Hammill Institute on Disabilities

Assessment for Effective Intervention
2021, Vol. 46(2) 99–109

DOI: 10.1177/1534508419857234
aei.sagepub.com
SAGE

# Early Screening for Decoding- and Language-Related Reading Difficulties in First and Third Grades

**Rebecca D. Silverman, EdD[1], Daniel McNeish, PhD[2], Deborah L. Speece, PhD[3], and Kristin D. Ritchey, PhD[4]**

## Abstract

The present study investigated the identification of end of first grade ($n$ = 125) and end of third grade ($n$ = 77) reading comprehension difficulties using beginning of first grade decoding-related and language-related predictors. Reading comprehension was defined using a composite of three standardized reading comprehension measures. Students at or below the 25th percentile on the reading comprehension composite were considered to have difficulty with reading comprehension. Sight word reading efficiency and sentence repetition predicted end of first grade reading comprehension difficulties. Sight word reading efficiency, sentence repetition, and oral discourse predicted end of third grade reading comprehension difficulties. Both screening batteries yielded high classification accuracy (area under the receiver operating characteristic curve [AUC] = 94.7% in Grade 1 and AUC = 90.5% in Grade 3). Results are discussed in light of the research base on early screening for reading comprehension difficulties.

## Keywords

early literacy, reading/literacy, screening/benchmarking

Research shows that early screening can identify students at risk of reading difficulties later in school (e.g., Compton, Fuchs, Fuchs, & Bryant, 2006; O'Connor & Jenkins, 1999; Speece et al., 2011). The prevalence of early screening in schools has increased since the reauthorization of the Individuals with Disabilities Education Improvement Act (2004), which allows for Response to Intervention (RTI) to be used in identifying children with specific learning disabilities. A typical model of RTI includes three tiers of assessment and instruction. Tier 1 typically includes (a) universal screening (i.e., assessment of all students) to identify children at risk of difficulties, (b) quality evidence-based general education instruction, and (c) frequent progress monitoring to identify students who are not demonstrating expected growth. Students who are at risk of reading difficulties and who do not respond to evidence-based general education instruction as demonstrated by expected growth are provided with intervention in subsequent tiers. As the first step in the process of identifying difficulties, universal screening is an essential component of RTI. Although there has been extensive research on decoding-related predictors (e.g., phonological awareness, letter knowledge, and word reading) of reading difficulties (e.g., Compton et al., 2006; O'Connor & Jenkins, 1999), there has been relatively less research investigating whether language-related assessments would facilitate identifying students at risk of reading difficulties.

However, as reviewed by Florit and Cain (2011), research has established that reading comprehension is a product of both decoding- *and* language-related skills. The relative importance of decoding-related and language-related skills seems to change over time (e.g., Adlof, Catts, & Little, 2006). Initial reading activities place high demands on decoding-related skills, but, as students become more efficient decoders, language-related skills become more influential in reading comprehension. To effectively identify students who may have difficulty in decoding-related skills, language-related skills, or both, early screening batteries may need to include language measures in addition to decoding measures. The purpose of this study was to investigate whether assessment of a range of language-related skills in

[1]Stanford University, CA, USA
[2]Arizona State University, Tempe, AZ, USA
[3]Virginia Commonwealth University, Richmond, VA, USA
[4]University of Delaware, Newark, DE, USA

**Corresponding Author:**
Rebecca D. Silverman, Stanford University, 205 Barnum Center, 505 Lasuen Mall, Stanford, CA 94305, USA.
Email: rdsilver@stanford.edu

addition to decoding-related skills at the beginning of first grade would contribute to the identification of students experiencing reading comprehension difficulties at the end of first and third grades. Further details of the study are provided following a brief overview of the research base.

To effectively identify students, universal screening batteries must assess the skills essential to predicting reading difficulties and have high classification accuracy (Speece et al., 2011). Screening batteries have high classification accuracy when they identify most of the students who would ultimately experience reading difficulty (i.e., true positives) and minimize over-identification (i.e., false positives) and under-identification (i.e., false-negatives) so that resources can be targeted to students most in need of intervention (O'Connor & Jenkins, 1999). Thus, research on universal screening has commonly investigated which combination of measures yields the highest classification accuracy in terms of (a) sensitivity (i.e., the ability of the screener to correctly identify those students with reading difficulty) and (b) specificity (i.e., the ability of the screener to correctly identify those students without reading difficulty).

As noted, decoding-related measures such as phonological processing, letter-sound knowledge, and word reading fluency have surfaced as important skills to assess in universal screening batteries. For example, O'Connor and Jenkins (1999) found that phoneme segmentation and rapid letter naming assessed at the beginning and end of kindergarten and at the beginning of first grade were reliable predictors of a reading disability at the end of first grade. In addition, Speece et al. (2011) identified that two measures of word reading fluency and a teacher rating of reading problems collected in the beginning of first grade accurately identified students with reading challenges at the end of first grade. Recently, Catts, Nielsen, Bridges, Liu, and Bontempo (2015) found that various combinations of measures of letter naming fluency, phonological awareness, rapid naming, and nonword repetition at the beginning of kindergarten accurately identified students with reading difficulties at the end of first grade.

