School Psychology Review,
2016, Volume 45, No. 3, pp. 310–326

Evaluating the Interpretations and Use of Curriculum-Based Measurement in Reading and Word Lists for Universal Screening in First and Second Grade

Stacy-Ann A. January
University of South Carolina

Scott P. Ardoin
University of Georgia

Theodore J. Christ
University of Minnesota

Tanya L. Eckert
Syracuse University

Mary Jane White
University of Minnesota

Abstract. Universal screening in elementary schools often includes administering curriculum-based measurement in reading (CBM-R); but in first grade, nonsense word fluency (NWF) and, to a lesser extent, word identification fluency (WIF) are used because of concerns that CBM-R is too difficult for emerging readers. This study used Kane's argument-based approach to validation as a framework to evaluate the interpretations and use of scores resulting from screening 257 first- and second-grade students. First, scores from three word lists (decodable WIF, high-frequency WIF, and whole-word NWF) were examined as indicators of reading achievement. Then, the use of these word list scores was evaluated regarding their ability to classify at-risk readers accurately and as supplements to

---

Theodore J. Christ, PhD, has equity and royalty interests in, and will serve on the Board of Directors for, FastBridge Learning (FBL), a company involved in the commercialization of the Formative Assessment System for Teachers (FAST). The University of Minnesota also has equity and royalty interests in FBL. These interests have been reviewed and managed by the University of Minnesota in accordance with its conflict-of-interest policies.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A090038 to the University of Minnesota. The opinions expressed are those of the authors and do not necessarily represent the official views of the Institute or the U.S. Department of Education.

Correspondence concerning this article should be sent to Stacy-Ann A. January, Department of Psychology, University of South Carolina, 1512 Pendleton St., Columbia, SC 29208; e-mail: sajanuary@sc.edu

CBM-R during the winter universal screening period. Participants were also concurrently administered a norm-referenced measure of early reading skills and global reading achievement. Results suggested that the word lists were good indicators of reading achievement and provided support for using CBM-R or a word list in conjunction with CBM-R to discriminate among at-risk readers. Findings have implications for the administration of universal screeners in first and second grade.

Universal screening, a core component of a Multi-Tiered System of Supports framework, is used for early identification of students who may be at risk for learning disabilities. Resultant data are used to inform early intervention, which is an effective approach to prevent reading difficulties (Vellutino, Scanlon, Small, & Fanuele, 2006). Curriculum-based measurement in reading (CBM-R) and nonsense word fluency (NWF) are often used for universal screening in the early elementary grades (Deno et al., 2009). Word identification fluency (WIF), albeit less frequently used, is another universal screening measure available to schools. Although there are clear benefits to administering CBM-R, NWF, and WIF, there are limitations associated with the use of NWF and WIF, and concerns about the ability of NWF scores to classify at-risk early readers accurately (Clemens, Shapiro, & Thoemmes, 2011).

CURRICULUM-BASED MEASUREMENT IN READING

CBM-R is a task in which students read aloud from grade-level text as the examiner listens and records their performance to estimate oral reading rate, which is typically reported in the metric of words read correctly per minute (WRCM). One benefit of administering CBM-R is that as a general outcome measure, it indexes global reading performance across the academic year, instead of measuring the specific, hierarchically organized subskills of reading (Fuchs & Deno, 1991). Although many published studies exist indicating that CBM-R is useful for universal screening (January & Ardoin, 2015; Kilgus, Methe, Maggin, & Tomasula, 2014; Reschly, Busch, Betts, Deno, & Long, 2009), the procedure requires students to integrate the many components of skilled reading required to read connected text (Fuchs, Fuchs, Hosp, & Jenkins, 2001), including decoding and word identification. However, many students in the early elementary grades are not yet prepared to read connected text, so the task may be too difficult and may result in poor classification accuracy for emerging readers who are at risk for developing reading disabilities (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009; Hosp, Hosp, & Dole, 2011). It is potentially for this reason publishers of curriculum-based measurement (CBM) probes recommend that the earliest CBM-R should be administered for universal screening is in the winter of first grade, and even then, NWF should be administered in conjunction with CBM-R for the remainder of the year (Good & Kaminski, 2007; Pearson, 2012).

NONSENSE WORD FLUENCY

In contrast to CBM-R, NWF is a subskill mastery measure that combines sound identification and blending of vowel–consonant (VC) and consonant–vowel–consonant (CVC) pseudowords to measure students' letter–sound correspondence, decoding skills, and progress as emerging readers (Good, Baker, & Peyton, 2009). Evidence indicates NWF scores account for a large portion of the variance in word reading and pseudoword decoding (Burke & Hagan-Burke, 2007; Oslund et al., 2012). Research also has demonstrated that NWF scores have moderate to strong concurrent and predictive associations with CBM-R performance (Burke & Hagan-Burke, 2007; Cummings, Dewey, Latimer, & Good, 2011; Harn, Stoolmiller, & Chard, 2008) and reading achievement (Fien et al., 2008, 2010).

