. *Scott Foresman Reading K-2 Individual Reading Inventory and Running Record*. Pearson School; Chandler, AZ: 2006.
- Skindrud K, Gersten R. An evaluation of two contrasting approaches for improving reading achievement in a large urban district. *The Elementary School Journal* 2006;106(5):389–407.
- Snijders, TAB.; Boskers, RJ. *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage; 1999.
- Spector JE. Predicting progress in beginning reading: Dynamic assessment of phonemic awareness. *Journal of Educational Psychology* 1992;84:353–363.
- Swanson HL. Effects of dynamic testing on the classification of learning disabilities: The predictive and discriminant validity of the Swanson-Cognitive Processing Test. *Journal of Psychoeducational Assessment* 1995;13:204–229.
- Swanson HL, Howard CB. Children with reading disabilities: Does dynamic assessment help in the classification? *Learning Disability Quarterly* 2005;28:17–34.
- Swets JA. The science of choosing the right decision threshold in high-stake diagnostics. *American Psychologist* 1992;47:522–532. [PubMed: 1595983]
- Torgesen, JK. Empirical and theoretical support for direct diagnosis of learning disabilities by assessment of intrinsic processing weaknesses. In: Bradley, R.; Danielson, L.; Hallahan, DP., editors. *Identification of learning disabilities: Research to practice*. Mahwah, NJ: Erlbaum; 2002a. p. 565-613.
- Torgesen JK. The prevention of reading difficulties. *Journal of School Psychology* 2002b;40:7–26.
- Torgesen, JK.; Wagner, RK.; Rashotte, CA. *Test of Word Reading Efficiency*. Austin, TX: Pro-Ed; 1997.
- Vadasy PF, Sanders EA, Peyton JA. Code-oriented instruction for kindergarten students at risk for reading difficulties: A randomized field trial with paraeducator implementers. *Journal of Educational Psychology* 2006;98(3):508–528.
- VanDerHeyden AM, Witt JC, Gilbertson DA. Multi-year evaluation of the effects of a response to intervention (RTI) model on identification of children for special education. *Journal of School Psychology* 2007;45:225–256.
- Wagner, RK.; Torgesen, JK.; Rashotte, CA. *Comprehensive Test of Phonological Processing*. Austin, TX: Pro-Ed; 1999.
- Wagner RK, Torgesen JK, Rashotte CA, Hecht SA, Barker TA, Burgess SR, Donahue J, Garon T. Changing causal relations between phonological processing abilities and word-level reading as children develop from beginning to fluent readers: A five-year longitudinal study. *Developmental Psychology* 1997;33:468–479. [PubMed: 9149925]
- Woodcock, RW. *Woodcock Reading Mastery Test–Revised/Normative Update*. Circle Pines, MN: AGS; 1998.
- Woodcock, RW.; McGrew, KS.; Mather, N. *Woodcock-Johnson III Tests of Psychoeducational Ability*. Itasca, IL: Riverside; 2001.
- Zeno, SM.; Ivens, SH.; Millard, RT.; Duvvuri, R. *The educator’s word frequency guide*. Touchstone Applied Science Associates; 1995.

Table 1

Demographic and First-Grade Screening Performance for the Sample that Moved Versus the Sample that Completed the Study on Demographic Variables and First-Grade Screening Measures

Variable	Moved		Completed Study		χ^2	<i>p</i>	
	<i>f</i>	%	<i>f</i>	%			
Sex	(n=120)		(n=348)		3.6528	.056	
Male	52	43.33	186	53.45			
Female	68	56.67	162	46.55			
Race	(n=119)		(n=349)		0.0758	.963	
African American	53	44.54	160	45.98			
Caucasian	48	40.34	136	39.08			
Other	18	15.13	53	15.23			
Lunch Status	(n=112)		(n=325)		1.3221	.250	
No Free lunch	44	39.29	148	45.54			
Free lunch	68	60.71	177	54.46			
Entry Status	(n=131)		(n=354)		6.8436	.033	
Low Student Entry	96	73.28	214	60.45			
High Student Entry	17	12.98	66	18.64			
Average Student Entry	18	13.74	74	20.90			
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>d</i>	<i>t</i> (483)	<i>p</i>
	(n=131)		(n=354)				
Rapid Letter Naming	38.98	16.85	43.01	16.54	2.37	40.66	.0182
Rapid Sound Naming	26.11	12.07	29.74	12.27	2.91	27.61	.0038
WIF_N Screen	18.27	20.60	24.72	23.55	2.77	23.92	.0059

Note. WIF_N Screen = Word Identification Fluency screening measure.

