# HHS Public Access

# Early Identification of Reading Disabilities within a RTI Framework

**Hugh W. Catts, Ph.D.**, **Diane Corcoran Nielsen, Ph.D.**, **Mindy Sittner Bridges, Ph.D.**, **Yi Syuan Liu, BA**, and **Daniel E. Bontempo, Ph.D.**
University of Kansas, 1000 Sunnyside Ave, Lawrence, KS 66045

Hugh W. Catts: catts@ku.edu; Diane Corcoran Nielsen: dnielsen@ku.edu; Mindy Sittner Bridges: msittner@ku.edu; Yi Syuan Liu: michelle.liu@ku.edu; Daniel E. Bontempo: deb193@ku.edu

## Abstract

Early and accurate identification of children at risk for reading disabilities (RD) is critical for the prevention of RD within a RTI framework. In this study, we investigated the use of universal screening and progress monitoring for the early identification of RD in kindergarten children. Three-hundred sixty-six children were administered a battery of screening measures at the beginning of kindergarten and progress monitoring probes across the school year. A subset of children who showed initial risk for RD also received a 26-week Tier 2 intervention. Participants' achievement in word reading accuracy and/or fluency was assessed at the end of first grade. Results indicated that a screening battery containing measures of letter naming fluency, phonological awareness, rapid naming or nonword repetition accurately identified good and poor readers at the end of first grade. Findings also showed that children's response to supplemental and/or classroom instruction measured in terms of growth in letter naming fluency added significantly to the prediction of reading outcomes.

Response to intervention (RTI) is a model for the early identification and prevention of reading disabilities (D. Fuchs & Fuchs, 2005; Fuchs & Vaughn, 2012; Haager, Klinger, & Vaughn, 2007). According to this model, children can be identified as having a learning or reading disability (RD) if their response to scientifically-based instruction, including targeted intervention, is substantially below that of their peers. Response to instruction/ intervention is assessed by universal screening and/or progress monitoring measures. All children participate in periodic universal screening to identify children who are potentially at-risk for RD. Those who "fail" universal screening receive supplemental instruction (Tier 2), and their response is assessed by progress-monitoring measures to further gauge risk for RD. Children who continue to show poor response may be provided with more intensive intervention (Tier 3), and in some settings, be considered for special education placement.

For RTI to be maximally successful, it is critical that identification procedures (i.e., universal screening and progress monitoring) are carried out in a timely and accurate manner. Preferably, identification would take place in kindergarten or first grade, prior to at-risk children experiencing significant reading problems. This would allow for the opportunity to provide early intervention to prevent RD or significantly reduce its impact.

Early identification procedures should also be accurate. Accuracy is often assessed in terms of sensitivity (i.e., correctly identifying those who will have RD) and specificity (i.e., correctly identifying those who will not have RD). Screening procedures that result in sensitivity levels at or above 90% and specificity levels of at least 80% are generally deemed acceptable (Jenkins, 1993). An alternative index of accuracy is area under the receiver operating characteristic (ROC) curve (Metz, 1978; Swets, 1979). A ROC curve is a plot of the true-positive rate (sensitivity) against the false-positive rate (1-specificity) for each of the cut points of a decision making instrument. As such, the area under the curve (AUC) is an overall estimate of the accuracy an assessment. Values above .80 are considered good and values above .90 are excellent.

Research has begun to examine the use of universal screening and progress monitoring within RTI for the early identification of RD (Al Otaiba et al., 2011; Compton, Fuchs, Fuchs, & Bryant, 2006; O'Connor & Jenkins, 1999). In a particularly noteworthy study, Compton and colleagues (Compton et al., 2006) administered a multivariate screening battery (i.e., a set of measures tapping different pre-literacy skills) and short-term progress monitoring measures to 252 beginning first-grade children with low initial reading abilities. Children's reading outcomes were subsequently measured at the end of second grade. Logistic regression analyses showed that a screening model that included measures of phonological awareness, rapid digit naming, and oral vocabulary predicted reading outcome with a high degree of accuracy. AUC was .84 when reading outcome was based on individual component measures of reading and .86 when reading outcome was based on a composite score for reading. When growth parameters (level and slope) from 5 weeks of progress monitoring of word reading fluency was added to the original model prediction was significantly improved (AUC = .89 or .91). Finally, when classification tree methodology was used rather than logistic regression analyses, the prediction was improved further (AUC = .94 or .98). However, some caution is warranted in drawing conclusions from the classification tree analyses since this analysis can over fit the data when the number of decision nodes is as high as it was in this study (Breiman, Friedman, Olshen, & Stone, 1984).

Evidence of the ability to accurately identify children at risk for RD from the beginning of first grade is a positive sign for the RTI approach. Then again, it raises the question of whether or not accurate identification could take place earlier. In the United States, most children are enrolled in kindergarten, often in a full day program. Furthermore, kindergarten curricula have changed dramatically in recent years, and in the majority of settings, include formal reading instruction (Al Otaiba et al., 2008; Al Otaiba et al., 2011). Most kindergarten children come to school with some literacy knowledge (e.g., letter knowledge), and by the end of the school year, can read and spell some words. In response to growing expectations in literacy, many schools have implemented universal screening in kindergarten to identify at-risk children.

Research suggests that screening for risk for RD in kindergarten can have acceptable levels of accuracy. Studies have generally found moderate correlations between literacy (letter knowledge) or language abilities (phonological awareness, vocabulary) in kindergarten and reading in the early school grades (National Institute of Child Health and Human

Development, 2000; Scarborough, 1998; Schatschneider, Fletcher, Francis, Carlson, & Foorman, 2004). A small number of studies have also used multivariate-screening tests in kindergarten to predict reading achievement in the primary grades (Catts, Fey, Zhang, & Tomblin, 2001; Felton, 1992; Ritchey, 2004; O'Connor & Jenkins, 1999; Vellutino, Scanlon, Zhang, & Schatschneider, 2008). These studies have shown that kindergarten screening can lead to accurate identification. For example, both O'Conner and Jenkins (1999) and Catts et al. (2001) reported sensitivity levels above 90% and specificity levels between 80–90%.

An important component of early identification in a RTI framework is the use of response to Tier 2 instruction as a further indicator of risk for RD. As noted above, in the typical RTI model, children who are deemed to be at risk based on universal screening are provided with supplemental Tier 2 instruction. In many situations, some of these children will be falsely identified (e.g., false positives), perhaps due to their lack of experience or inadequate classroom instruction (Vellutino et al., 2008). However, the latter children would be predicted to respond positively to Tier 2 instruction and demonstrate satisfactory performance on progress-monitoring measures. Those children who are truly at risk would be expected to respond less well to supplemental instruction. Thus, in this framework, response to Tier 2 intervention theoretically becomes another indicator of risk for RD.

Findings from Compton et al. (2006) offer some indirect support for the usefulness of response to instruction/intervention in the prediction of RD. As reported above, they found that a first-grade prediction model that included multivariate-screening measures and growth parameters from a 5-week progress monitoring of word reading fluency was a significantly better predictor of reading outcome than was a model including only multivariate-screening measures. It is important to note that because children had not yet been placed in Tier 2 intervention, growth in progress monitoring in this study is better characterized as response to classroom instruction (Tier 1) and not response to Tier 2 intervention. Al Otaiba et al. (2011) also examined response to classroom instruction as a predictor of reading achievement. They administered measures of reading and reading-related skills (e.g., letter naming, vocabulary) periodically during kindergarten and assessed reading achievement at the end of first grade. Their results showed that end-of-kindergarten year scores in reading and reading-related skills were good predictors of reading achievement. Once end-of year scores were controlled, growth in these skills across kindergarten added to the prediction. However, rapid growth was associated with a higher likelihood of reading problems. In other words, students who grew more rapidly to achieve the same point at the end of kindergarten had poorer reading outcomes in first grade than those who grew less. Al Otaiba et al. argued that the former children likely came to school less prepared and had more room to grow. Data were not reported concerning how growth interacted with beginning-kindergarten scores in predicting reading outcomes.

The most direct evidence of the added predictive value of response to Tier 2 intervention is provided by Vellutino et al. (2008). They administered a battery of screening measures to a large sample of kindergarten children at the beginning of the school year. Children who scored below the 30th percentile on a letter-naming task were designated as at risk. Half of these children were randomly assigned to receive Tier 2 intervention by project personnel.