Much of the research that has identified various decoding-related measures as important to identifying reading difficulties has focused primarily on decoding-related predictors or decoding-related outcomes, and these studies have typically been conducted in the early grades when decoding-related abilities tend to dominate the reading process. For example, McNamara, Scissons, and Gutknecth (2011) used kindergarten measures of phonological awareness to identify risk of reading disabilities. Furthermore, in the kindergarten and first grade study by O'Connor and Jenkins (1999), reading difficulties were defined through formal school classification or mean performance on measures of word identification and word attack. Other studies that have identified various decoding-related measures as important to identifying reading difficulties have included one or more language-related predictors alongside several decoding-related predictors or included language-related skills or general measures of reading comprehension in a composite reading outcome that is dominated by decoding-related skills. For example, Compton et al. (2006) included one predictor of oral language alongside five decoding-related predictors in their first grade assessment battery. Furthermore, Speece et al. (2011) included two comprehension measures alongside six decoding-related measures in the reading outcome under investigation in their first grade sample. In studies that include more than one language measure (e.g., Cooper, Roth, Speece, & Schatschneider, 2002), the authors typically assess decoding-related skills as the outcome instead of comprehension ability. Language-related predictors have not surfaced as important in this line of research, but it could be that the strong reliance on decoding-related measures and the focus on reading comprehension in the early grades overshadow the contribution of language-related predictors to reading difficulties.

Other research suggests that language-related skills are important to reading comprehension, especially as students master decoding and encounter evermore complex text in upper elementary school and beyond (e.g., Hoover & Gough, 1990). There is also evidence that students who have early language difficulties may experience later reading problems (Catts, Compton, Tomblin, & Bridges, 2012). Although some students with more severe early language difficulties will be identified through speech-language services, other students with less severe early language delays that could affect later reading achievement may go unidentified. If these children could be identified early and provided with intervention to ameliorate language-related challenges in areas such as vocabulary, syntax, and listening comprehension, attempts to prevent reading difficulties may be more successful. This may be especially important considering that, without intervention, students with lower levels of language skills may not grow as fast as their peers with higher language skills, resulting in a widening gap between children with lower and higher levels of language skills across the elementary school years that could make later intervention more difficult to implement than early intervention (McNamara et al., 2011).

Among studies exploring the relationship between language skills and reading comprehension, there have been divergent findings. For example, Wise, Sevcik, Morris, Lovett, and Wolf (2007) administered a battery of assessments to 279 second and third grade students with reading disabilities in the areas of receptive vocabulary, expressive vocabulary, listening comprehension, prereading skills, and reading comprehension and found that language measures contributed to reading comprehension only through their influence on prereading and word identification skills. The authors acknowledged that the study was limited because it included only one measure of reading comprehension that

may have been more reliant on decoding- rather than language-related skills. Findings from Wise et al. (2007) are contrasted by findings from a recent study by Catts et al. (2015) who investigated the relative importance of language-related measures in predicting reading difficulties from kindergarten through the end of third grade. As part of a larger study, these researchers administered decoding-related measures (i.e., letter knowledge, phonological awareness, rapid automatized naming, and nonword repetition) as well as language-related measures (i.e., receptive and expressive vocabulary, syntax as assessed via sentence imitation, and receptive and expressive narrative language) in kindergarten and reading comprehension in third grade. Vocabulary and narrative language added to the prediction of reading comprehension difficulties over and above the decoding-related measures, suggesting that language-related measures may be important in universal screening to prevent later difficulties in reading comprehension. Findings held even after controlling for second grade word reading. This work builds on other research by Catts et al. (2012) suggesting that more than 13% of children emerge as poor readers in later elementary school and many of these children could have been identified with language-related difficulties in the early elementary years.

The study by Catts et al. (2015) makes an important contribution to the research base on universal screening, but further research is needed to substantiate these findings across samples and additional research is needed to identify the types of language-related assessments that might be most predictive of later reading difficulties. The purpose of the present study is to address these research needs. The present study differs from the Catts et al. (2015) study in several ways. Whereas Catts et al. (2015) used decoding-related and language-related skills at the beginning of kindergarten, we assessed these skills at the beginning of first grade. There is some debate in the research about when is the best time to screen students for reading difficulties. Some evidence suggests that kindergarten screening would be preferable to first grade screening because the earlier students can be identified as at risk of reading problems, and the earlier schools can provide intervention to prevent difficulty (Catts et al., 2015). In contrast, given that many students enter formal school for the first time in kindergarten and, therefore, may have had limited exposure to sounds, letters, and academic language, kindergarten screening may lead to the over- or under-identification of students at risk of reading difficulties (Compton et al., 2006; O'Connor & Jenkins, 1999; Speece et al., 2011). Providing intervention to students identified as at risk in kindergarten may unnecessarily waste resources that would be better used when it is more clear whether children are at risk of experiencing difficulty based on their response to the general education kindergarten curriculum. In addition, whereas Catts et al. (2015) defined reading comprehension difficulties at the end of third grade only, we defined these challenges at the end of first and third grades to be able to determine whether the predictors that mattered most in identifying students would be similar or different at the two grade levels.

The present study also departs from the Catts et al. (2015) study in the measures used. Although Catts et al. (2015) used measures of letter naming fluency, phonological awareness, rapid automatized naming, and nonword repetition as the decoding-related measures, we used measures of sight word reading efficiency and pseudoword decoding efficiency. These measures are more proximal to reading comprehension outcomes and have been found to be important predictors of reading difficulties in previous screening research in first grade (Compton et al., 2010; Speece et al., 2011). In addition, although Catts et al. (2015) used measures of receptive and expressive vocabulary, syntax, and receptive and expressive narrative language as the language-related measures, we used a measure of awareness of semantic relationships, a measure of syntax, measures of receptive and expressive vocabulary, and two measures of listening comprehension, one using a cloze sentence format and one using a paragraph-level multiple-choice format. The decision to use two measures of general listening comprehension instead of one measure of receptive and expressive paragraph-level narrative language ability was based on research identifying the importance of listening comprehension and the objective of comparing two different measures of listening comprehension as predictors of reading comprehension difficulties. Finally, although Catts et al. (2015) identified students as having a reading difficulty based on two, separate, paragraph-level multiple-choice measures of reading comprehension, we identified students as having a reading difficulty based on a composite that included data from three norm-referenced measures: a silent reading efficiency and comprehension measure, a cloze sentence passage comprehension measure, and a paragraph-level open-ended reading response measure of reading comprehension. Representing reading comprehension using multiple measures tapping different aspects of the construct is important given that different variables predict reading comprehension depending on how it is measured (Cutting & Scarborough, 2006).