Given that NWF is a decoding task, students can use different approaches to correctly decode each word. That is, students are able to say the individual sounds in each word, partially blend the word, or say the word as a unit. Thus, NWF has the potential of providing more information about students' decoding skills and potential risk for reading problems than other CBM measures. For instance, students who decode pseudowords as units (as opposed to sound by sound or partial blending with or without recoding) score higher on NWF probes and subsequent measures of oral reading (Harn et al., 2008). Furthermore, students who blend nonsense words as units generally have better phonemic skills and have improved automaticity than students who decode the individual letter sounds or use a combination of strategies (Cummings et al., 2011; Harn et al., 2008).

Despite the potential of gaining more descriptive information about students' decoding skills, a limitation is introduced when students use different decoding strategies. More specifically, variability in the strategies used results in a lack of consistency in the skill or construct measured within and across NWF assessments (Ritchey, 2008), which may affect its relation to measures of reading achievement (Harn et al., 2008). Therefore, by allowing students to choose their decoding strategy, educators cannot be certain which decoding skill (e.g., unitization, letter–sound correspondence) is measured by the NWF probes they administer.

Another limitation of existing NWF research is that it has almost exclusively examined the utility of NWF measures for assessing kindergarten and first-grade students' skills, despite that decoding skills continue to be an important element of reading instruction beyond these grade levels, particularly for struggling readers. As such, the potential benefit of using NWF scores to differentiate at-risk second-grade students has not been examined empirically. Extant research indicates that for students in kindergarten, scores from NWF adequately discriminate between those who do and those who do not later meet oral reading benchmarks (Clemens, Hilt-Panahon, Shapiro, & Yoon, 2012), but for first-grade students, NWF scores fail to predict which students later underachieve in reading (Clemens et al., 2011; Vanderwood, Linklater, & Healy, 2008). The inability of NWF performance to discriminate among poor readers in first grade and the lack of NWF research in later grades may be due to existing NWF probes assessing a narrow set of skills (i.e., decoding VC and CVC pseudowords). It is possible that NWF probes that include more complex word types, such as consonant–vowel–consonant–*e* (e.g., *vate*), have the potential to provide educators with information about students' advanced phonics and decoding skills, as well as better discriminate among at-risk readers.

WORD IDENTIFICATION FLUENCY

WIF is yet another alternative for universal screening. WIF probes require that students read a list of high-frequency and/or decodable words in 1 min, directly measuring students' accuracy and speed of real word reading (Fuchs, Fuchs, & Compton, 2004). Existing research suggests moderate to strong associations between WIF scores and performance on norm-referenced measures of word identification and decoding, passage reading, and reading achievement (Clemens et al., 2011; Fuchs et al., 2004; Zumeta, Compton, & Fuchs, 2012). Additionally, research by Clemens et al. (2011) and Fuchs et al. (2004) suggested that in the fall of first grade, WIF, as compared with NWF, better predicts later reading achievement and is a better indicator of risk for reading difficulties. Furthermore, although WIF was the single most accurate early literacy measure for identifying first-grade students at risk for reading problems, adding one or two additional early reading measures, such as NWF or phoneme segmentation fluency, provided a more accurate screening battery that identified first-grade students at risk for reading failure (Clemens et al., 2011). Both Clemens et al. (2011) and Fuchs et al. (2004) used investigator-developed WIF probes consisting of words sampled from popular high-frequency word lists (e.g., Dolch Word List) made up of both decodable

and nondecodable words. Thus, the extent to which the WIF probes used in those studies measured students' decoding skills likely varied (Ritchey, 2008).

Despite WIF demonstrating superiority over other first-grade CBM measures such as NWF, there are disadvantages that preclude its widespread use. First, unlike NWF, structured, reliable, and valid WIF probes are not available from most publishers of CBM probes. Therefore, educators must resort to developing their own measures of high-frequency word reading or simply using generic lists that are available (e.g., from interventioncentral.org). Although educator-developed high-frequency word lists may provide valuable information, they lack structure, are not validated as indicators of reading achievement, lack adequate norms to compare student performance for benchmarking, and do not have equivalent forms for progress monitoring. Furthermore, if structured and validated WIF probes were developed and made widely available, educators could have greater confidence when using them to make decisions about which students are not meeting reading benchmarks. Unfortunately, to date, researchers have not examined (a) whether there is any added benefit of concurrently administering WIF or NWF probes with CBM-R or (b) whether there is any benefit to administering WIF probes that are composed solely of decodable words.