The remainder received whatever remedial services were routinely provided by their home school (i.e. business as usual). At the beginning of first grade, children who had received Tier 2 intervention were divided into those who continued to be at risk and those who no longer were at risk (based on letter identification and word reading abilities). Analyses were carried out to predict the latter group membership. Results showed that the initial kindergarten screening battery resulted in a less than optimal prediction of whether children continued to be at risk or not (AUC=.79). Vellutino et al. suggested that many of the children with initial risk may not have had enough literacy experience by the beginning of kindergarten for a literacy-based screening battery to be predictive of reading outcomes. Further analyses, however, demonstrated that when measures of response to Tier 2 intervention (growth in letter knowledge and word reading abilities across the kindergarten year) were added to the initial prediction model, the accuracy of the model was quite high (AUC=.96

The present study was carried out to further investigate the usefulness of an RTI approach for the early identification of RD in kindergarten children. We administered a multivariate-screening battery to a group of children at the beginning of kindergarten. This battery included measures that are commonly used in kindergarten screening (e.g., Letter Naming Fluency) as well as other less frequently used measures (e.g., nonword repetition). We also included both short-term progress monitoring over the first six weeks of school and longer-term progress monitoring over the entire school year. In addition, children deemed to be at initial risk for RD based on beginning-of-year progress monitoring probes were randomly assigned to a Tier 2 intervention or a business as usual control condition. At the end of first grade, we assessed all children's reading achievement. Because reading achievement at this grade is primarily influenced by word reading abilities, our outcome assessments focused on word reading accuracy and fluency. Analyses were undertaken to determine what combination of screening measures and/or progress monitoring probes best predicted reading achievement. We also investigated whether or not response to Tier 2 intervention and/or classroom instruction added to this prediction.

## Method

### Participants

The participants in this study were 366 kindergarten children from a medium-sized school district. This district is diverse in terms of ethnicity (approximately 63% Caucasian, 11% African-American, 6% Hispanic, 7% American Indian/Alaskan native, 6% Asian/Pacific Islander, and 7% multi-racial) and family SES (24% free & 11% reduced lunch). The district contained 15 elementary schools; 8 with full day kindergarten classes and 7 with half day kindergarten classes. Participants entered the study in two cohorts, one year apart. In selecting our participants, we oversampled children with increased risk for reading disabilities. This oversampling was necessary to provide the opportunity to examine the added predictability of response to Tier 2 intervention in at-risk children. Although these children are referred to as "at risk," this designation is based on beginning-of-year progress monitoring probes and not on our screening battery. Specifically, at-risk status was determined by performance on two subtests from the Dynamic Indicators of Basic Early

Literacy Skills (DIBELS; Good & Kaminski, 2002) that were administered to all district kindergarteners by school personnel in the first week of school. To be considered at risk, children had to perform in the "Some risk" or "At-risk" categories on both the Letter Name Fluency and Initial Sound Fluency subtests. Approximately 20% of kindergarteners in the district (across the two years) met this criterion. The majority of these children (N=263; 150 boys, 113 girls) served as participants in the study. The remaining were excluded because they had severe disabilities such as autism or behavior disorders (13), limited English proficiency (23) or were unavailable for testing because they moved before testing had begun or parent/teacher requested nonparticipation (5). In addition to the at-risk participants, we randomly selected 103 children (53 boys, 50 girls) who did not meet the risk criteria on the DIBELS subtests.

Between the time the screening battery was administered at the beginning of kindergarten and the end of first grade, 49 children (13.4%) were dropped from the study. Most of the children moved out of the district (42) and were unavailable for testing. Other children were dropped because of parental request (3), later diagnosis of autism or other special needs (3), or excessive absences (1). In addition to the above attrition, 4 children were missing one or more of the screening, progress-monitoring, or reading-outcome measures. Thus, a complete data set was available for 313 children through the end of first grade.

### Measures and Procedures

All participants were administered a battery of screening measures at the beginning of kindergarten and progress-monitoring measures across the year. A portion of the at-risk participants were provided with Tier 2 intervention. Finally, measures of reading achievement were administered at the end of first grade. The specific measures and procedures used in each component of the study are listed below.

**Screening and progress monitoring—**In mid to late September of kindergarten, all participants were administered a battery of screening assessments. Some of these assessments also served as progress monitoring measures and were given periodically across the kindergarten year or at the end of the school year. All assessments were administered by trained examiners from our research team with one exception. As described below, the participating school district provided data on several measures collected by trained school personnel as part of district-wide progress monitoring.

The selection of assessments was based on practical and theoretical bases. Two subtests of DIBELS (Good & Kaminski, 2002) were administered for screening and progress monitoring. These measures, Letter Naming Fluency and Initial Sound Fluency, have been widely used in schools and measure abilities (i.e. letter knowledge and phonological awareness) shown to be related to early reading achievement (Catts et al., 2002; O'Conner & Jenkins, 1999; Schatschneider et al., 2004). Because of the low reliability of the Initial Sound Fluency subtest, two other measures of phonological awareness were also administered. One measure was comparable to Initial Sound Fluency in the aspect of phonological awareness that was measured (i.e., sound identity) but was untimed and had higher reliability. The other measure was a dynamic assessment of sound elision that

provided children with feedback and instruction during administration and also had acceptable reliability. In addition, measures of sentence imitation, rapid naming, and nonword repetition abilities were administered as part of the screening battery. Previous research has documented that measures of these abilities are predictive of early reading achievement (Catts et al., 2001; Schatschneider et al., 2004). Each of the screening and progress monitoring measures are listed in Table 1 and are described below.

**Letter Naming Fluency (LNF):** In this subtest of DIBELS, the participant is shown a stimulus card containing 11 rows of randomly presented upper- and lower-case letters. The child names as many letters as he/she can in 1 minute. A different form was available for each administration, and the published alternate form reliability was .88. LNF was administered on 9 occasions across the kindergarten year. These occurred approximately during school weeks 1, 3, 5, 7, 15, 23, 29, 35, and 38. For ease of presentation, we refer to the administration by number rather than week (i.e., LNF1–9). Administration of LNF1, LNF6, and LNF9 were conducted by school personnel as part of district-wide assessment, and all other LNF assessments were administered by study personnel. LNF1 was used in part to identify participants at initial risk (as described above). LNF3 was given concurrently with other screening measures and was the primary measure of letter knowledge used in the screening models. We also used growth from initial biweekly progress monitoring (LNF1–4) in screening models. Additional progress monitoring (LNF6, LNF9) was used to evaluate response to instruction for all participants. Further assessments of letter knowledge (LNF5, 7 & 8) were given only to at-risk children participating in Tier 2 intervention (see below).

**Initial Sound Fluency (ISF):** In the ISF task, the participant is shown a series of stimulus cards containing four pictures. The examiner provides the names of the four pictures and asks the participant to identify the picture that begins with a particular sound. The child is also asked to produce the beginning sounds of words presented orally by the examiner. The amount of time taken to identify/produce the correct sounds is converted into the number of initial sounds correct in a minute. A different form was available for each administration, and the published alternate form reliability was .72. ISF was administered according to the same schedule as LNF and was used for screening/progress monitoring in the same manner as LNF with one exception. ISF9 was unavailable because it was not part of district-wide assessment in week 38.

**Sound Matching:** The Sound Matching subtest from the Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgesen, & Rashotte, 1999) is an untimed test of the ability to identify the sounds in words (i.e., phonological awareness). The participant is shown a series of stimulus cards, each with a target picture and three test pictures. The examiner provides the name of each picture and the participant is asked to identify which of the three test pictures starts or ends with the same sound as the target picture. Test-retest reliability is .83 and internal consistency is .93. Sound Matching was administered at the beginning of the year as part of the screening battery and at the end of the year for progress-monitoring purposes. These assessments are designated as Sound Matching and Sound Matching 2, respectively.

**Dynamic Screening of Phonological Awareness:** In this task (Bridges & Catts, 2010), the participant is required to delete a portion of a word and say the remaining word. Unlike static phonological awareness measures, in this dynamic task, the child is provided with feedback and instruction throughout the task. This feedback/instruction consists of standardized prompts. According to the test procedures, when a child gives a correct response, the response is acknowledged as so. Alternatively, when a child gives an incorrect response to an item, the examiner provides a series of prompts until the item is answered correctly or the answer is given. The score for each item decreases by one point for each successive prompt that is needed. Test-retest reliability is .89 and internal consistency is .86.

**Rapid Automatized Naming (RAN):** On this subtest of the CTOPP, the participant is presented with two forms displaying pictured arrays of 6 common objects repeated 6 times in a random order. The child named all objects from each form as quickly as possible. The number of seconds required to name the objects from each form was combined to derive the score for this measure. The alternate-form reliability is .82.

**Nonword Repetition:** In this his task (Dollaghan & Campbell, 1998), the participant is required to repeat 16 nonwords ranging from one to four syllables in length (four words at each length). Each of the nonwords was composed of early-developing phonemes and contained syllables that did not correspond to English lexical items. Nonwords were presented to children via headphones and a high-quality audio-recorder and participants' responses were recorded. An examiner scored the audio-recorded responses in terms of the number of consonants in error across the 16 words. A second examiner re-scored approximately 13% of the data and interjudge reliability was 93%.

**Sentence Imitation:** In this subtest from Test of Language Development-P:3 (TOLD-P:3; Hammill & Newcomer, 1997), the child is presented with a series of spoken sentences that increase in length and grammatical complexity. The participant is required to repeat each sentence as accurately as possible. Test-retest reliability is .90 and internal consistency is .92.