## The Present Study

Students were assessed on decoding- and language-related skills at the beginning of first grade. Students were also assessed at the end of first and third grades on three different measures of reading comprehension. Analyses were conducted to identify the best predictors of reading comprehension difficulty at the end of first and third grades. The research question guiding this study was:

**Research Question 1**: Which measures of beginning of first grade decoding- and language-related skills are important to identifying reading comprehension difficulties at the end of first and third grades?

Note that, to be used in real-world settings, screening batteries ultimately need to be efficient. It is unreasonable to expect teachers and students to spend an inordinate amount of time assessing children at the beginning of every year. However, there are few efficient measures of language-related skills to be used in screening. Thus, we determined that although many of the measures used in the present study would be inappropriate for a standard school-based screening battery, measures deemed important to include in a screening battery could either be used as models for developing more efficient measures of those constructs or be used in a gated screening process.

## Method

### Participants

*First grade.* The participants were 125 students from three public schools in a Mid-Atlantic suburban area. Consent forms were sent to the parents or guardians of all first grade students ($n = 268$) in those schools with a 68.3% return rate. We obtained parental consent for 146 students. Of those, 16 students declined assent prior to testing, two students were withdrawn from participation due to difficulty understanding assessment tasks, and two students moved out of the school in first grade. The final sample included 65 girls and 61 boys in first grade, and the mean age of students at fall first grade testing was 6.58 years ($SD = .35$). According to parent questionnaires, the sample was 64% White, 17% African American, and 6% Asian. We were unable to obtain race/ethnicity data for 9% of the sample. We used mother's education as a socio-economic descriptor: 9% reported holding graduate degrees, 21% reported holding college degrees, 64% reported holding high school diplomas, and 6% reported not completing high school. We were unable to obtain mother's education data for 1.5% of the sample.

*Third grade.* In third grade, we followed 77 students (62%) of the 125 students we assessed in first grade. We were not able to follow the full sample because many of the students moved out of the school district between first and third grades. In general, the district had high student mobility. Tests for differential attrition indicated no statistically significant differences in age, $F(1, 124) = .32$, $p = .5734$, gender, $\chi^2 = 1.91$, $df = 1$, $p = .1673$, or race, $\chi^2 = 5.42$, $df = 4$, $p = .2587$, or on the *Test of Silent Reading Efficiency and Comprehension* (TOSREC), $F(1, 124) = .01$, $p = .9437$ (Wagner, Torgesen, Rashotte, & Pearson, 2010), at the beginning of first grade.

### Measures

Measures were administered in the fall of first grade and the spring of first and third grades. See descriptive statistics in Table 1. See Supplemental Material for correlations.

*Fall first grade measures.* Nine measures were administered in the fall of first grade.

*The Test of Word Reading Efficiency (TOWRE).* The TOWRE consists of two subtests, Phonemic Decoding Efficiency (PDE) and Sight Word Efficiency (SWE), which measure nonword and real word reading fluency, respectively (Torgesen, Wagner, & Rashotte, 1999). The authors report high alternate-form reliability ($r = .93$) for both subtests. In addition, concurrent validity for the PDE with the Word Attack subtest and SWE with the Word Identification subtest of the *Woodcock Reading Mastery Test–Revised Normative Update* is .85 and .87, respectively.

*Wechsler Individual Achievement Test, Third Edition (WIAT-III).* Five subtests from the WIAT-III were administered in the fall of first grade (Wechsler, 2009). The technical manual provides information about the content and convergent validity of the subtests. Spearman–Brown split-half reliability coefficients for first grade are provided here. WIAT-III Oral Word Fluency measures efficiency of word retrieval by requiring students to say as many words as they can that are related to a verbal prompt (i.e., animals) in 1 min. Split-half reliability is .69. WIAT-III Sentence Repetition measures oral syntactic knowledge and short-term memory. Students are asked to repeat sentences that get progressively more complex. Split-half reliability is .86. WIAT-III Expressive Vocabulary measures speaking vocabulary and word retrieval ability. After viewing a picture prompt (e.g., toothbrush, butterfly, closet), students are expected to name the picture and supply a brief description. Split-half reliability is .71. WIAT-III Receptive Vocabulary measures listening vocabulary. Students are asked to point to a picture (from four choices) that matches an orally presented word. Split-half reliability is .71. WIAT-III Oral Discourse Comprehension measures the ability to make inferences and remember details from oral discourse. Students are asked to answer literal and inferential questions about sentences and passages played on an audio recorder. Split-half reliability is .84.