Table 2

Differences in Screening Measures by Entry Status

Measure	LSE (n = 214)		ASE (n = 75)		HSE (n = 66)		F	d [†]
	M	SD	M	SD	M	SD		
RLN	34.53	14.04	53.14	10.52	59.14	9.84	124.99*	1.41
RLS	25.14	11.32	36.00	10.79	37.65	9.44	48.79*	0.97
WIF_N Screen	9.51	5.99	29.81	5.26	68.35	11.78	1662.87*	3.49

Note. Low, Average, and High study entry groups were identified using Latent Class Analysis. LSE = Low Study Entry; ASE = Average Study Entry; HSE = High Study Entry; RLN = Rapid Letter Name; RLS = Rapid Letter Sound; WIF= Word Identification Fluency.

[†] Effect size calculated between ASE and LSE groups.

* p<.001.

Table 3

Correlation Coefficients Among First-Grade Predictors and the Second-Grade Running Record and Composite Outcome of Reading Performance

Measure	1	2	3	4	5	6	7	8	9	10	11	12
1. RDN ^a	—											
2. Sound Matching ^a	.41**	—										
3. Oral Vocabulary ^a	.32**	.57**	—									
4. WIF_N Screen ^a	.49**	.50**	.47**	—								
5. WIF_N INT ^a	.63**	.60**	.49**	.87**	—							
6. WIF_N SLP ^a	.24**	.12*	.05	-.04	.21**	—						
7. WIF_B INT ^a	.51**	.57**	.50**	.94**	.91**	.05	—					
8. WIF_B SLP ^a	.51**	.50**	.42**	.75**	.82**	.30**	.83**	—				
9. DA ^a	-.45**	-.58**	-.48**	-.59**	-.65**	-.11*	-.62**	-.54**	—			
10. ORF ^a	.49**	.52**	.48**	.94**	.86**	-.01	.96**	.92**	-.65**	—		
11. Running Record ^a	.41**	.50**	.51**	.82**	.72**	-.03	.84**	.63**	-.55**	.85**	—	
12. Composite Reading ^b	.59**	.68**	.63**	.72**	.83**	.21**	.79**	.71**	-.69**	.68**	.66**	—
Mean	8.66	9.78	95.96	24.87	41.92	0.47	25.64	0.36	9.92	26.83	8.13	
SD	2.25	2.19	15.18	23.59	26.18	0.53	22.46	0.49	3.31	24.26	7.65	

Note. $N = 355$. RDN = Rapid Digit Naming; WIF_N = Word Identification Fluency_Narrow; WIF_B = Word Identification Fluency_Broad; INT = Intercept at assessment wave 5; SLP = Slope across 5 assessment waves; DA = Dynamic Assessment. Correlations for Running Record are based on $n=320$.

^aFirst-grade screening measure.

^bSecond-grade outcome.

* $p < .05$.

** $p < .001$.

Table 4
Coefficients and Classification Indices for the Original and Replicated Logistic Regression Models

Measure	B	SE	Wald	p	TN	FN	TP	FP	Hit rate	Sensitivity	Specificity	AUC	SE
Model 1: Original Model (n = 206)													
WIF_N_Screen	-0.210	0.180	1.369	0.244	151	3	17	30	83.6	90.0	82.9	0.913	0.032
RDN	-0.221	0.146	2.280	0.131									
Sound Matching	-0.567	0.199	8.123	0.004									
Oral Vocabulary	-0.010	0.025	0.160	0.690									
WIF_N_Level	0.017	0.192	0.008	0.930									
WIF_N_Slope	-0.501	0.852	0.348	0.557									
Constant	7.033	2.534	7.728	0.006									
Model 2: Replication (n = 355)													
WIF_N_Screen	-0.210				253	5	49	48	84.8	90.7	84.0	0.925	0.015
RDN	-0.221												
Sound Matching	-0.567												
Oral Vocabulary	-0.010												
WIF_N_Level	0.017												
WIF_N_Slope	-0.501												
Constant	7.033												

Note. Hit rate, sensitivity, and specificity are expressed as percentages. WIF_N_Screen = Word Identification Fluency screening measure. WIF_N_Level = Level of Word Identification Fluency_Narrow at week 5; WIF_N_Slope = Slope of Word Identification Fluency_Narrow over 5 weeks; TN = true negatives; FN = false negatives; TP = true positives; FP = false positives; Hit rate = (TP + TN)/N; sensitivity = TP/(TP + FN); specificity = TN/(TN + FP); AUC = area under the curve; SE = standard error of AUC.