**Tier 2 intervention**—Participants selected into the study based on initial risk on DIBELS (at-risk children, N=263) were randomly assigned to one of two conditions: (1) Tier 2 Intervention condition ($n = 156$) or (2) at-risk control condition ($n = 107$). Proportionally more children were assigned to the intervention condition than the at-risk control condition in order to ensure sufficient sample size for potential analyses. Because children received intervention in small groups at the school they attended, it was not possible to use a completely random approach in group selection. Rather, for each of the two cohorts, all at-risk children at a given school were randomly assigned into groups of three (or two/four to ensure that all at-risk children were grouped). Then, for each cohort, we randomly assigned these small groups across the district to the intervention or no-intervention condition at a 3 to 2 rate. This resulted in a total of 47 small groups (59.5%) in the Tier 2 intervention condition and 32 (40.5%) in the no intervention control condition. Children in the intervention condition received the Tier 2 intervention described below. At- risk control children participated in business as usual practice within the district. In many cases, these

children received some supplemental small-group intervention from school-based reading specialists or paraprofessionals. We will return to this issue in a later section.

**Intervention:** For the Tier 2 intervention group, intervention began in the third week of October and continued for 26 instructional weeks. The intervention consisted of three 30-minute sessions a week. Approximately one-half of each session was devoted to training in phonological awareness, letter-name/sound knowledge, and the alphabetic principle. The phonological activities used in this instruction were drawn from Schuele and Dayton (2000) as well as Blachman, Ball, Black, and Tangel (2000). Activities followed a scope and sequence that moved quickly from working at a syllable level to primarily working at the single phoneme level. The schedule for instruction was predetermined but allowed for some flexibility based on group progress. Activities included sound sorting and segmenting and blending of speech sound units. Manipulatives such as letter tiles were used to isolate sound units and provide visual support in guided practice for segmenting and blending. Letter names and their corresponding sounds were systematically introduced and explicitly taught following the letter sequence used in Animated Literacy (Stone, 2006), which was the phonics program used in the district. Instruction of each letter began with a clear connection between the letter's sound (/p/) and its name and was linked to a character (e.g., Polly Panda) and key words related to the character's action (e.g., painting purple Ps), which were the same as those used in Animated Literacy. Students were actively involved during guided practice and review activities. The sound-letter connection was constantly reinforced ("What *sound* do you hear, what *letter* makes that sound?*")* in activities included in other parts of the lesson. During the last four weeks of intervention, students were taught to read and spell one-syllable CV, VC, or CVC words.

In addition to the above instruction, the other half of each 30-minute session included activities directed at improving vocabulary and language comprehension/production, factors particularly related to reading comprehension. In the current paper, we are primarily concerned with instruction related to word reading outcomes rather than comprehension. Therefore, the details of our vocabulary and language instruction will be included in a future paper that will examine the relationship between response to this instruction and outcomes in reading comprehension.

The intervention was carried out by educators (e.g., substitute teachers) and paraeducators on our research team. The interventionists attended a two-day workshop that provided theoretical background and training on lesson specific strategies. They also met with trainers (second and third authors who wrote the lessons) biweekly for lesson-related training. A procedural fidelity checklist was developed for each lesson plan that was used to document the instructor's use of directions, pacing and sequencing of activities, monitoring of student engagement in and completion of activities, and use of any necessary materials. Following workshop training, interventionists were observed by professional research staff to ensure the integrity of implementation of lesson plans using the fidelity checklists until they had achieved 95% or higher fidelity on three consecutive sessions. Subsequently, approximately 20% of the lessons randomly selected were observed by research staff to monitor fidelity for drift.

**Assessment of response to Tier 2 intervention:** Participants' response to Tier 2 intervention was assessed in several ways. As noted above, progress-monitoring probes involving the LNF and ISF were administered across the school year. Because ISF was not available at the end of the year, Sound Matching was re-administered, and pretest-posttest performance on this measure served as an index of response to phonological awareness intervention. Response to instruction was also assessed for children in the at-risk and typical control groups. This involved LNF (1–4,1–6,1–9) and ISF (1–4,1–6) probes and pretest-posttest performance on Sound Matching. Raw scores and growth curve model-derived scores were employed in data analyses. Stata's *xtmixed* procedure was used to fit random coefficient linear growth models (Rabe-Hesketh & Skrondal, 2008) and subsequently obtain (i.e., predict) individual level and slope scores capturing each individual's model-derived LNF and ISF growth trajectories[1]. In these models, random slope variation was separately estimated for at-risk and non at-risk children. We centered growth estimates at the third administration because this administration was used in the primary screening model. The use of this administration also had distributional advantages, as greater floor effects were observed in the first two administrations. Reliability estimates for rates of change (random slope scores) were generally good [2], varying around alpha =.7. Across our 4-, 6-, and 9-adminstration growth models, reliability estimates for slope scores were lower for ISF than LNF, which is consistent with the poorer test-retest reliability of ISF.

**Reading outcome in first grade—**At the end of first grade, all participants were administered measures of reading achievement by trained examiners on our research team. Because reading achievement in first grade is primarily based on word reading, our assessments included measures of word reading accuracy and/or fluency.

**Woodcock Reading Mastery Tests-Revised: Normative Update (WRMT-R:NU; Woodcock, 1998):** Two subtests from this measure were administered: Word Identification and Word Attack (Woodcock, 1998). The Word Identification subtest measures a participant's ability to accurately pronounce printed English words ranging from high to low frequency of occurrence. The Word Attack subtest assesses participant's ability to read pronounceable nonwords varying in complexity. These subtests were combined to form the Basic Skills Cluster score that served as the index of performance for the WRMT-R. The split-half reliability of the Basic Skills Cluster for first grade is .98.

**Test of Word Reading Efficiency-Second Edition (TOWRE-2; Torgesen, Wagner & Rashotte, 2011):** This measure is composed of two subtests: Sight Word Efficiency and Phonemic Decoding Efficiency. The Sight Word subtest measures how many printed

---

[1]Linear models were used rather than non-linear (e.g., quadratic) models because growth curvature was slight when plotted and unlikely to have impacted results. Linear models also had the advantage of a single, easily interpreted slope coefficient that could be used in predictive models.

[2]LNF growth score reliability coefficients for at-risk versus non at-risk children were.61 vs .78, .80 vs .72, and .82 vs .71 respectively. These across-status differences are the result of modeling independent random slopes for each risk-status, as well as fewer assessments of the nonintervention children (missing assessment 5, 7, and 8). For the ISF 1–4 scores, reliability estimates were low (. 47 at-risk vs. .64 non at-risk), but improved in the ISF1–6 model (.74 at-risk vs .67 non at-risk). The reliability of slopes is more complex than the reliability of assessments themselves. Slope reliability increases with greater precision (smaller within-person residual variance), decreases with less between-person variation in random slopes, and increases with more assessments or more widely-spaced assessments. Therefore, variation in slope reliabilities can be difficult to interpret (see Singer & Willett, 2003).

English words, which range from high to low frequency of occurrence, a participant can accurately pronounce in 45 seconds. The Phonemic Decoding subtest assesses how many pronounceable nonwords, which vary in complexity, a participant can accurately pronounce in 45 seconds. Scores from each subtest were combined to form a standard score for overall performance. The test-retest reliability of this measure in first grade is .92.

**Florida Assessment of Instruction for Reading: Oral Reading Fluency (FAIR: ORF; Florida Department of Education, 2009):** The participant read aloud two grade-appropriate passages (155 and 190 words in length) and the number of words read correctly from each passage in a minute was adjusted for passage dependency based on normative data provided with the measure. Scores were then averaged to form the index of performance. The alternate-form reliability for passages in first grade is .95.

**Classification of reading outcomes**—For each of the above measures, participants were classified as reading disabled (RD) or non-RD. RD was defined as performance equal to or below the 20th percentile, and non-RD as above this cut-score. This cut-score is comparable to that of other researchers who have investigated RD in primary grade children (Lovett, Steinbach, & Frijters, 2000; Speece, Mills, Ritchey, & Hillman, 2003; Torgesen, 2009). Researchers have generally chosen liberal definitions of RD in these grades to assure that children with moderate but potentially significant reading problems are identified. This is especially the case within an RTI framework where early identification can lead to Tier 2 intervention (Vellutino et al., 2008). We used local data to calculate standard scores and percentiles rather than normative data from reading achievement measures. This decision was based on the fact that our reading achievement measures varied considerably in terms of when they had been normed and how applicable the norms were for our sample (see below). Because we oversampled at-risk children, we calculated local norms by using a weighting procedure. Recall that all kindergarteners in the school district were administered LNF and ISF during the first week of the school year to determine initial risk status. As a result, we were able to ascertain the percentage of children in the district who met our criteria for risk. Using the district rate (18.9%) and the rate in our sample (69.3%), we created a weighting variable that allowed us to adjust raw scores and to calculate percentile and standard scores that would be expected if all children in the district had been assessed on a given reading achievement measure. Weighted standard scores showed that our sample had a mean of 97.6 (SD=14.6) on the TOWRE-2 and a mean of 112.9 (SD=11.0) on the WRMT-R: NU Basic Skills Cluster. Standard scores were not available for the FAIR: ORF. The mean score for the TOWRE-2 was near the expected normative mean of 100 (SD=15), whereas the mean for the WRMT-R:NU was much higher than the expected score. Similar high scores on the WRMT-R:NU have been reported by others in recent studies (e.g., Al Otaiba et al, 2011). The discrepancy in the mean standard scores between the TOWRE-2 and the WRMT-R:NU is likely the result of when these measures were normed. The norms for the TOWRE-2 were published in 2011 whereas those for the WRMT-R:NU were published in 1998, prior to programs such as Reading First and Early Reading First that were directed at improving young children's word reading skills.