*The Woodcock-Johnson Tests of Achievement, Third Edition (WJIII).* The Oral Comprehension subtest of the WJIII was administered in the fall of first grade (Woodcock, McGrew, Mather, & Schrank, 2001). The subtest measures students' oral language skill. Students are asked to complete oral cloze sentences after listening to a short audio-recorded passage (i.e., Water looks blue and grass

**Table 1.** Standard Scores on Measures Used in the Study.

| | Grade 1 Fall | | Grade 1 Spring | | Grade 3 Spring | |
|---|---|---|---|---|---|---|
| Measure | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| TOWRE Phonemic Decoding Efficiency | 103.14 | 10.33 | | | | |
| TOWRE Sight Word Efficiency | 100.72 | 13.01 | | | | |
| WIAT-III Oral Discourse Comprehension | 101.82 | 14.64 | | | | |
| WIAT-III Receptive Vocabulary | 99.08 | 12.74 | | | | |
| WIAT-III Expressive Vocabulary | 92.19 | 16.04 | | | | |
| WIAT-III Oral Word Fluency | 102.23 | 15.72 | | | | |
| WIAT-III Sentence Repetition | 97.88 | 12.32 | | | | |
| WJIII Oral Comprehension | 106.82 | 12.33 | | | | |
| TOSREC | | | 93.97 | 16.99 | 98.83 | 13.03 |
| WIAT-III Reading Comprehension | | | 99.74 | 12.21 | 95.81 | 11.10 |
| WJIII Passage Comprehension | | | 104.44 | 12.98 | 94.92 | 9.28 |

*Note*. TOWRE = *Test of Word Reading Efficiency*; WIAT-III = *Wechsler Individual Achievement Test*; WJIII = *Woodcock Johnson Test of Achievement, Third Edition*; TOSREC = *Test of Silent Reading Efficiency and Comprehension*.

looks __.). According to the technical manual, there is strong evidence for validity and the test–retest correlation is .82 and the split-half reliability is .78 for 7-year-olds.

*Spring first and third grade measures.* Three measures of reading comprehension were administered in the spring of first and third grade.

*TOSREC.* On this measure, students are given 3 min to read and respond *true* or *false* to a series of sentences (e.g., *A doughnut is made of very hard steel*) (Wagner et al., 2010). According to the technical manual, alternate-form reliability coefficients exceed .85 across all forms and grade levels, and reliability coefficients with other reading measures such as the WJIII and the Group Reading Assessment and Diagnostic Evaluation (GRADE) exceed .70.

*WIAT-III Reading Comprehension.* This untimed measure of reading comprehension requires students to read various types of text and answer orally presented literal and inferential questions about these texts. According the technical manual, the split-half reliability is .89 for spring of first grade and .82 for spring of third grade. The test–retest reliability is .93 for PK-5. Criterion-related validity is .67 with the Verbal Comprehension Index of the Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV).

*WJIII Passage Comprehension.* This untimed measure of reading comprehension requires students to read sentences and short passages and supply missing words via the cloze format (i.e., There was a rabbit sitting in the _.). According to the technical manual, test–retest correlation is .86 to .89 for ages 4 to 10 and split-half reliability ranges from .91 to .96 for ages 6 to 9. The measure has concurrent validity with the *Kaufman Test of Educational Achievement* Reading Comprehension subtest (.62) and Reading Composite score (.82).

## Analysis

As a first step, we combined the scores on the three measures of reading comprehension into a single composite variable. Rather than take a simple average, we formed a composite score using a principal component analysis (PCA) using Proc Factor in SAS 9.3. The relatively small sample size ($n = 125$) may be a cause for concern when conducting data reduction. However, de Winter, Dodou, and Wieringa (2009) and MacCallum, Widaman, Zhang, and Hong (1999) have noted that the strength of the loadings is far more relevant for determining the suitability of data reduction methods compared with the number of observations. For instance, de Winter et al. (2009) found that sample sizes as small as 20 are sufficient for uncovering the proper component structure and unbiased estimates of component loadings if the loadings in the population are .90. The loadings in our PCA were quite high, so the sample size is not an issue for scoring the composite variable. Results from a Horn's (1965) parallel analysis, in which the eigenvalue of a second extracted component did not exceed threshold, support the notion that the three measures could be reasonably reduced to a single component.

To address how well the predictors are able classify students who have difficulty with reading comprehension, a binary indicator variable was created from the reading comprehension component score. Following a common cut-point in many other studies of learning disabilities (e.g., Cirino, Fuchs, Elias, Powell, & Schumacher, 2015), students at or below the 25th percentile of the reading comprehension composite were considered to have difficulty with reading comprehension. A logistic regression model with

adaptive variable selection was then implemented to determine which of the eight candidate predictors measured in the fall of first grade were best able to predict which students would have reading comprehension difficulties in (a) the spring of first grade and (b) the spring of third grade. Given the widely cited criticism of stepwise and all subsets regression such as incorrect degrees of freedom and deflated standard errors (e.g., Harrell, 2001) and the fact that the ratio of the number of candidate predictor variables ($p = 8$) to the sample size was relatively small (15.8 for first grade spring scores and 9.8 for third grade spring scores; Babyak, 2004), the least absolute selective shrinkage operator (Lasso; Efron, Hastie, Johnstone, & Tibshirani, 2004) was used for predictor selection using the glmnet R package. Once model estimates were obtained, we assessed how successfully the model classified students in the 25th percentile or below using a receiver operator characteristic (ROC) curve and sensitivity and specificity analyses using the pROC R package (Robin et al., 2011). As a comparison, we compared the Lasso-selected models with a baseline model that includes only the two decoding variables (TOWRE PDE and SWE). Confidence intervals (CIs) for area under the ROC curve (AUC), sensitivity, and specificity are reported via the asymptotic DeLong method (DeLong, DeLong, & Clarke-Pearson, 1988) when possible. Otherwise, CIs were obtained via percentile bootstrapping (Carpenter & Bithell, 2000).