Table 5

Variance Components of the Cross-Classified Logistic Regression Models

Component	Parameter	Estimate	IUCC	df	χ^2	p
Model 3: Extension Base Model ($n = 355$)						
Between-1st-grade classrooms	τ_{600}	0.2876	0.0802	55	50.74	> 0.500
Between-2nd-grade classrooms	τ_{c00}	0.0037	0.0010	104	89.40	> 0.500
Model 4: Extension using WIFA ($n = 355$)						
Between-1st-grade classrooms	τ_{600}	0.3158	0.0876	55	49.36	> 0.500
Between-2nd-grade classrooms	τ_{c00}	0.0041	0.0011	104	73.82	> 0.500
Model 5: Extension using WIFB ($n = 355$)						
Between-1st-grade classrooms	τ_{600}	0.4318	0.1160	55	53.44	> 0.500
Between-2nd-grade classrooms	τ_{c00}	0.0026	0.0007	104	71.67	> 0.500
Model 6: Extension using DA ($n = 355$)						
Between-1st-grade classrooms	τ_{600}	0.3258	0.0901	55	51.82	> 0.500
Between-2nd-grade classrooms	τ_{c00}	0.0025	0.0007	104	81.19	> 0.500
Model 7: Base Model using ($n = 320$)						
Between-1st-grade classrooms	τ_{600}	0.39912	0.1089	51	50.83	> 0.500
Between-2nd-grade classrooms	τ_{c00}	0.0013	0.0003	98	77.58	> 0.500
Model 8: Extension using RR ($n = 320$)						
Between-1st-grade classrooms	τ_{600}	0.4009	0.1087	51	50.83	> 0.500
Between-2nd-grade classrooms	τ_{c00}	0.0013	0.0004	98	77.50	> 0.500
Model 9: Extension using ORF ($n = 320$)						
Between-1st-grade classrooms	τ_{600}	0.2765	0.0776	51	46.95	> 0.500
Between-2nd-grade classrooms	τ_{c00}	0.001	0.0003	98	61.21	> 0.500

Note. IUCC = Intraunit correlation coefficient, $\tau_{600}/(\tau_{600} + \pi^2/3)$ and $\tau_{c00}/(\tau_{c00} + \pi^2/3)$.

Table 6
Coefficients and Classification Indices for the Extension Models using Progress Monitoring and Dynamic Assessment

Measure	B	SE	Wald	p	TN	FN	TP	FP	HR	Sensitivity	Specificity	AUC	SE	z
Model 3: Extension Base Model (<i>n</i> = 355)														
WIF Screen	-0.120	0.037	10.693	0.002	257	5	49	44	86.0	90.7	85.4	0.948	0.012	—
RDN	-0.246	0.095	6.750	0.010										
Sound Matching	-0.372	0.115	10.537	0.002										
Oral Vocabulary	-0.051	0.017	9.193	0.003										
Constant	9.320	1.818	26.276	0.000										
Model 4: Extension using WIF_N (<i>n</i> = 355)														
WIF Screen	0.028	0.059	0.228	0.633	265	5	49	36	88.2	90.7	88.0	0.957	0.010	2.51*
RDN	-0.146	0.101	2.117	0.147										
Sound Matching	-0.300	0.124	5.856	0.016										
Oral Vocabulary	-0.053	0.018	8.521	0.004										
WIF_N Level	-0.128	0.041	9.885	0.002										
WIF_N Slope	-0.264	0.701	0.142	0.706										
Constant	9.417	1.982	22.610	0.000										
Model 5: Extension using WIF_B (<i>n</i> = 355)														
WIF Screen	0.067	0.059	1.300	0.256	274	5	49	27	90.7	90.7	91.0	0.963	0.009	3.29*
RDN	-0.170	0.104	2.676	0.102										
Sound Matching	-0.207	0.128	2.628	0.106										
Oral Vocabulary	-0.056	0.019	9.205	0.003										
WIF_B Level	-0.416	0.102	16.695	0.000										
WIF_B Slope	5.289	1.958	7.295	0.008										
Constant	9.856	2.025	23.688	0.000										
Model 6: Extension using DA (<i>n</i> = 355)														
WIF Screen	-0.102	0.038	7.285	0.008	264	5	49	37	88.2	90.7	87.7	0.953	0.011	2.24*
RDN	-0.215	0.098	4.814	0.029										
Sound Matching	-0.292	0.122	5.770	0.017										
Oral Vocabulary	-0.052	0.017	8.934	0.003										
DA	0.211	0.099	4.541	0.034										