## Data Analysis

The first set of analyses were carried out to determine if a battery of screening measures administered at the beginning of kindergarten could accurately predict reading outcomes at the end of first grade. Because reading outcomes in school settings are often treated as binary in nature (i.e., reading disabled vs. non-disabled), we used binary logistic regression analysis. This analysis predicts a dichotomous dependent variable based on a set of independent variables. It provides a rank order of the relative importance of the predictor variables and the amount of variance in the dependent variable accounted for by these variables (Pseudo-$R^2$). In our initial set of models, we entered LNF3, ISF3, RAN, Nonword Repetition, and Sentence Imitation to predict outcomes in WRMT-R:NU Basic Skills, TOWRE-2, or FAIR: ORF. In subsequent models we replaced ISF3 with Sound Matching or Dynamic Screening of Phonological Awareness to investigate if models with alternative measures of phonological awareness were appreciably better. In further analyses, we examined if models including growth in LNF or ISF over the first 4 administrations (six weeks) provided a better prediction than models with only a single measurement of these assessments. In this latter analysis, we used linear growth scores (see above) from each individual's model-derived LNF and ISF growth trajectories over the first to fourth administrations.

To determine if response to instruction (as measured by growth in progress monitoring over the school year) added significantly to the predictive models, we used two different approaches In one approach, growth was explicitly modeled, and in the other, growth was estimated using an autoregression approach involving pretest-posttest residuals. We explicitly modeled growth for LNF1–6, LNF1–9, and ISF 1–6 by using each individual's growth model-derived slope scores. *Slope* scores were subsequently added to screening models and their impact on these models were evaluated. Growth model-derived *level* scores (centered administration 3) were not added to screening models for this set of analyses because raw LNF3 and ISF3 scores were already in the models to which these scores would be added. This also allowed a more direct comparison to the results of the second approach in which growth was examined. In second approach, we directly added raw scores for each of the progress monitoring measures (i.e., LNF6, LNF9, ISF6, Sound Matching 2) to the screening models. Because, each of the screening models constituted an autoregressor for the corresponding progress monitoring measure (i.e., contained ISF3 and/or LNF3), a significant entry by a given progress-monitoring measure could be attributed to growth in that measure. The latter approach to estimating growth has the advantage of being more easily applied since it does not require advanced modeling. On the other hand, the growth model approach has the advantage of potentially higher reliability due to the use of more than two time points and due to the statistical borrowing of information from other participants when predicting individual trajectories (Singer & Willett, 2003).

To further compare models and estimate the accuracy of prediction, results from logistic regression were used to calculate the area under the ROC curve for each logistic model (Metz, 1978; Swets, 1979). As noted above, an ROC curve is a plot of the true-positive rate (sensitivity) against the false-positive rate (1-specificty) for each of the cut points of a decision making instrument. The area under the curve or plot (AUC) can be used as an

overall estimate of the accuracy of the instrument. Measures of sensitivity and specificity provide estimates of the accuracy of an instrument using a given cut-score, while the AUC is an estimate of the accuracy across cut-points. As such, it is not unduly impacted by the selection of a specific cut-score. Differences in predictive accuracy of models can also be interpreted by evaluating AUC differences (Hanley & McNeil, 1983). A critical ratio $z$ is calculated between two AUCs and a value greater than 1.65 is designated as significant (one-tailed comparison). Critical ratio values are corrected for the correlation introduced by using the same sample of participants to derive two AUCs. In many cases, the models that were compared were nested models in which a model with one additional variable (e.g., LNF1–6 slope) was compared to a model without this variable. In such a case, the chi square statistic was used to judge the statistical significance of the difference.

Finally, weighted analyses were used when appropriate to control for the fact that we oversampled children who were at an increased risk for RD. As described above, we were able to use district-wide data to determine the likelihood that an at-risk or non at-risk child would have been selected randomly from our schools. This knowledge allowed us to create a weighting variable that when applied to our analyses, reduced the impact of oversampling of at-risk children and allowed us to better approximate the results that would have been obtained if we had used a large random sample from our school district or a district similar to it.

## Results

### Screening

Table 2 displays the correlations between screening, progress monitoring, and reading-outcome measures. All screening and progress-monitoring measures were significantly correlated with reading-outcome measures. Table 3 shows the results of weighted logistic regression analyses involving the screening battery. Recall that three initial versions of a screening model were run for each reading-outcome measure. These versions differed by which measure of phonological awareness (i.e., ISF3, Sound Matching, or Dynamic Screening of Phonological Awareness) was used. Only those with Sound Matching are shown in Table 3. This measure proved to be the best phonological awareness measure in predicting reading outcome. ISF3 did not add significantly to models predicting WRMT-R Basic Skills or TOWRE-2. It was a significant unique predictor in the model for FAIR: ORF (p=.015) but not as good a predictor as Sound Matching. The Dynamic Screening of Phonological Awareness was not a unique predictor in any model, regardless of the outcome. Results indicated that screening models demonstrated good-to-excellent prediction of each of the three reading-outcome measures. AUC values ranged from .85 to .92. For each of the models, LNF3 was the strongest predictor of reading outcomes followed by Sound Matching, Nonword Repetition, or RAN, depending on the specific model.

In a second set of models, we examined if the use of growth in LNF or ISF over the first 4 administrations provided a better prediction than did a single measurement of these assessments. In other words, we asked if the rate at which children made progress in letter naming or initial sound judgments in the first six weeks of school added anything more than a single measurement of these abilities during that time period. For these models, LNF1–4

and ISF1–4 model-derived *slope* scores were added to each of the models. Model-derived *level* scores were not added because of their redundancy with LNF3 or ISF3 (or its proxy Sound Matching) that was already in the initial screening models. We also wanted to directly compare models with growth to these initial models in a nested fashion. Results showed that neither LNF1–4 nor ISF1–4 slope added significantly to the initial screening models for any of the measures of reading outcome (p > .05). Models were also run with model-derived levels in addition to slopes and the results concerning the additive effects of slope were not appreciably different.

### Response to Instruction

Further analyses were carried out to determine if response to instruction over a longer period of time added significantly to the prediction of reading outcome over and above screening measures. Initially, out primary interest was the contribution of response to Tier 2 intervention in the prediction of reading outcome. Recall that a portion of the at-risk children were provided with Tier 2 instruction focusing on phonological awareness/letter knowledge and vocabulary/narration. The remainder of the at-risk children was placed in a control group that did not receive our Tier 2 intervention but was not restricted from receiving supplemental instruction as part of business as usual practice in the district.

Before examining the added contribution of response to Tier 2 intervention, it was necessary to investigate the impact of the intervention. To do this, we compared the performance of the at-risk intervention group to that of the at-risk control group and the typical control group on pretest-posttest measures of phonological awareness (i.e., Sound Matching, Sound Matching 2) and letter naming (i.e., LNF3, LNF9). Comparisons were restricted to these measures because they were most directly linked theoretically to our letter knowledge/phonological awareness intervention and to the word-reading outcomes examined in this study. Table 4 shows that the at-risk intervention group made gains in phonological awareness and letter knowledge but these gains were comparable to those of the at-risk control group. ANOVAs showed a significant pretest-posttest effect for phonological awareness, $F(1, 310) = 436.2$, $p<.001$, $\eta^2 = .585$ and letter naming, $F(1, 309) = 1205.8$, $p<.001$, $\eta^2 = .796$ and a significant group effect for phonological awareness, $F(2, 310) = 42.9$, $p<.001$, $\eta^2 = .217$ and letter naming, $F(2, 309) = 77.8$, $p<.001$, $\eta^2 = 335$. The interaction was not significant for phonological awareness $F(2, 310) = .00$, $p>.05$, $\eta^2 = .00$ but was significant for letter naming $F(2, 309) = 8.52$, $p<.001$, $\eta^2 = .052$. Follow-up Tukey HSD tests showed that the at-risk groups performed significantly less well than the typical control group at both pre- and posttest (p<.001). The significant interaction for letter naming resulted from less of a difference between at-risk groups and the typical control group at posttest than at pretest. Finally, the primary finding was that the at-risk intervention group did not differ significantly from the at-risk control group on either measure at either time point (p>.05). Thus, these results did not support an intervention effect for our Tier 2 instruction. There are several possible reasons for the lack of an intervention effect. One possible reason is that risk status was determined in the first week of school when some children might have underperformed because of their lack of familiarity with the school setting and/or testing materials. While there is no strong reason to believe that these effects might differentially impact groups (intervention vs. control), they could have obscured group differences. To

rule out this possibility, participants were re-classified as at risk based on LNF4 and ISF4 (week 7) scores. Group comparisons using only re-classified at-risk children again showed no significant differences in letter knowledge or phonological awareness following intervention (p>.05). Other possible reasons for a lack of an intervention effect will be considered in the discussion section.