Lasso is a method used in machine learning to select meaningful variables in high-dimensional problems (models where there are many predictors relative to the sample size). Although eight predictors is short of the consideration for being "high dimensional" by traditional definitions, the moderate sample size in this study presents a small *n:p* ratio for which Lasso was originally intended (Finch, 2014). Lasso handles overfitting by applying a penalty term to the likelihood function so that regression coefficients are not inflated from overfitting the model. Also, regression coefficients for null predictors are zeroed out and removed from the model as in stepwise or all subsets methods (McNeish, 2015), though the mechanism for doing so in Lasso differs from these traditional methods. Regression coefficients are interpreted identically to a standard regression model. The differences with Lasso lie in the selection of the predictors and the estimation of the effects—the end result, however, is a standard regression model.

The selection of the penalty term can be determined by a few competing methods; we used cross-validation and selected the penalty term within 1 *SE* of the minimum value as recommended in the literature (Friedman, Hastie, & Tibshirani, 2010). Although a method for obtaining *p* values has recently been derived for Lasso with continuous outcomes (Lockhart, Taylor, Tibshirani, & Tibshirani, 2014), statistical theory has not yet advanced a method for computing *p* values with binary outcomes. Therefore, one cannot talk about "significance" of predictors because, unlike stepwise methods, predictors are not selected based on *p* values and *p* values are not computed or capable of being computed. Instead, predictors retained by the Lasso algorithm are considered to be meaningful, regardless of whether the effect may or may not be significantly different from zero (McNeish, 2015). To help contextualize the relative importance of predictors, we report odds ratios and standardized regression coefficients based on *z*-scored predictors. Standardized coefficients are also easier to interpret for our data because a one-unit change on the raw scale of the predictors is rather small and would yield small changes, even if the predictor is important.

Although the data are from a clustered structure, our interest is solely in prediction for lower level units. For this context, random effects models such as hierarchical linear models or cluster corrections do not necessarily affect the results. As noted by Raudenbush and Bryk (2002) regarding the use of single-level methods for multilevel data, "In general, the OLS estimates are unbiased but not as efficient as the hierarchical linear model estimators" (p. 141). Because our prediction models are narrowly focused on the regression coefficients and have no interest in standard errors or inferential tests (in fact, Lasso does not produce either of these), differences in models that do or do not account for clustering would be moot. Bouwmeester et al. (2013) further emphasize this point. Although they note that random effects models can help prediction in some contexts, they state, "Accurate predictions are not necessarily achieved with a random effects model (having different regression parameters compared with a standard model), because the random effects are not readily applicable in new data with new clusters." Later, they note, "The different predictor effects [from a random effects model], however, did not result in clear improvements in model performance (discrimination and calibration) between the marginal risk calculation and the standard model [that does not account for clustering]."

## Results

### PCA

A one-component solution accounted for 87.2% of the variance in the original set of first grade reading comprehension variables and 76.3% of the variance in the original set of third grade reading comprehension measures. Component loadings, communalities, and component reliability as calculated by Coefficient *H* (Hancock & Mueller, 2001) are reported in Table 2; component loadings and communalities were fairly consistent across the three measures for both grades indicating that they contribute relatively equal amounts of information to the overall reading comprehension component. Coefficient *H* is reported instead of the

**Table 2.** Component Loadings, Communalities, and Component Reliabilities for Composite Variable.

| | First Grade ($n$ = 125) | | Third Grade ($n$ = 77) | |
|---|---|---|---|---|
| Measures | Component Loading | Communality | Component Loading | Communality |
| WIAT-III | .92 | .85 | .82 | .68 |
| WJIII | .95 | .90 | .91 | .83 |
| TOSREC | .93 | .87 | .88 | .78 |
| Component reliability | .95 | | .91 | |

*Note.* WIAT-III = *Wechsler Individual Achievement Test*; WJIII = *Woodcock Johnson Test of Achievement*; TOSREC = *Test of Silent Reading Efficiency and Comprehension*.

traditional Cronbach's alpha because it is more appropriate for component-scored variables and it does not require an assumption of *tau* equivalence (McNeish, 2018). The interpretation of Coefficient $H$ is on the same scale as Cronbach's alpha, so it is interpreted similarly. Component reliabilities exceeded .90 for both grades. By default, component scores are reported on a $z$-score scale ($M = 0$ and $SD = 1$). Because the original reading comprehension measures were standardized ($M = 100$ and $SD = 15$), we linearly transformed the principal component scores to match this scale by multiplying component scores by 15 and adding 100 to facilitate interpretation.

$$\text{Coefficient } H = \frac{1}{1+\frac{1}{\frac{l_1^2}{(1-l_1^2)}+\frac{l_2^2}{(1-l_2^2)}+\frac{l_3^2}{(1-l_3^2)}}},$$

where $l$ is the component loading.

## *Classifying At-Risk Students*

*First grade.* Two predictors were selected by Lasso as being non-null (TOWRE SWE and WIAT-III Sentence Repetition) for classifying students who were in the bottom 25% of the reading comprehension component in the spring of first grade. Recall that there are no $p$ values in a Lasso analysis and the selection of the two predictors (from the original eight) implies that they are meaningful to retain in the model together. The standardized coefficient for TOWRE SWE was −.155 meaning that for each 1-*SD* increase, the odds of having difficulty with reading comprehension decrease multiplicatively by .856, holding all other predictors in the model constant. That is, for a 2-*SD* increase, the odds will be .856 × .856 = .733 lower. The standardized coefficient for WIAT-III Sentence Repetition was −.055 meaning that for each 1 *SD* increase, the odds of having difficulty with reading comprehension decrease multiplicatively by .946. A smoothed ROC curve showed that the model performed well in classifying students who do and do not have reading comprehension difficulties in the spring of first grade. The AUC was high at 94.7% (DeLong 95% CI = [90.7, 98.9]). AUC values greater than 70% are considered good, whereas AUC values exceeding 90% are considered excellent (Rice & Harris, 2005). This indicates that TOWRE SWE and WIAT-III Sentence Repetition from the fall of first grade do an excellent job of classifying students with reading comprehension difficulties in the spring of first grade. At the optimal threshold of a .593 probability of having reading comprehension difficulty, 88.9% of students were correctly classified, 90.0% of students without reading comprehension difficulties were correctly classified (specificity; bootstrapped 95% CI = [83.3, 95.6]), and 85.7% of students with reading comprehension difficulties were correctly classified (sensitivity; bootstrapped 95% CI = [74.2, 97.1]).