Measure	B	SE	Wald	p	TN	FN	TP	FP	HR	Sensitivity	Specificity	AUC	SE	z
Constant	5.757	2.407	5.722	0.017										

Note. Hit rate, sensitivity, and specificity are expressed as percentages. WIF Screen = Word Identification Fluency screening measure; RDN = rapid digit naming; DA = dynamic assessment; RR = running records level; WIF_N Level = Level of Word Identification Fluency_Narrow at week 5; WIF_N Slope = Slope of Word Identification Fluency_Narrow over 5 weeks; WIF_B Level = Level of Word Identification Fluency_Broad at week 5; WIF_B Slope = Slope of Word Identification Fluency_Broad over 5 weeks; TN = true negatives; FN = false negatives; TP = true positives; FP = false positives; HR = Hit rate (TP + TN)/N; sensitivity = TP/(TP + FN); specificity = TN/(TN + FP); ROC = receiver operating characteristic; AUC = area under the curve. z = test for difference in AUC compared to Model 3; Extension Base Model.

* p < 0.05

Table 7
Coefficients and Classification Indices for the Extension Models using Paragraph Reading Measures

Measure	B	SE	Wald	p	TN	FN	TP	FP	HR	Sensitivity	Specificity	AUC	SE	z
Model 7: Base Model (n = 320)														
WIF Screen	-0.120	0.038	9.809	0.002	235	5	45	35	87.2	90.0	87.0	0.950	0.012	—
RDN	-0.243	0.100	5.852	0.016										
Sound Matching	-0.380	0.122	9.672	0.002										
Oral Vocabulary	-0.047	0.018	6.985	0.009										
Constant	9.013	1.903	22.439	0.000										
Model 8: Extension using RR (n = 320)														
WIF Screen	-0.122	0.042	8.375	0.005	235	5	45	35	87.2	90.0	87.0	0.950	0.012	0.00
RDN	-0.244	0.101	5.852	0.016										
Sound Matching	-0.381	0.123	9.610	0.003										
Oral Vocabulary	-0.047	0.018	6.959	0.009										
RR	0.008	0.089	0.008	0.928										
Constant	9.025	1.908	22.382	0.000										
Model 9: Extension using ORF (n = 320)														
WIF Screen	-0.031	0.050	0.393	0.531	237	5	45	33	87.9	90.0	87.8	0.951	0.011	0.22
RDN	-0.193	0.103	3.486	0.062										
Sound Matching	-0.332	0.123	7.279	0.008										
Oral Vocabulary	-0.048	0.019	6.355	0.012										
ORF	-0.048	0.019	6.325	0.013										
Constant	8.697	1.979	19.307	0.000										

Note. Hit rate, sensitivity, and specificity are expressed as percentages. WIF Screen = Word Identification Fluency screening measure; RDN = Rapid Digit Naming; RR = Running Record; ORF = Oral Reading Fluency; TN = true negatives; FN = false negatives; TP = true positives; FP = false positives; HR = Hit rate (TP + TN)/N; sensitivity = TP/(TP + FN); specificity = TN/(TN + FP); ROC = receiver operating characteristic; AUC = area under the curve. z = test for difference in AUC compared to Model 7; Extension Base Model. Models 8 and 9 were compared to Model 7.

Table 8

Number of True Negatives Eliminated, with Associated Cut-scores, Using Timed and Untimed Measures of Word-level Reading Skill in a Two-Step Gated Procedure

Measure	Cut-score	TN Eliminated	Percent Reduction
Sight Word Efficiency	111	144	40.5*
Phonemic Decoding Efficiency	108	154	43.4*
Word Identification	117	103	29.0*
Word Attack	113	136	38.3*

Note. $N = 355$. Cut-scores represented as standard scores. TN = true negatives

* $p < .001$