AN ARGUMENT-BASED APPROACH TO VALIDATION

Kane's (2013a, 2013b) argument-based approach to validation is a practical framework for evaluating the decisions that are made based on observed (test) scores, including results from universal screenings. This framework posits that the interpretations and uses of observed scores must be explicitly stated (referred to as the *interpretation/use argument* [IUA]) and then evaluated systematically (validation). When the IUA and its assumptions are sufficiently supported by evidence, the uses and interpretations of test scores can be regarded as valid. The IUA includes a set of three hierarchically organized inferences that should be examined empirically: scoring, generalization, and extrapolation. *Scoring inferences* are based on the process by which an observed performance (e.g., a student reading connected text aloud) is transformed into an observed score (e.g., WRCM) through scoring rules (e.g., a word that is misread counts as an error). Evidence (i.e., validation) that scoring rules are applied appropriately includes adequate interscorer agreement/interrater reliability. The *generalization inference* refers to the assumption that scores at one point in time generalize across several observation conditions (e.g., occasions, raters). Reliability metrics such as the α coefficient, alternate-form reliability, and test–retest reliability provide evidence of the generalizability of observed scores. *Extrapolation inferences* consider how well the observed score indicates performance in a larger domain, either concurrently or in the future. An example of a validated extrapolation inference is that CBM-R scores are indicative of the larger domain of reading achievement (e.g., January & Ardoin, 2015; Reschly et al., 2009). Indeed, the argument-based approach is well suited as a framework for validating the interpretations and use of results from universal screening in schools (Christ & Nelson, 2014).

In the case of universal screening in reading, the IUA is all of the inferences and decisions that are made based on the resultant data. That is, it is assumed that (a) scoring rules were applied accurately (scoring inference), (b) observed scores generalize across observations (generalization inference), and (c) scores from universal screening assessments are indicators of students' reading achievement (extrapolation inference). On the basis of these *inferences,* universal screening data are *used* to make decisions regarding whether a student may benefit from more intensive instruction in reading. Schools typically conduct universal screenings three times per year; thus, inferences and decisions are made based on the resultant data from each universal screening period. Therefore, it is important that the interpretations and use of screening data are validated within the context of universal screening.

THE CURRENT STUDY

The present study aimed to replicate and extend the existing universal screening literature by examining procedures for evaluating early elementary students' achievement in reading. The current study extended this research by using NWF probes that measured skills beyond the decoding of VC and CVC words and requiring students to read the pseudowords that make up NWF probes as units. By requiring students to read nonwords as units, NWF probes assess the same skill for all students, as opposed to data reflecting some students' letter–sound knowledge and other students' blending skills. We also added to the existing literature by evaluating students' word-reading fluency on a WIF probe consisting of solely decodable words, in addition to a probe consisting of high-frequency words. This is in contrast to previous studies (e.g., Clemens et al., 2011; Fuchs et al., 2004; Zumeta et al., 2012), in which a single WIF probe consisted of both decodable and nondecodable high-frequency words. By administering both a solely decodable word list and a high-frequency word list, we explored whether the type of words used in WIF probes is meaningful in predicting students' reading achievement.

The current study used Kane's (2013a, 2013b) argument-based framework to evaluate the validity evidence for universal screening data in the early elementary grades. Although evidence for the scoring and generalization inferences is not a primary focus of this study, this information will be presented in the Method section. Thus, the interpretation of interest is as follows: NWF and WIF scores are good indicators of the larger domain of reading achievement for first- and second-grade students (i.e., extrapolation inferences). Previous research suggests a moderate to strong relation between the subskills of decoding (as measured by NWF) and word-reading skills (as measured by WIF) with students' reading achievement, as measured by their performance on CBM-R probes and norm-referenced measures. Therefore, the first purpose of this study was to evaluate whether NWF and WIF scores are adequate indicators of word analysis skills (i.e., decoding, phonological awareness) and global reading achievement, as measured by CBM-R and a nationally norm-referenced test that was administered concurrently. We also determined which word list (WIF, NWF) was a better indicator of early reading skills and reading achievement.

Kane's (2013a, 2013b) argument-based approach to validation was also used to evaluate the use of universal screening data to identify students who are at risk. Extant research examining the classification accuracy of NWF and WIF scores has suggested that WIF might be more accurate for identifying at-risk readers in first grade (Clemens et al., 2011) and has questioned the utility of administering CBM-R when screening early readers (Catts et al., 2009; Hosp et al., 2011). However, because it is a general outcome measure, CBM-R may be most appropriate for universal screening, instead of using subskill mastery measures that assess the component skills of reading. Thus, the second purpose of the current study was to evaluate the accuracy of the decisions that are made with CBM-R, NWF, and WIF scores regarding whether first- and second-grade students are at risk for reading difficulties. We were also interested in determining if classification accuracy could be improved when either an NWF or WIF probe is administered in conjunction with CBM-R. To address the second purpose of this study, we evaluated the classification accuracy of each screening measure alone and then with CBM-R to identify at-risk students, as measured by a concurrently administered norm-referenced measure of global reading achievement.