The failure to find an intervention effect compromises potential conclusions that might be made about the additive effects of response to Tier 2 intervention in the prediction of reading outcomes. However, a lack of an intervention effect does not undermine our ability to address the important question of whether growth in general (i.e., growth related to Tier 2 intervention and/or classroom instruction) across the kindergarten year adds to prediction models. The notion of response to instruction is a basic tenet of the RTI model, and thus, evidence that growth in reading and/or reading-related skills predicts future reading outcomes would be an important finding. Therefore, we examined the contribution of growth in phonological awareness and letter naming for all participants across the kindergarten year.

As noted above, two approaches were used to determine if response to instruction and/or intervention added significantly to the prediction models. In both approaches, we used growth in phonological awareness and/or letter knowledge as measured from the beginning of the year to mid-year (January) or end of year (March/April) as the indicator of response to instruction. In one approach, growth was evaluated directly using growth model-derived slope scores. LNF1–6, ISF1–6, or LNF1–9 slopes were added to the initial screening model for each of the reading-outcome measures. ISF1–9 slope scores were not available because ISF was not administered by school personnel as part of the prescribed end-of-year progress monitoring. Again analyses were run without model-derived *level* scores added to screening models because LNF3 and ISF3 (or its proxy Sound Matching) were already in the models to which these scores would be added. Using LNF3 and ISF3 (or Sound Matching) rather than level scores also allowed for a more direct comparison to the results of the second approach that we used to examine growth (see below). Results shown in Table 5 indicated that ISF1–6 slope did not add significantly to any of the initial models. Chi-square analyses showed that models with ISF 1–6 were not significantly different than initial screening models. There was also no appreciable change in the AUC. On the other hand, when LNF1–6 or LNF1–9 model-derived slope scores were added to each of the screening models, there was a significant reduction in the log likelihood (p. < 001). There was also a significant increase in the AUC for each of the models containing LNF1–9 slope compared to those without this variable in the model (z = 1.80–2.20, p. <.05). Critical values for comparable comparisons involving LNF1–6 slope were significant for FAIR: ORF (z = 1.81, p. <.05), and approached but did not reach significance for WRMT-R Basic Skills (z = 1.54, p. >.05) and TOWRE-2 (z = 1.34, p. >.05).

In a second approach to examining the contribution of response to instruction, we added the raw scores for progress-monitoring measures from midyear (i.e., LNF6, ISF6) and end of year (LNF9, Sound Matching 2) to the initial screening models. Because each of the screening models constituted an autoregressor for the addition of the corresponding progress-monitoring measure, a significant effect by a given progress-monitoring measure

could be attributed to growth in that measure. Results showed that ISF6 did not add significantly to initial screening models ($p. > .05$). On the other hand, when either LNF6 or LNF9 was added to the initial models, there was a significant reduction in the log likelihood ($p. < .001$). There was also a significant increase in the AUC for each of the models containing LNF9 compared to those without this variable in the model ($z = 1.87–2.72$, $p. < .05$). Critical values for comparable comparisons involving LNF6 were significant for WRMT-R Basic Skills ($z = 2.11$, $p. < .05$), and FAIR ($z = 2.61$, $p. < .05$), and approached but did not reach significance for TOWRE-2 ($z = 1.54$, $p. > .05$). Adding Sound Matching 2 to the screening models also resulted in a significant reduction in the log likelihood for TOWRE-2 ($p. = .012$) and FAIR ($p. = .007$) but not WRMT-R Basic Skills ($p. > .05$). However, changes in the AUC were small and non-significant ($z = .65–.68$, $p. > .05$). In addition, Sound Matching 2 was not a significant predictor when entered in models along with LNF9 ($p. > .05$).

## Discussion

Early and accurate identification of children at risk for RD is critical for the prevention of RD within a RTI framework. Our results indicate that a screening battery containing a small number of assessments that were administered at the beginning of kindergarten accurately predicted word reading accuracy and/or fluency at the end of first grade. We further found that response to supplemental and/or classroom instruction, in some cases by January of the kindergarten year, added significantly to the prediction of reading outcomes.

Overall, our kindergarten screening battery performed well in classifying good and poor readers at the end of first grade. AUC values varied from .85 to .92 depending on which measure of reading achievement is considered. These AUC values are comparable to, or just slightly lower than, those of logistic regression models for first-grade screening that were reported by Compton et al. (2006). It may seem unlikely that a screening battery administered at the beginning of kindergarten would be nearly as accurate in predicting reading outcomes two years later as one given at the beginning of first grade. However, in recent years, preschool-age children have been exposed to increasing levels of literacy. This increase has come, in part, from national initiatives and funding such as Early Reading First that have brought scientifically-based literacy instruction to many preschool classrooms. There has also been an increasing trend for more children to be enrolled in full-day preschool classrooms (Aud et al., 2012). As a result, most children now arrive at kindergarten with some literacy knowledge. Thus, it is not surprising that kindergarten screening measures involving this knowledge, at least in part, accurately predict later reading outcomes. Because much of the increase in preschool literacy exposure or instruction has occurred in recent years, it may not be fair to compare our results to those of a first-grade screening study that took place more than six years earlier than ours. A better comparison would be possible if a study screened the same children in kindergarten and then again in first grade and compared the predictability of these screenings for a later reading outcome.

Among the screening measures, an assessment of letter knowledge (DIBELS: LNF) proved to be the strongest single predictor of reading outcomes. LNF had a moderate correlation

with first-grade reading achievement (.58 to .66) and was the strongest predictor in all of the screening models. Numerous other studies have found letter knowledge to be among the better predictors of early reading achievement (Catts et al., 2001; Schatschneider et al., 2004; Vellutino et al., 2008). In addition, several recent studies have specifically shown that LNF, when administered in kindergarten, was a significant predictor of first-grade reading achievement (e.g., Burke, Crowder, Hagan-Burke & Zou, 2009). Currently, many schools use LNF, along with other measures, for universal screening in kindergarten, and our results support this practice.

Our screening battery also contained a second DIBELS measure, ISF. Whereas this measure did add to the prediction of reading outcome in a few models, it did not perform as well as Sound Matching, which was one of the stronger predictors of reading outcomes. The most recent edition of DIBELS has replaced ISF with a similar measure, First Sound Fluency (FSF; Good et al., 2011). Some initial research suggests that FSF may be more reliable than ISF, and as such, could be an appropriate choice rather than Sound Matching for a phonological awareness measure in a kindergarten screening battery (Cummings, Kaminski, Good, & O'Neil, 2011). The Dynamic Screening of Phonological Awareness did not add significantly to any of the models. However, other research suggests that this measure may be useful as a secondary measure of phonological awareness in a gated screening model (Bridges & Catts, 2011).

We also found that measures of rapid naming (RAN) and nonword repetition provided unique prediction in some screening models. RAN explained additional variance for TOWRE-2, and Nonword Repetition explained additional variance for WRMT-R Basic Skills and FAIR: ORF. The relationship between RAN and reading achievement is well documented (Pennington, Cardoso-Martins, Green, & Lefty, 2001; Wolf et al., 2002), and RAN is often used in clinical or educational setting for diagnostic or screening purposes (e.g., Wiig, Zureich, & Chan, 2000). It is also important to note that RAN's contribution to predicting reading outcomes in the current study was likely reduced by including LNF, which clearly has a speed component. Nonword repetition has been employed much less often in screening batteries. Other investigations have found an association between nonword repetition and reading achievement (Baird, Slonims, Simonoff, & Dworzynski, 2011; Catts, Adlof, Hogan, & Weismer, 2005; van Weerdenburg, Verhoven, van Balkom, & Bosman, 2009). Diagnostic batteries have also included measures of nonword repetition for the assessment of language and reading disorders (Miles, 1982; Wagner, Torgesen, & Rashotte, 1999), but these measures, or comparable measures, have not generally been employed in universal screening. Some may question whether or not a nonword repetition task could be reliably scored by an examiner in a face-to-face setting. There is some indication that it can be for children 7–9 years of age (Bishop, North & Donlan, 1996) but further investigation is needed to determine if this will be the case for kindergarten-age children.

We also investigated the issue of whether short-term progress monitoring at the beginning of the year could add to the prediction provided by the screening battery. During the first six weeks of school, LNF and ISF were administered on four occasions (prior to Tier 2 intervention), approximately two weeks apart. Growth curves were estimated for

participants' performance across these measures, and growth model-derived scores were added to the prediction models. Our results showed that growth in LNF and ISF over the first six weeks did not add significantly to the prediction of reading outcomes. These results appear to indicate that it may not be how fast or slow children acquired letter knowledge or phonological awareness, but how much knowledge they have at the time of screening (e.g., third time point). Of course, the failure of growth to add to the prediction may have been the result of only considering growth over a very short period of time. Other data, discussed below, indicated that growth over a longer time span did add significantly to the prediction of reading outcomes.