By comparison, a model with only decoding variables (TOWRE PDE and SWE) yielded an AUC of 91.4% (DeLong 95% CI = [86.6, 96.3]). At the optimal threshold of a .112 probability of having reading comprehension difficulty, 76.9% of students were classified correctly, 68.9% of students without reading comprehension difficulty were classified correctly (specificity; bootstrapped 95% CI = [58.9, 77.8]), and 97.1% of students with reading comprehension difficulty were classified correctly (sensitivity; bootstrapped 95% CI = [91.4, 100]). Figure 1 (Supplemental Material) compares ROC curves for the decoding variables–only model with the Lasso-selected model for first grade. A Venkatraman test for paired ROC curves (Venkatraman, 2000) indicated that the ROC curves were significantly different ($E = 264$, $p = .03$). The most substantively relevant portion of the ROC curves is sensitivity and specificity intervals between .80 and 1.0 because values outside of this range are unacceptably inaccurate (Jiang, Metz, & Nishikawa, 1996). A bootstrapped paired partial area under the ROC curve ($p$AUC; Hanley & McNeil, 1983) test over the [.80, 1.0] specificity interval was statistically significant ($\Delta p\text{AUC} = 3.0$, $D = 2.50$, $p = .013$), indicating that the Lasso-selected model was significantly better at classifying students who had reading comprehension difficulties, conditional on students without reading difficulties being accurately classified at least 80% of the time. A bootstrapped paired $p$AUC test of the [.80, 1.0] sensitivity interval was

not statistically significant ($\Delta p$AUC = 1.0, $D$ = 0.77, $p$ = .44), indicating no difference in the number of misclassified students without reading comprehension difficulties, conditional on students with difficulties being correctly classified at least 80% of the time. In other words, the Lasso model better classifies students with reading comprehension difficulties while minimizing misclassification of students without difficulties. In order for the decoding variable model to classify students with difficulties as well as the Lasso model, the decoding model risks higher misclassification of students without difficulties (as reflected by the much lower optimal threshold and specificity in the decoding variables–only model).

*Third grade.* Three predictors measured in the fall of first grade were selected by Lasso as being non-null (TOWRE SWE, WIAT-III Sentence Repetition, WIAT-III Oral Discourse) for classifying students who were in the bottom 25% of the Reading Comprehension component measured at the spring of third grade. The standardized coefficient for TOWRE SWE was –.079, meaning that for each 1-*SD* increase, the odds of having difficulty with reading comprehension decrease multiplicatively by .924, holding all other predictors in the model constant. The standardized coefficient for WIAT-III Sentence Repetition was −.068 meaning that for each 1-*SD* increase, the odds of having difficulty with reading comprehension decrease multiplicatively by .933. The standardized coefficient for WIAT-III Oral Discourse was −.021 meaning that for each 1-*SD* increase, the odds of having difficulty with reading comprehension decrease multiplicatively by .979. An ROC curve showed that the model performed well in classifying students who do and do not have reading comprehension difficulties in the spring of third grade. The AUC was again quite high at 90.5% (DeLong 95% CI = [83.7, 97.2]), indicating that TOWRE SWE, WIAT-III Sentence Repetition, and WIAT-III Oral Discourse in the fall of first grade excellently classify students with reading comprehension difficulties in the spring of third grade. At the optimal threshold of a .560 probability of having reading comprehension difficulty, 80.8% of students were correctly classified, 77.0% of students of students without reading comprehension difficulties were correctly classified (specificity; bootstrapped 95% CI = [65.6, 86.7]), and 93.8% of students with reading comprehension difficulties were correctly classified (sensitivity; bootstrapped 95% CI = [81.3, 100]).

By comparison, a model featuring only decoding variables (TOWRE PDE and TOWRE SWE) yielded an AUC of 83.6% (DeLong 95% CI = [73.4, 93.7]). At the optimal threshold of a .299 probability of having reading comprehension difficulty, 80.8% of students were classified correctly, 82.0% of students without reading comprehension difficulty were classified correctly (specificity; bootstrapped 95% CI = [72.1, 90.2]), and 75.0% of students with reading comprehension difficulty were classified correctly (sensitivity; bootstrapped 95% CI = [56.3, 93.8]). Figure 2 (Supplemental Material) compares the ROC curves for the decoding variables–only model with the final Lasso-selected model for the third grade sample. Although the magnitude of differences in the AUC values was larger in third grade (6.9%) than in first grade (3.3%), a Venkatraman test for paired ROC curves indicated that the ROC curves were not significantly different ($E$ = 142, $p$ = .31). This pattern is most likely attributable to the smaller sample at third grade compared with first grade. For the most substantively interesting intervals of the curves, a bootstrapped paired $p$AUC test over the [.80, 1.0] specificity interval was not statistically significant ($\Delta p$AUC = 2.8, $D$ = 0.93, $p$ = .36). A bootstrapped paired $p$AUC test of the [.80, 1.0] sensitivity interval was also not statistically significant ($\Delta p$AUC = 4.0, $D$ = 1.59, $p$ = .11), indicating that there was no difference in the number of misclassified students without reading comprehension difficulties, conditional on students with difficulties being correctly classified at least 80% of the time.