METHOD

Potential participants were initially recruited to be a part of a large study validating CBM-R for universal screening and progress monitoring in Grades 1–5 (Pratt et al., 2011; White et al., 2011). For the purposes of the current study, data were collected as part of two elementary schools' routine assessment procedures. Participating schools were part of

two school districts located within the Southeastern United States. There were 10 first-grade classrooms (4 in School A, 6 in School B) and 9 second-grade classrooms (4 in School A, 5 in School B) represented in this study. School-wide, 19% of students in School A qualified for free or reduced-price meals and approximately 71% of students in School B qualified for free or reduced-price meals.

Participants

All students ($N = 287$) who were present during the winter universal screening window were recruited as participants. However, students who were English learners (5 in first grade, 25 in second grade) were excluded from the analyses for this study because of potential bias in using CBM-R scores for universal screening (Hosp et al., 2011). Thus, all remaining participants ($n = 257$) were native English speakers. There were 135 first-grade students (69 from School A, 66 from School B), who were primarily male (59.3%) and ranged in age from 6.41 to 8.31 years ($M = 7.02$ years, $SD = 0.38$ years). The racial and ethnic composition of the first-grade students was 82.2% White, 6.7% African American, 4.4% Hispanic or Latino, 2.2% Asian, and 4.4% other or not specified. Approximately 3.7% of first-grade students were eligible for special education services. Just over half of the 122 second-grade students (60 from School A, 62 from School B) were male (54.1%); the second-grade students ranged in age from 6.85 to 9.27 years ($M = 7.99$ years, $SD = 0.41$ years). The racial and ethnic composition of the second-grade students was 73.8% White, 8.2% Hispanic or Latino, 7.4% African American, 4.1% Asian, and 6.6% other or not specified. Of the second-grade students, 5.7% were eligible for special education services.

Measures

All participants were administered two CBM-R probes; one decodable WIF probe (WIF-D); one high-frequency WIF probe (WIF-HF); a whole-word NWF probe (NWF-whole) that required blending words as units; and the Iowa Test of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2001). The word lists used in this study are similar to those developed and published by a screening and progress-monitoring assessment system (Christ et al., 2014) with measures demonstrating adequate reliability and validity. Unless otherwise noted, the dependent measure for each universal screener was WRCM.

Decodable WIF

The authors developed separate WIF-D probes for each grade, with each list consisting of 304 phonetically-regular words. In the development of both lists, decodability guidelines set forth by Menon and Hiebert (1999) were employed, which include CV words at Level 1; the words become increasingly difficult, based on linguistic decoding patterns, with multisyllabic words at Level 8. The first-grade WIF-D consisted of words that met the guidelines for decodability Levels 1 through 5. For example, words on the first-grade WIF-D included *pop* (Level 2), *dent* (Level 3), *cape* (Level 4), and *breeze* (Level 5). The second-grade list included 159 words from Levels 1–5 and 145 words from Levels 6 (e.g., *car*), 7 (e.g., *south*), and 8 (e.g., *problem*). Evidence for the generalization inference for the WIF-D probe is reflected in adequate internal consistency ($\alpha = .98$), test–retest reliability ($r = .94$), and alternate-form reliability ($r = .94$; Christ et al., 2014).

High-Frequency WIF

Separate first- and second-grade WIF-HF probes were also developed by the authors, with each word list consisting of 304 high-frequency words that were decodable and non-decodable. Words were selected from two commonly used high-frequency word lists, the 315 Dolch Word List (Johns, 1971) and the New Instant Word List (Fry, 1980). The WIF-HF probe for each grade included words from both lists, but the second-grade list included a greater number of less frequent words from the New Instant Word List. WIF-HF probes have adequate alternate-form reliability ($r = .94$), internal consistency ($\alpha = .99$), and test–retest reliability ($r = .97$), providing ev-

idence of the generalization inference (Christ et al., 2014).

Whole-Word NWF

Similar to the WIF-D probes, separate grade-level NWF-whole probes were developed using the decodability levels outlined by Menon and Hiebert (1999). NWF-whole probes consisting of 304 decodable pseudowords were developed for each grade level. The first-grade probe included 304 pseudowords from Levels 1–5, and the second-grade probe included pseudowords from Levels 2–8. The generalization inference for NWF-whole probes is supported by adequate alternate-form reliability ($r = .85$), test–retest reliability ($r = .76$), and internal consistency ($\alpha = .96$; Christ et al., 2014).