In addition to universal screening at the beginning of the year, we were interested in whether growth in literacy skills across the year provided insight into reading outcomes. A primary tenet of the RTI model is that response to instruction offers useful information for early identification and prevention of RD. One of the goals of this study was to investigate the additive effects of response to Tier 2 intervention. To this end, we randomly assigned children with initial risk for RD to an intervention or control condition. Children in the intervention condition were provided with 26 weeks of supplemental instruction in phonological awareness and letter knowledge by our research team. Children in the control condition received business as usual practice, which could have included similar supplemental instruction by school personnel. All children received classroom instruction in literacy skills. Our results indicated that whereas children in the intervention condition did show growth in literacy skills, they did not outperform the at-risk control children. There may be several reasons why this occurred. First, the majority of the children in the at-risk control group (85%) were reported to have received supplemental instruction as part of business as usual practice in the schools. This instruction was provided by trained professionals in the schools, and while not generally as systematic as our instruction, it contained many of the same phonological awareness and letter knowledge activities. Thus, the effects of our intervention may have been masked by many in the at-risk control group also receiving supplemental intervention as part of business as usual practices.

An alternative explanation is that the intervention effects were diluted by high-quality classroom instruction. That is, because children regularly received a "heavy dose" of high-quality instruction in phonological awareness and letter naming as part of regular classroom instruction, they gained little benefit from the relatively brief Tier 2 intervention. Others have offered a similar argument for a lack of intervention effects under comparable conditions (Bailet, Repper, Piasta, & Murphy, 2009; Denton, Cirino, & Fletcher, 2010; Lonigan & Philips, 2009). The school district in which our study took place had a strong literacy curriculum in kindergarten and primary grades. The kindergarten curriculum consisted of a mandated 90-minute literacy block that included the use of an explicit and systematic phonological awareness and phonics program (Animated Literacy, Stone, 2006). In addition, teachers engaged children in read alouds, guided reading, and writing activities. A strong code-based emphasis was also inherent in the first-grade curriculum with a continuation of explicit instruction in phonics (Animated Literacy) and application of phonics during guided reading instruction (Fountas & Pinnell, 1996) and in supplemental Tier 2 interventions that were provided to children deemed to be at risk. As a result of the high-quality code-based instruction, children in the district appeared to be making good

progress in word reading skills. Recall, we found that first graders were scoring on average quite high on the WRMT-R Basic Skills Cluster. Thus, in light of the high-quality classroom instruction, it may have been difficult for children to benefit much from the relatively brief doses of intervention we provided. Both the intervention group and at-risk control group did make progress in literacy skills across the year, and in the case of letter naming, made somewhat more growth than the typical control group. Nevertheless, by year's end, at-risk groups were still performing, on average, nearly a standard deviation below the typical groups' performance. Additional evidence that our intervention in phonological awareness and letter naming may have been diluted by classroom instruction in these areas comes from other intervention results not reported here. Those results have demonstrated strong intervention effects for the portion of our intervention that focused on vocabulary and narration (Bridges, Catts, & Nielsen, 2012; Catts, Bridges, Nielsen, & Chan, 2011). Vocabulary and narration were not areas of instructional concentration in kindergarten classrooms in our school district. Thus, when our intervention was directed at skills/ knowledge that received little classroom instruction, it was more effective.

If Tier 2 intervention effects concerning phonological awareness and letter naming were in fact significantly diluted by classroom instruction, these results have implications for RTI in kindergarten. Specifically, they question the role of Tier 2 intervention directed at phonological awareness and letter naming under certain circumstances. They suggest that brief supplemental instruction in these skills may not lead to significant improvement and/or aid in the identification process when combined with high-quality Tier 1 classroom instruction. It may be necessary to implement more intensive intervention or tailor the intervention more specifically to a given child to achieve the benefits from Tier 2 intervention in these situations. Future investigations are needed to better understand the interplay between Tier 1 and Tier 2 instruction in kindergarten.

Whereas the lack of intervention effects (for phonological awareness and letter knowledge) compromised potential conclusions that we might have made about the additive effects of response to Tier 2 intervention, they did not limit our ability to address the important RTI tenet of whether response to instruction in general predicts reading outcome. All children in our study received explicit instruction in phonological awareness and letter knowledge as part of our Tier 2 intervention or business as usual supplemental instruction and/or classroom instruction. Thus, we were able to examine response to instruction/intervention across the kindergarten year in relation to later reading outcomes. Overall, our results indicated that growth in literacy skills, the index of response to instruction, predicted reading outcomes over and above that of the screening battery. Both growth in LNF and Sound Matching were significant predictors of reading achievement. However, when these were added together in models, only LNF growth was a unique predictor. ISF growth was not a significant predictor of reading achievement in any of our models. Other results showed that for two of our three reading outcomes (TOWRE-2 & FAIR), there were no significant differences between the additive effects of growth measured through January versus that measured through March/April. Finally, we found that when growth was assessed using the "autoregressor" approach as opposed to a growth model approach, it accounted for more unique variance in predicting reading outcomes. This finding is significant in that the

regression approach is much simpler to utilize and could more easily be put into practice in the field.

Our results concerning the usefulness of response to instruction in the prediction of reading achievement are generally in line with those of Vellutino et al. (2008). The primary difference between the studies is that Vellutino et al. explicitly examined response to Tier 2 intervention, whereas we investigated response to instruction more generally. Nevertheless, both studies showed that response to instruction during kindergarten added to the prediction of first grade reading achievement over and above initial kindergarten screening. Our results concerning kindergarten growth are also in line with those of Al Otaiba et al. (2011). This is the case even though the latter study looked at growth from a different perspective, and at first glance, appears to have different results. Recall, Al Otaiba and colleagues examined literacy skills at the end of kindergarten as well as growth in response to instruction (not specifically Tier 2 instruction) across the year. They reported that children who ended the year with higher reading and reading-related skills had better reading outcomes in first grade than those who had lower end-of-year skills. They also found that when end-of-year skills were controlled, children who grew the most had the poorest first-grade reading outcomes. We, on the other hand, found that children who grew the most had the best reading outcomes. However, we examined growth in relation to the beginning of kindergarten rather than the end of kindergarten. We found that children who had higher screening scores at the beginning of the year had better outcomes. But once beginning of the year scores were controlled, we found that children who grew more quickly had better reading outcomes. Because growing more quickly translates to higher end-of-year scores and having a higher initial score is associated with less growth, our results are actually the mirror image of those of Al Otaiba et al.

The observation that growth or response to instruction added to the prediction of reading outcomes is not surprising in the case of kindergarten children. Children likely come to kindergarten with varying levels of literacy skills and other cognitive abilities, and these skills and abilities influence future reading achievement. However, at the beginning of kindergarten, children most likely have not fully differentiated themselves in terms of their potential for reading achievement. They need more experience with literacy to show their potential in this regard. Our results suggest kindergarten provides that experience and that the differential growth that is observed can be useful in gauging risk for RD.

### Implications

Results of this study provide support for the use of an RTI model in the early identification of RD. Our findings show that universal screening at the beginning of kindergarten can identify children at risk for RD with an acceptable level of accuracy. Several of the measures we used already have widespread use in kindergarten screening batteries. These assessments might further be supplemented with other measures such as rapid naming and/or nonword repetition to improve their accuracy. Our findings also support an additional principle of RTI, that is, measures of growth in response to instruction provide useful information for forecasting reading outcomes. Our work further shows that in many cases growth through January is sufficient to capture this information, and waiting until the end of

the year to make additional instructional decisions may not be necessary. Thus, a combination of where a child starts at the beginning of kindergarten and where he/she is at midyear might give practitioners a good indication of risk for RD in first grade.

## Acknowledgments

## References

Al Otaiba S, Connor C, Lane H, Kosanovich ML, Schatschneider C, Dyrlund AK, et al. Reading First kindergarten classroom instruction and students' growth in phonological awareness and letter naming-decoding fluency. Journal of School Psychology. 2008; 46:281–314. [PubMed: 19083361]

Al Otaiba S, Folsom JS, Schatschneider C, Wanzek J, Greulich L, Meadows J, et al. Predicting first-grade reading performance from kindergarten response to tier I instruction. Exceptional Children. 2011; 77:453–470. [PubMed: 21857718]

Aud, S.; Hussar, W.; Johnson, F.; Kena, G.; Roth, E.; Manning, E.; Wang, X.; Zhang, J. The condition of education 2012 (NCES 2012-045). Washington, DC: U.S. Department of Education, National Center for Education Statistics; 2012. Retrieved [August 19, 2012] from http://nces.ed.gov/pubsearch

Bailet LL, Repper KK, Piasta SB, Murphy SP. Emergent literacy intervention for prekindergarteners at risk for reading failure. Journal of Learning Disabilities. 2009; 42:336–355. [PubMed: 19398614]

Baird G, Slonims V, Simonoff E, Dworzynski K. Impairment in non-word repetition: a marker for language impairment or reading impairment? Developmental Medicine and Child Neurology. 2011; 53:711–716. [PubMed: 21585364]

Bishop DVM, North T, Donlan C. Nonword repetition as a behavioural marker for inherited language impairment: Evidence from a twin study. Journal of Child Psychology and Psychiatry. 1996; 37:391–403. [PubMed: 8735439]

Blachman, BA.; Ball, EW.; Black, R.; Tangel, DM. Road to the Code: A Phonological Awareness Program for Young Children. Baltimore, MD: Brookes Publishing Company; 2000.