## Discussion

The purpose of this study was to investigate whether language-related predictors in addition to decoding-related predictors would add to the identification of reading comprehension difficulties among first and third grade students. In both grades, language-related predictors proved useful in improving the identification of reading comprehension difficulties. Measures of sight word reading efficiency and sentence repetition predicted end of first grade difficulties. Measures of sight word reading efficiency, sentence repetition, and oral discourse predicted end of third grade difficulties. Both screening batteries yielded high classification accuracy (AUC = 94.7% in Grade 1 and AUC = 90.5% in Grade 3). The finding that a measure of sight word reading efficiency was important to predicting reading comprehension difficulties at both grade levels is not surprising given the extensive research indicating that decoding-related skills are important predictors of reading (e.g., Compton et al., 2006; O'Connor & Jenkins, 1999). This measure captures the ability to read real words with fluency, which is an essential component of reading comprehension (e.g., Silverman, Speece, Harring, & Ritchey, 2013). Students who have difficulty with this decoding-related skill will not be able to access the words on the page regardless of whether they can infer their meaning. Findings from this study confirm that sight word reading efficiency is an integral part of an early first grade screening battery. In fact, in the present study, sight word reading efficiency was more important for predicting reading difficulty than PDE, which may have implications for which measures are ultimately

included if screening batteries need to be winnowed down to be as brief as possible.

The importance of sentence repetition to the prediction of reading difficulty at the end of first grade and third grade is intriguing. Measures of sentence repetition have been widely used in identifying students with early speech and language difficulties. Sentence repetition is a task that purportedly measures syntactical knowledge and short-term memory, but there is some indication that the measure may tap students' general language skills (Klem et al., 2015). In the Catts, Nielsen, Bridges, and Liu (2016) study, kindergarten sentence imitation was an important predictor of third grade reading comprehension, controlling for word reading measures, vocabulary, and narrative language on one measure, the Measures of Academic Progress: Reading (MAP; Northwest Evaluation Association, 2009), but not the other, an experimental measure similar in format to that found in informal reading inventories such as the *Qualitative Reading Inventory-5* (QRI-5; Leslie & Caldwell, 2011). In the present study, reading comprehension was measured by a composite of three measures representing three different formats of reading comprehension assessment: an open-ended reading response task, a cloze sentence task, and a silent reading efficiency and comprehension task. Combining these measures into one composite makes capturing the underlying construct of reading comprehension more likely. The fact that sentence repetition was an important predictor of first and third grade reading comprehension difficulties using this measurement model suggests it may be an important predictor of the underlying construct of reading comprehension that may or may not be evaluated by individual assessments of reading comprehension. Of note, the sentence repetition task is a relatively efficient task to administer. Although administering an extensive battery of language-related assessments to first graders would be impractical given the amount of time that these assessments typically take to administer, it may be feasible to administer a fairly quick measure of sentence repetition, which could be derived from the WIAT-III task, to children in first grade as a screener.

One other measure of oral language was found to be important in the prediction of reading comprehension difficulty in third grade. However, unlike in previous studies, this measure was not a vocabulary measure (e.g., Catts et al., 2016; Wise et al., 2007). Instead, oral discourse contributed to the prediction of reading difficulties in third grade. The oral discourse measure used in this study asks students to listen to sentences and passages and then orally respond to literal and inferential comprehension questions. This measure may be an indicator of global inference-making skills (Oakhill, Cain, & Bryant, 2003). Global inference making involves integrating information across a text and, often, with background knowledge as well. To adequately identify students who have difficulty at the end of third grade, it may be important to consider early screening and intervention for early global inference-making skills. It is impractical to administer the specific oral discourse measure used in the present study to all first grade children. However, new, more efficient measures that tap oral discourse skills could be developed for screening purposes or oral discourse could be used in a second stage in a two-stage gated screening process (e.g., Compton et al., 2010). To adequately identify and prevent later reading difficulties, further research on assessing early oral discourse skills is needed.

There are several limitations to the study that should be noted. First, the sample size was small; it was likely unrepresentative of the populations in many school districts around the country, and we had high attrition from first to third grade. These limitations undermine the generalizability of the study. Second, some of the measures we used had relatively low reported reliability (.69–.71) which could bias coefficients and affect power. Therefore, further research with more reliable measures is needed. Third, this study focused narrowly on classification accuracy. Additional research is needed to fully explore the validity of screening batteries including sentence repetition and oral discourse tasks. Fourth, as we were focused on classification, we used a binary outcome (i.e., reading difficulty or not) and a cut-point that, while used extensively in previous research, is somewhat arbitrary. Future research on early screening using decoding- and language-related measures should investigate other ways of defining reading difficulty. Furthermore, future research should include measures of growth as well as cognitive skills, which have been shown to improve identification of students with reading difficulties (e.g., Compton et al., 2006). Finally, this study was not implemented in the full context of RTI implementation, which would likely require sentence repetition and oral discourse to be assessed via more efficient measures or within a gated screening process. Also, if these constructs are included in screening within an RTI framework, future research should investigate implications for intervention and progress monitoring.