Curriculum-Based Measurement in Reading

First-grade students were administered two CBM-R probes—one investigator-developed preprimer probe and one first-grade level CBM-R probe—selected from the easyCBM passage set (www.easycbm.com). The preprimer probe developed by the authors included 88 unique words (258 total words), 57% of which were high-frequency words. Second-grade students were administered the first-grade level probe that was administered to the first-grade students and a second-grade level probe from the easyCBM passage set. By administering a passage that was below grade level, we were attempting to increase the possibility that CBM-R scores could be used to distinguish among struggling students. We hoped that an easier passage might result in greater differences among those students who had difficulty reading their grade-level passage. Furthermore, given the number of probes that were administered to participants, we administered only one grade-level and one below grade-level CBM-R probe as opposed to the three CBM-R probes that are traditionally administered as part of universal screenings. Additionally, previous research suggests that administering one CBM-R probe instead of three is appropriate for universal screening purposes (Ardoin et al., 2004). The reliability and validity of easyCBM passages are adequate (Jamgochian et al., 2010; Lai et al., 2010), and are similar to other commonly used CBM probes. The average WRCM across the two probes was used as the dependent measure.

Iowa Test of Basic Skills

The ITBS is a group-administered and nationally norm-referenced assessment for kindergarten through eighth-grade students (Hoover et al., 2001). Students were administered either Form A, Level 7 (first grade) or Form A, Level 8 (second grade). For the purposes of this study, the ITBS–Total Reading composite (ITBS-TR), which estimates students' vocabulary and reading comprehension skills, and the ITBS–Word Analysis subtest (ITBS-WA), which assesses students' phonological awareness, decoding, and understanding of word parts, were used. The ITBS-WA was selected for the current study given that it measures students' early reading skills. The ITBS-TR and ITBS-WA have adequate Kuder–Richardson Formula 20 internal consistency in first grade (.93 and .85, respectively) and second grade (.94 and .85, respectively; Hoover et al., 2001). The content-related validity of the ITBS was established through an extensive development process that included a curriculum review, preliminary item tryout, national item tryout, fairness review, and development of individual tests (Hoover et al., 2001).

The current study used ITBS Developmental Standard Scores (SSs) as the dependent measure. Developmental SSs were created using 200 as the median score for fourth-grade students and 250 as the median score for eighth-grade students. Thus, students' SSs indicate their performance along an achievement continuum from kindergarten through Grade 8. In the standardization sample, first-grade students' ITBS-TR SSs averaged 151.3 ($SD = 13.15$) and ITBS-WA SSs averaged 152.2 ($SD = 18.4$). For the second-grade students in the standardization sample, ITBS-TR SSs averaged 170.0 ($SD = 19.1$) and ITBS-WA SSs averaged 171.0 ($SD = 23.7$).

Procedures

Students were administered the ITBS during winter of the academic year by their classroom teachers, who followed standardized administration procedures. Within 1 week of ITBS administration, examiners individually administered the WIF-D, WIF-HF, NWF-whole, and CBM-R probes in random order, counterbalanced across all participants during one session. For the CBM-R probes, standardized administration and scoring procedures were followed as students were instructed to read across the page and down, were instructed to do their best reading, and were instructed that if they did not know a word, it would be told to them. Substitutions, skipped words, misread words, and words that were not read within 3 s were counted as errors and used to calculate WRCM. With the exception of students being told they would be reading a list of words, the administration and scoring procedures were identical for the WIF probes. NWF-whole administration procedures were modified from typical NWF procedures. That is, students were told they would be reading a list of pseudowords and were instructed to read the words as whole words and not sound by sound. To ensure that students understood the instructions, they were administered practice items and were provided with corrective feedback prior to being administered the word list. Scoring procedures were modified also, as only pseudowords read accurately as units were scored as correct. WRCM for the NWF-whole task was calculated by subtracting the total number of words read by the total number of errors (i.e., words read sound-by-sound, skipped words, misread words, and words that were not read within 3 s).

Procedural Integrity and Interscorer Agreement

Examiners were school psychology graduate students and undergraduate research assistants who participated in an hour-long training session led by the second author. Examiners were trained until they were 100% reliable on three consecutive probes. Prior to collecting data independently, examiners observed the second author complete an administration, were observed as they conducted an administration, and then were provided with feedback. If examiners completed 100% of the procedures accurately, they transitioned to collecting data independently. Otherwise, on-site training procedures were repeated until examiners accurately completed all required steps. All experimental sessions were audio recorded, and recordings were used to calculate procedural integrity and interscorer agreement of 15% of experimental sessions. Examiners adhered to a procedural checklist, and procedural integrity was calculated by dividing the number of correctly completed steps by the total number of steps (40), multiplied by 100 to obtain a percentage. Across examiners, procedural integrity averaged 98% (range = 83%–100%). Interscorer agreement was calculated by dividing the number of agreements by the number of agreements plus disagreements, multiplied by 100 to obtain a percentage. Interscorer agreement averaged 99% for CBM-R (range = 91%–100%), 98% for WIF-HF (range = 91%–100%), 94% for WIF-D (range = 74%–100%), and 90% for NWF-whole (range = 67%–100%), providing evidence of appropriate scoring inferences. Although the interscorer agreement for the NWF-whole probes was lower than expected, there were only a few outliers (i.e., four fell below 75%).