Breiman, L.; Friedman, J.; Olshen, R.; Stone, CJ. Classification and regression trees. New York: Chapman & Hall; 1984. 1984

Bridges, MS.; Catts, HW. Dynamic Screening of Phonological Awareness. East Moline, IL: LinguiSystems; 2010.

Bridges MS, Catts HW. The use of a dynamic screening of phonological awareness to predict risk for reading disabilities in kindergarten children. Journal of Learning Disabilities. 2011; 44:330–338. [PubMed: 21571700]

Bridges, MS.; Catts, H.; Nielsen, DC. Response to narrative instruction in Tier 2; Poster presented at the annual conference of the American Speech-Language-Hearing Association; Atlanta, GA. 2012.

Burke M, Crowder W, Hagan-Burke S, Zou Y. A comparison of two path models for predicting reading fluency. Remedial and Special Education. 2009; 30:84–95.

Catts HW, Adlof SM, Hogan TP, Weismer SE. Are specific language impairment and dyslexia distinct disorders? Journal of Speech, Language, and Hearing Research. 2005; 48:1378–1396.

Catts, HW.; Bridges, MS.; Nielsen, DC.; Chan, Y. Response to vocabulary instruction in Tier 2; Poster presented at the Pacific Coast Research Conference; San Diego, CA. Feb. 2011

Catts HW, Fey ME, Zhang X, Tomblin JB. Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. Language, Speech, and Hearing Services in Schools. 2001; 32:38–50.

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.

Cummings KD, Kaminski RA, Good RH, O'Neil M. Assessing phonemic awareness in early kindergarten: Development and initial validation of first sound fluency (FSF). Assessment for Effective Intervention. 2011; 36:94–106.

Denton, CA.; Cirino, P.; Fletcher, J. The impact of instructional variables on outcomes in Tier 2 first-grade reading intervention; Paper presented at the Pacific Coast Research Conference; San Diego, CA. 2010.

Dollaghan C, Campbell TF. Nonword repetition and child language impairment. Journal of Speech, Language, and Hearing Research. 1998; 41:1136–1146.

Felton RH. Early identification of children at risk for reading disabilities. Topics in Early Childhood Special Education. 1992; 12:212–229.

Florida Department of Education. Florida Assessments for Instruction in Reading (FAIR). Tallahassee, FL: 2009.

Fountas, IC.; Pinnell, GS. Guided reading: Good first teaching for all children. Portsmouth, NH: Heinemann; 1996.

Fuchs D, Fuchs LS. Peer-assisted learning strategies: Promoting word recognition, fluency, and comprehension in young children. Journal of Special Education. 2005; 39:34–44.

Fuchs LS, Vaughn S. Responsiveness-to-intervention: A decade later. Journal of Learning Disabilities. 2012; 45:195–203. [PubMed: 22539056]

Good, RH.; Kaminski, RA., editors. Dynamic Indicators of Basic Early Literacy Skills. 6th ed.. Eugene, OR: Institute for the Development of Educational Achievement; 2002.

Good, RH.; Kaminski, RA.; Cummings, K.; Dufour-Martel, C.; Petersen, K.; Powell-Smith, K.; Stollar, S.; Wallin, J. DIBLES Next Assessment Manual. Dynamic Measurement Group; 2011.

Haager, D.; Klinger, J.; Vaughn, S. Evidence-based reading practices for response to intervention. Baltimore, MD: Paul H. Brooks Publishing Company; 2007.

Hammill, DD.; Newcomer, PL. Test of Language Development-Primary-Third Edition. Austin, TX: Pro-Ed; 1997.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Diagnostic Radiology. 1983; 143:29–36.

Jenkins, JR. Candidate measures for screening at-risk students; Paper presented at the NRCLD Responsiveness-to-Intervention Symposium; Kansas City, MO. 2003.

Lovett MW, Steinbach KA, Frijters JC. Remediating the core deficits of developmental reading disability: A double deficit perspective. Journal of Learning Disabilities. 2000; 33:334–358. [PubMed: 15493096]

Metz CE. Basic principles of ROC analysis. Seminars in Nuclear Medicine. 1978; 8:283–298. [PubMed: 112681]

Miles, TR. The Bangor Dyslexia Test. Wisbech, England: Learning Development Aids; 1982.

National Institute of Child Health and Human Development. Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office; 2000.

O'Connor RE, Jenkins JR. Prediction of reading disabilities in kindergarten and first grade. Scientific Studies of Reading. 1999; 3:159–197.

Pennington BF, Cardoso-Martins C, Green PA, Lefly D. Comparing the phonological and double deficit hypotheses for developmental dyslexia. Reading and Writing: An Interdisciplinary Journal. 2001; 14:707–755.

Rabe-Hesketh, S.; Skrondal, A. Multilevel and longitudinal modeling using Stata. College Station, Tex: Stata Press Publication; 2008.

Ritchey, KD. Early identification of reading disabilities: How accurately can at-risk kindergarten children be identified?; Paper presented at the annual conference of the International Dyslexia Association; Philadelphia, PA. 2004.

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.

Schatschneider C, Fletcher JM, Francis DJ, Carlson CD, Foorman BR. Kindergarten prediction of reading skills: A longitudinal comparative analysis. Journal of Educational Psychology. 2004; 96:265–282.

Schuele; Dayton. Intensive Phonological Awareness Program. Nashville, TN: Authors; 2000.

Singer, JD.; Willett, JB. Applied longitudinal data analysis : modeling change and event occurrence. Oxford ; New York: Oxford University Press; 2003.

Speece DL, Mills C, Ritchey KD, Hillman E. Initial evidence that letter fluency tasks are valid indicators of early reading skill. The Journal of Special Education. 2003; 36:223–233.

Stone, J. The animated–alphabet story, song, and action book. La Mesa, CA: J. Stone Creations; 2006.

Swets JA. Analysis applied to the evaluation of medical imaging techniques. Investigative Radiology. 1979; 14:109–121. [PubMed: 478799]

Torgesen JK. The response to intervention instructional model: Some outcomes from a largescale implementation in Reading First schools. Child Development Perspectives. 2009; 3:38–40.

Torgesen, JK.; Wagner, RK.; Rashotte, CA. Test of Word Reading Efficiency-Second Edition. Austin, TX: PRO-ED; 2011.

van Weerdenburg M, Verhoeven L, Balkom H, Bosman A. Cognitive and linguistic precursors of early literacy achievement in children with specific language impairment. Scientific Studies in Reading. 2009; 13:484–507.

Vellutino FR, Scanlon DM, Zhang H, Schatschneider C. Using response to kindergarten and first grade intervention to identify children at-risk for long-term reading difficulties. Reading & Writing: An Interdisciplinary Journal. 2008; 21:437–480.

Wagner, RK.; Torgesen, JK.; Rashotte, CA. Comprehensive Test of Phonological Processes. Austin, TX: Pro-Ed; 1999.

Wiig EH, Zureich P, Chan NH. A clinical rationale for assessing rapid automatized naming in children with language disorders. Journal of Learning Disabilities. 2000; 33:359–370. [PubMed: 15493097]

Wolf M, O'Rourke AG, Gidney C, Lovett MW, Cirino P, Morris R. The second deficit: An investigation of the independence of phonological and naming-speed deficits in developmental dyslexia. Reading & Writing: An Interdisciplinary Journal. 2002; 15:43–72.

Woodcock, RW. Woodcock-Reading Mastery Tests-Revised/Normative Update. Circle Pines, MN: American Guidance Service; 1998.

**Table 1**

Screening, Progress Monitoring, and Reading-Outcome Measures

| Measure | Screening | Progress Monitoring | Reading Outcome |
|---|---|---|---|
| Letter Naming Fluency: DIBELS | X | X | |
| Initial Sound Fluency: DIBELS | X | X | |
| Sound Matching: CTOPP | X | X | |
| Dynamic Screening of Phonological Awareness | X | | |
| RAN:CTOPP | X | | |
| Nonword Repetition | X | | |
| Sentence Imitation: TOLD-2P | X | | |
| Woodcock Reading Mastery Tests-Revised: Basic Skills | | | X |
| Test of Word Reading Efficiency-2 | | | X |
| Florida Assessment of Instruction in Reading: Oral Reading Fluency | | | X |

**Table 2**

Correlations Between Kindergarten Screening, Progress Monitoring, and First Grade Reading-Outcome Measures