Despite these limitations, the present study adds to the research base on early screening by investigating a broad range of language-related measures in addition to decoding-related measures, by using a composite of reading comprehension measures to identify reading difficulties, and by comparing results across first and third grades. This research also builds upon findings that early screening is a powerful tool to identify later reading difficulties (e.g., Compton et al., 2006; O'Connor & Jenkins, 1999; Speece et al., 2011). Findings of this study move the field forward by suggesting the possibility of developing screening batteries composed of decoding-based, language-based, and comprehension measures that are sensitive enough to detect potential reading failure. Finally, this study suggests it is important to consider decoding-related *and* language-related measures,

including sentence repetition and oral discourse measures, in early screening of reading difficulties. Research along these lines is needed to improve identification and intervention for students at risk of reading difficulties.

## Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by grants from the National Institute of Child Health and Human Development (Grant No. R01 HD 046758) and the U.S. Department of Education (Grant No. H325D070082).

## ORCID iD

Rebecca D. Silverman https://orcid.org/0000-0002-9785-0313

## Supplemental Material

Supplemental material for this article is available online at https://journals.sagepub.com/doi/suppl/10.1177/1534508419857234.

## References

Adlof, S. M., Catts, H. W., & Little, T. D. (2006). Should the simple view of reading include a fluency component? *Reading and Writing: An Interdisciplinary Journal*, *19*, 933–958.

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*, 411–421.

Bouwmeester, W., Twisk, J. W., Kappen, T. H., van Klei, W. A., Moons, K. G., & Vergouwe, Y. (2013). Prediction models for clustered data: Comparison of a random intercept and standard regression model. *BMC Medical Research Methodology*, *13*, 19.

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, *19*, 1141–1164.

Catts, H. W., Compton, D., Tomblin, J. B., & Bridges, M. S. (2012). Prevalence and nature of late-emerging poor readers. *Journal of Educational Psychology*, *104*, 166–181.

Catts, H. W., Nielsen, D. C., Bridges, M. S., & Liu, Y. (2016). Early identification of reading comprehension difficulties. *Journal of Learning Disabilities*, *49*, 451–465.

Catts, H. W., Nielsen, D. C., Bridges, M. S., Liu, Y. S., & Bontempo, D. E. (2015). Early identification of reading disabilities within an RTI framework. *Journal of Learning Disabilities*, *48*, 281–297.

Cirino, P. T., Fuchs, L. S., Elias, J. T., Powell, S. R., & Schumacher, R. F. (2015). Cognitive and mathematical profiles for different forms of learning difficulties. *Journal of Learning Disabilities*, *48*, 156–175.

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., . . . Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology*, *102*, 327–340.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, *98*, 394–409.

Cooper, D., Roth, F., Speece, D., & Schatschneider, C. (2002). The contribution of oral language to the development of phonological awareness. *Applied Psycholinguistics*, *23*, 399–416.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, *10*, 277–299.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845.

de Winter, J. D., Dodou, D. I. M. I. T. R. A., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, *44*, 147–181.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*, 407–499.

Finch, H. (2014). A comparison of methods for group prediction with high dimensional data. *Journal of Modern Applied Statistical Methods*, *13*, 5.

Florit, E., & Cain, K. (2011). The simple view of reading: Is it valid for different types of alphabetic orthographies? *Educational Psychology Review*, *23*, 553–576.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 195–216). Lincolnwood, IL: Scientific Software International.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, *148*, 839–843.

Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York, NY: Springer.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal*, *2*, 127–160.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.

Individuals with Disabilities Education Improvement Act, Pub. L. No. 108-446. (2004).

Jiang, Y., Metz, C. E., & Nishikawa, R. M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*, *201*, 745–750.

Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S. H., Gustafsson, J., & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science*, *18*, 146–154.

Leslie, L., & Caldwell, J. A. (2011). *Qualitative Reading Inventory: 5*. Boston, MA: Pearson/Allyn & Bacon.

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics*, *42*, 413–468.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84–99.

McNamara, J. K., Scissons, M., & Gutknecth, N. (2011). A longitudinal study of kindergarten children at risk for reading disabilities: The poor really are getting poorer. *Journal of Learning Disabilities*, *44*, 421–430.

McNeish, D. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*, 471–484.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*, 412–433.

Northwest Evaluation Association. (2009). *Measures of Academic Progress*. Lake Oswego, OR: Author.

Oakhill, J. V., Cain, K., & Bryant, P. E. (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes*, *18*, 443–468.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and Data analysis methods*. Thousand Oaks, CA: SAGE.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's ***d***, and ***r***. *Law and Human Behavior*, *29*, 615–620.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Speece, D. L., Schatschneider, C., Silverman, R. D., Case, L. P., Jacobs, D., & Cooper, D. (2011). Early identification of reading problems in a response to intervention framework. *Elementary School Journal*, *111*, 585–607.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A. (1999). *Test of Word Reading Efficiency: Examiner's manual*. Austin, TX: PRO-ED.

Venkatraman, E. S. (2000). A permutation test to compare receiver operating characteristic curves. *Biometrics*, *56*, 1134–1138.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2010). *Test of Silent Reading Efficiency and Comprehension (TOSREC)*. Austin, TX: Pro-Ed.

Wechsler, D. (2009). *Wechsler Intelligence Scale for Children* (3rd ed.). San Antonio, TX: Pearson.

Wise, J., Sevcik, R., Morris, R., Lovett, M., & Wolf, M. (2007). The relationship among receptive and expressive vocabulary, listening comprehension, pre-reading skills, word identification skills, and reading comprehension by children with reading disabilities. *Journal of Speech, Language, and Hearing Research*, *50*, 1093–1109.

Woodcock, R. W., McGrew, K., Mather, N., & Schrank, F. (2001). *Woodcock-Johnson Tests of Achievement* (3rd ed.). Itasca, IL: Riverside.