Data Analyses

Evidence for the extrapolation inferences was obtained by using Pearson product–moment correlations to examine the concurrent relation between the WIF-D, WIF-HF, and NWF-whole scores and students' ITBS-WA, ITBS-TR, and CBM-R performance. The magnitude of correlation coefficients was compared by use of Cohen's (1988) general guidelines, wherein point estimates $\leq .29$ are considered *small*; .30 to .49, *moderate*; .50 to .69, *large*; and coefficients $\geq .70$, *very large*. Then, the extent to which scores from WIF-D, WIF-HF, or NWF-whole were better indicators of early reading skills (ITBS-WA) and global reading achievement (ITBS-TR) was

evaluated using guidelines for comparing correlation coefficients that were delineated by Steiger (1980). That is, each correlation coefficient was transformed into a *z* score, and statistical significance between pairs of coefficients was then evaluated using equations detailed by Steiger (1980) that accounted for the fact that correlations are dependent (i.e., from the same sample) and have one variable in common (i.e., ITBS-WA or ITBS-TR).

To address the second purpose of this study, which was to evaluate the classification accuracy of each screening measure (WIF-D, WIF-HF, NWF-whole, CBM-R) and to determine whether adding a subskill mastery measure (WIF-D, WIF-HF, or NWF-whole) to CBM-R would improve classification accuracy, students in each grade were classified as at risk or not at risk, based on their ITBS-TR scores. For these analyses, students with scores at or below the 25th percentile were classified as at risk and those scoring above the 25th percentile were classified as not at risk. Therefore, risk was used as a dichotomous variable. The 25th percentile was selected because it corresponds with below-average performance on the ITBS. Next, several regression analyses were conducted with the screening measures predicting students' risk status. First, each predictor was entered separately in series of logistic regressions. Then, in a series of sequential logistic regressions, CBM-R and each subskill mastery measure were entered together to determine the classification accuracy gained by adding WIF-D, WIF-HF, or NWF-whole to CBM-R. For each logistic regression, the associated predicted probabilities were saved so that receiver operating characteristic curves could be conducted to further evaluate the classification accuracy of the screeners to predict risk status.

Several statistics were used to evaluate the classification accuracy of the universal screening measures (see Christ & Nelson, 2014, for a review). In the present study, *sensitivity* is the percentage of students determined to be at risk on the ITBS-TR (i.e., scored at or below the 25th percentile) who were accurately classified by the screener as being at risk. *Positive predictive value* (PPV) refers to the percentage of students accurately predicted to be at risk by the screener and can be viewed as how much the screener overidentifies students as at risk. *Specificity* is the percentage of students determined to be not at risk on the ITBS-TR who were correctly classified by the screener as not at risk. *Negative predictive value* (NPV) is the percentage of students accurately predicted as not at risk by the screener. Researchers have suggested that screening measures should be able to identify at least 90% of students at risk (Jenkins, Hudson, & Johnson, 2007). Researchers also have suggested that a good screener should have at least 80% specificity (Compton et al., 2010). As such, sensitivity values were set as close to 90% as possible and then the specificity, PPV, and NPV of each measure or combination of measures were obtained and compared. Finally, the area under the curve (AUC) is a measure of the overall classification accuracy of the predictors, as .50 indicates a screener (or set of screeners) has a classification accuracy that is no greater than chance and 1.0 represents perfect classification accuracy. It is generally accepted that AUC values of .90–1.0 are excellent and .85–.89 are good (Christ & Nelson, 2014); however, screeners with AUC values $<$.85 are not recommended for making screening decisions (Center on Response to Intervention, 2015).

RESULTS

Prior to analyses being conducted, it was determined that all variables were normally distributed. Descriptive statistics and correlations for all study variables are presented in Table 1. Chi-square analyses conducted to investigate potential differences in student performance on each measure as a function of school attended revealed no significant differences in first grade; however, second-grade students in School A had a significantly higher performance on ITBS-TR than second-grade students in School B ($p < .05$). No other significant differences in second-grade measures were observed. Additionally, results of the Fisher's exact test indicated no statistically significant differences across schools in the