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. LNF3 | - | | | | | | | | | | | | | |
| 2. ISF3 | .411** | - | | | | | | | | | | | | |
| 3. LNF6 | .703** | .314** | - | | | | | | | | | | | |
| 4. ISF6 | .425** | .415** | .494** | - | | | | | | | | | | |
| 5. LNF9 | .687** | .335** | .830** | .445** | - | | | | | | | | | |
| 6. SM | .524** | .571** | .474** | .408** | .454** | - | | | | | | | | |
| 7. SM2 | .483** | .445** | .527** | .443** | .553** | .525** | - | | | | | | | |
| 8. DSPA | .388** | .509** | .373** | .404** | .380** | .507** | .468** | - | | | | | | |
| 9. RAN | -.460** | -.145* | -.459** | -.331** | -.406** | -.250** | -.264** | -.227** | - | | | | | |
| 10. SI | .416** | .424** | .413** | .385** | .378** | .449** | .438** | .451** | -.229** | - | | | | |
| 11. NWR | -.291** | -.441** | -.214** | -.269** | -.302** | -.257** | -.376** | -.299** | .229** | -.564** | - | | | |
| 12. WRMT-R Basic Skills | .581** | .403** | .577** | .375** | .633** | .475** | .560** | .405** | -.335** | .497** | -.486** | - | | |
| 13. TOWRE-2 | .651** | .398** | .670** | .415** | .700** | .442** | .568** | .398** | -.379** | .465** | -.441** | .905** | - | |
| 14. FAIR: ORF | .663** | .373** | .649** | .408** | .666** | .481** | .486** | .336** | -.372** | .493** | -.401** | .833** | .898** | - |

*Note.* LNF = Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Letter Naming Fluency; ISF = DIBELS: Initial Sound Fluency; SM = Comprehensive Test of Phonological Processing (CTOPP): Sound Matching; DSPA = Dynamic Screening of Phonological Awareness; RAN = CTOPP: Rapid Automatized Naming; SI = Test of Language Development-P:3 (TOLD-P:3): Sentence Imitation; NWR = Nonword Repetition; WRMT-R: Basic Skills = Woodcock Reading Mastery Tests-Revised: Basic Skill Cluster; TOWRE-2 = Test of Word Reading Efficiency-Second Edition; FAIR: ORF = Florida Assessment of Instruction for Reading: Oral Reading Fluency.

*
p < .05;

**
p < .01

**Table 3**

Logistic Regression Analyses for Screening Measures

| | B | SE | Wald | p | Cox & Snell R² | Nagelkerke R² | AUC | SE |
|---|---|---|---|---|---|---|---|---|
| **Model 1: WRMT-R** | | | | | 0.261 | 0.409 | 0.852 | 0.026 |
| LNF3 | 0.078 | 0.019 | 17.449 | <0.001** | | | | |
| SM | 0.219 | 0.070 | 9.651 | 0.002** | | | | |
| NWR | −0.070 | 0.026 | 7.268 | 0.007** | | | | |
| RAN | −0.003 | 0.005 | 0.361 | 0.548 | | | | |
| SI | −0.011 | 0.041 | 0.071 | 0.790 | | | | |
| Constant | 0.805 | 1.046 | 0.593 | 0.441 | | | | |
| **Model 2: TOWRE-2** | | | | | 0.363 | 0.575 | 0.921 | 0.019 |
| LNF3 | 0.172 | 0.031 | 31.618 | <0.001** | | | | |
| SM | 0.172 | 0.080 | 4.564 | 0.033* | | | | |
| NWR | −0.009 | 0.030 | 0.091 | 0.763 | | | | |
| RAN | −0.013 | 0.006 | 4.603 | 0.032* | | | | |
| SI | 0.086 | 0.052 | 2.659 | 0.103 | | | | |
| Constant | −0.834 | 1.207 | 0.477 | 0.490 | | | | |
| **Model 3: FAIR: ORF** | | | | | 0.257 | 0.413 | 0.856 | 0.025 |
| LNF3 | 0.080 | 0.020 | 16.229 | <0.001** | | | | |
| SM | 0.209 | 0.073 | 8.159 | 0.004** | | | | |
| NWR | −0.077 | 0.027 | 8.182 | 0.004** | | | | |
| RAN | −0.007 | 0.005 | 1.666 | 0.197 | | | | |
| SI | −0.015 | 0.043 | 0.119 | 0.730 | | | | |
| Constant | 1.565 | 1.085 | 2.080 | 0.149 | | | | |

Note. LNF3 = Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Letter Naming Fluency; SM = Comprehensive Test of Phonological Processing (CTOPP): Sound Matching; RAN = CTOPP: Rapid Automatized Naming; SI = Test of Language Development-P:3 (TOLD-P:3): Sentence Imitation; NWR = Nonword Repetition; WRMT-R: Basic Skills = Woodcock Reading Mastery Tests-Revised: Basic Skill Cluster; TOWRE-2 = Test of Word Reading Efficiency-Second Edition; FAIR: ORF = Florida Assessment of Instruction for Reading: Oral Reading Fluency.

*
  p < .05;

**
  p < .01

**Table 4**

Pretests and Posttests Scores on the Sound Matching and Letter Naming Fluency by Group

| | At Risk Intervention (N = 129) | | At Risk Control (N = 88) | | Typical Control (N = 96) | |
|---|---|---|---|---|---|---|
| | **Pretest** | **Posttest** | **Pretest** | **Posttest** | **Pretest** | **Posttest** |
| | **M (SD)** | **M (SD)** | **M (SD)** | **M (SD)** | **M (SD)** | **M (SD)** |
| SM-SM2 | 3.05 (2.2) | 8.98 (5.0) | 3.12 (2.0) | 9.06 (5.2) | 7.29 (5.4) | 13.21 (5.6) |
| LNF3-LNF9 | 7.12 (6.7) | 35.54 (15.1) | 5.88 (5.6) | 36.31 (13.4) | 27.25 (18.3) | 49.81 (15.7) |

*Note.* SM = Sound Matching (pretest); SM2 = Sound Matching 2 (posttest); LNF3 = Letter Naming Fluency (pretest); LNF9 = Letter Naming Fluency (posttest).

**Table 5**

Logistic Regression Analyses for Screening Measures and Response to Instruction/Intervention

| | Log likelihood | p | Cox & Snell R² | Nagelkerke R² | AUC | SE |
|---|---|---|---|---|---|---|
| Model 1: WRMT-R | 224.314 | | 0.261 | 0.409 | 0.852 | 0.026 |
| Model 1 + ISF 1–6 slope | 222.713 | 0.206 | 0.265 | 0.415 | 0.858 | 0.025 |
| Model 1 + LNF1–6 slope | 209.393 | <0.001 | 0.296 | 0.462 | 0.874 | 0.024 |
| Model 1 + LNF1–9 slope | 196.176 | <0.001 | 0.325 | 0.508 | 0.891 | 0.023 |
| Model 1 + ISF6 | 223.781 | 0.465 | 0.261 | 0.408 | 0.852 | 0.026 |
| Model 1 + LNF6 | 204.901 | <0.001 | 0.305 | 0.476 | 0.886 | 0.023 |
| Model 1 + LNF9 | 188.477 | <0.001 | 0.341 | 0.532 | 0.903 | 0.021 |
| Model 1 + SM2 | 221.050 | 0.071 | 0.269 | 0.421 | 0.857 | 0.026 |
| Model 2: TOWRE-2 | 171.072 | | 0.363 | 0.575 | 0.921 | 0.019 |
| Model 2 + ISF 1–6 slope | 170.425 | 0.421 | 0.364 | 0.577 | 0.922 | 0.019 |
| Model 2 + LNF1–6 slope | 157.382 | <0.001 | 0.390 | 0.618 | 0.934 | 0.017 |
| Model 2 + LNF1–9 slope | 147.007 | <0.001 | 0.410 | 0.650 | 0.943 | 0.016 |
| Model 2 + ISF6 | 170.926 | 0.702 | 0.363 | 0.574 | 0.921 | 0.019 |
| Model 2 + LNF6 | 148.844 | <0.001 | 0.407 | 0.643 | 0.940 | 0.017 |
| Model 2 + LNF9 | 141.199 | <0.001 | 0.421 | 0.666 | 0.946 | 0.016 |
| Model 2 + SM2 | 164.796 | 0.012* | 0.375 | 0.595 | 0.926 | 0.018 |
| Model 3: FAIR: ORF | 213.003 | | 0.257 | 0.413 | 0.856 | 0.025 |
| Model 3 + ISF 1–6 slope | 212.257 | 0.388 | 0.259 | 0.415 | 0.858 | 0.025 |
| Model 3 + LNF1–6 slope | 196.082 | <0.001 | 0.297 | 0.475 | 0.883 | 0.023 |
| Model 3 + LNF1–9 slope | 193.352 | <0.001 | 0.303 | 0.485 | 0.890 | 0.022 |
| Model 3 + ISF6 | 212.255 | 0.387 | 0.258 | 0.413 | 0.856 | 0.025 |
| Model 3 + LNF6 | 184.423 | <0.001 | 0.322 | 0.515 | 0.902 | 0.021 |
| Model 3 + LNF9 | 177.283 | <0.001 | 0.321 | 0.521 | 0.902 | 0.021 |
| Model 3 + SM2 | 205.783 | 0.007** | 0.274 | 0.440 | 0.862 | 0.024 |

*Note.* p = alpha level for chi square test comparing log likelihoods of screening models (Model 1–3) to models including response to instruction. LNF = Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Letter Naming Fluency; ISF = DIBELS: Initial Sound Fluency; SM2 = Comprehensive Test of Phonological Processing (CTOPP): Sound Matching; WRMT-R = Woodcock Reading Mastery Tests-Revised: Basic Skill Cluster; TOWRE-2 = Test of Word Reading Efficiency-Second Edition; FAIR: ORF = Florida Assessment of Instruction for Reading: Oral Reading Fluency