Table 1. Descriptive Statistics and Intercorrelations Among Study Variables

Variable	*M*	*SD*	Range	*1*	*2*	*3*	*4*	*5*	*6*
First grade[a]									
1. Average CBM-R	63.39	40.47	2.5–156.5	—					
2. WIF-HF	38.44	26.49	2–97	.94*	—				
3. WIF-D	23.95	19.25	0–72	.90*	.94*	—			
4. NWF-whole	17.05	15.61	0–70	.85*	.86*	.93*	—		
5. ITBS-WA	153.84	18.43	124–202	.71*	.69*	.66*	.63*	—	
6. ITBS-TR	154.14	17.74	121–195	.89*	.83*	.81*	.77*	.77*	—
Second grade[b]									
1. Average CBM-R	100.31	39.63	7.5–202	—					
2. WIF-HF	52.68	23.77	5–110	.85*	—				
3. WIF-D	35.03	22.69	4–100	.86*	.91*	—			
4. NWF-whole	20.21	16.02	0–76	.79*	.82*	.91*	—		
5. ITBS-WA	167.94	23.05	121–233	.64*	.58*	.58*	.57*	—	
6. ITBS-TR	170.30	19.00	131–215	.81*	.69*	.71*	.64*	.72*	—

Note. CBM-R = curriculum-based measurement in reading; ITBS-TR = Iowa Test of Basic Skills–Total Reading composite; ITBS-WA = Iowa Test of Basic Skills–Word Analysis subtest; NWF-whole = whole-word nonsense word fluency; WIF-D = decodable word identification fluency; WIF-HF = high-frequency word identification fluency.
[a]$n = 135$.
[b]$n = 122$.
*$p < .001$.

percentages of students who were classified as at risk or not at risk in first and second grade.

Evidence for Extrapolation Inferences

Results indicated that WIF-D, WIF-HF, and NWF-whole scores had statistically significant ($p < .001$) associations with ITBS-WA, ITBS-TR, and CBM-R performance, with coefficients being slightly larger in magnitude for first-grade students than for second-grade students. With ITBS-WA, coefficients were large and ranged from .63 to .69 in first grade and from .56 to .58 in second grade. Associations between the word list scores and ITBS-TR performance were large to very large in magnitude, ranging from .77 to .83 in first grade and from .64 to .71 in second grade. A similar pattern was evident in the correlations between the word list scores and CBM-R performance in first grade ($r = .85–.83$) and second grade ($r = .79–.85$).

Although there were no significant differences in the associations between the word list and ITBS-WA scores in first and second grade ($p > .05$), results of the statistical tests indicated a few significant differences in coefficients between the word list and ITBS-TR scores. In first grade, WIF-HF scores had a significantly greater association with ITBS-TR performance than did NWF scores ($p = .019$) and WIF-D scores had a significantly larger association with ITBS-TR performance than did NWF-whole scores ($p = .037$). However, there was not a significant difference between WIF-D and WIF-HF scores in their relation to ITBS-TR performance ($p > .05$). For second-grade students, the association between WIF-D and ITBS-TR scores was significantly greater than the association between NWF-whole and ITBS-TR performance ($p = .012$). No significant differences were observed between WIF-HF and NWF-whole or WIF-HF and WIF-D in their relation to ITBS-TR performance ($p > .05$).

Classification Accuracy of Universal Screeners

In first grade, 18% of students ($n = 24$) scored at or below the 25th percentile on

Table 2. Classification Accuracy of Screening Measures to Predict ITBS-TR Risk Status

				Sensitivity ≈ 90%		
Screening Measure (s)	AUC	*SE*	95% CI	Specificity (%)	PPV (%)	NPV (%)
First grade[a]						
CBM-R	.973	.013	[.948, .997]	94	76	98
WIF-HF	.941	.022	[.898, .983]	88	63	98
WIF-D	.940	.024	[.893, .987]	85	58	98
NWF-whole	.885	.034	[.818, .952]	72	40	98
CBM-R + WIF-HF	.974	.012	[.950, .997]	94	71	98
CBM-R + WIF-D	.976	.014	[.948, 1.000]	96	85	98
CBM-R + NWF-whole	.972	.014	[.944, 1.000]	96	81	98
Second grade[b]						
CBM-R	.957	.027	[.905, 1.000]	87	59	98
WIF-HF	.927	.036	[.857, .997]	73	41	98
WIF-D	.968	.017	[.934, 1.000]	91	65	98
NWF-whole	.946	.023	[.901, .991]	86	55	98
CBM-R + WIF-HF	.956	.027	[.902, 1.000]	88	61	98
CBM-R + WIF-D	.965	.027	[.912, 1.000]	97	85	98
CBM-R + NWF-whole	.965	.028	[.910, 1.000]	99	94	98

Note. AUC = area under the curve; CBM-R = curriculum-based measurement in reading; ITBS-TR = Iowa Test of Basic Skills–Total Reading composite; NPV = negative predictive value; NWF-whole = whole-word nonsense word fluency; PPV = positive predictive value; WIF-D = decodable word identification fluency; WIF-HF = high-frequency word identification fluency.
[a]The at-risk base rate is 18% (n = 24).
[b]The at-risk base rate is 17% (n = 21).

ITBS-TR and were subsequently classified as at risk. As indicated in Table 2, all the screeners' individual classification accuracy was acceptable; however, CBM-R had the greatest AUC (.973), as compared with WIF-HF (.941), WIF-D (.940), and NWF (.885). The AUC for CBM-R + WIF-D (.976) was only slightly greater than that for CBM-R alone, CBM-R + WIF-HF (AUC = .974), and CBM-R + NWF-whole (AUC = .972). With sensitivity values set near 90%, NPVs were all 98% and CBM-R had the highest specificity and PPV (94% and 76%, respectively), followed by WIF-HF (88% and 63%, respectively), WIF-D (85% and 58%, respectively), and NWF (72% and 40%, respectively). When compared with CBM-R alone, the combination of CBM-R + WIF-D increased sensitivity by 2% and PPV by 9% and CBM-R + NWF-whole resulted in a 2% increase in specificity and a 5% increase in PPV. However, adding WIF-HF to CBM-R made no difference in specificity and reduced PPV by 5%.

In second grade, 17% of students (n = 21) were classified as at risk. When the overall classification accuracy of the measures in predicting second-grade students' ITBS-TR risk status was compared (see Table 2), each screener was adequate, as WIF-D had the greatest AUC (.968) as compared with CBM-R (.957), NWF-whole (.946), and WIF-HF (.927). The AUC for CBM-R + WIF-D and CBM-R + NWF-whole was the same (.965), which was greater than the AUC for CBM-R alone and CBM-R + WIF-HF (AUC = .956). With sensitivity values set near 90%, WIF-D had the greatest specificity and PPV (91% and 65%, respectively), followed by CBM-R (87% and

59%, respectively), NWF-whole (86% and 55%, respectively), and WIF-HF (73% and 41%, respectively). Furthermore, adding NWF-whole to CBM-R resulted in the greatest increase in specificity (12%) and PPV (35%) compared with CBM-R alone, and CBM-R + WIF-D yielded a 10% increase in specificity and 26% increase in PPV. Conversely, CBM-R + WIF-HF resulted in a 1% increase in specificity and a 2% increase in PPV.

DISCUSSION

Schools often use CBM-R and NWF probes for universal screening in first grade and CBM-R exclusively in second grade, even though there might be benefits to administering WIF probes in first grade (Clemens et al., 2011; Fuchs et al., 2004) and second grade. Recent research in fact suggests that WIF scores explain variance in student achievement beyond NWF scores and WIF was the single most accurate screening measure in first grade (Clemens et al., 2011). Such findings may be due to WIF probes assessing skills not measured by NWF probes, including students' recognition of high-frequency words, their skills in decoding words that are more complex than CVC words, and their ability to decode words as units. In an attempt to address these issues, we used subskill mastery probes developed to measure students' advanced decoding skills (NWF-whole, WIF-D) and students' reading of high-frequency words (WIF-HF) as well as CBM-R. Kane's (2013a, 2013b) argument-based approach to validation was used to evaluate the interpretations (i.e., extrapolation inferences) and use (i.e., decisions regarding at-risk status) of WIF-D, WIF-HF, NWF-whole, and CBM-R to identify at-risk readers in first and second grade. First, we evaluated the extent to which WIF-D, WIF-HF, and NWF-whole scores were indicators of early reading skills, as measured by ITBS-WA, and global reading achievement, as measured by ITBS-TR (i.e., extrapolation inferences). Next, we evaluated the decisions made based on scores from universal screening measures by examining the classification accuracy of each screener and determined whether administering a WIF-D, WIF-HF, or NWF-whole probe with CBM-R would yield improvements in identifying students at risk for reading difficulties.

Evidence for Extrapolation Inferences

Our findings provide evidence for the extrapolation inferences that WIF-D, WIF-HF, and NWF-whole scores are good indicators of the larger domains of reading achievement and early reading skills. That is, as in previous research with first-grade students (Clemens et al., 2011; Cummings et al., 2011; Fien et al., 2010), strong associations between WIF-D, WIF-HF, NWF-whole, and norm-referenced measures of early reading skills and global reading achievement were observed. We also extended those findings to second-grade students, with the relations among variables being similar in magnitude to those observed in first grade. In first grade, the WIF-D and WIF-HF scores demonstrated a statistically larger association with ITBS-TR performance than the relation between NWF-whole and ITBS-TR, suggesting that the WIF measures were better indicators of global reading achievement than NWF-whole. However, in second grade, WIF-D had a significantly greater association with ITBS-TR than did NWF-whole, whereas WIF-HF performance was similar to NWF-whole and WIF-D in their relation to ITBS-TR. Moreover, a slightly larger relation between each screener and ITBS-WA (e.g., decoding, phonological awareness) and ITBS-TR was observed in first grade as compared with second grade. Results also extend prior research in that the word lists administered differed from those used in previous studies, which used WIF probes consisting of words that were not controlled for decodability. We administered a similar WIF probe (WIF-HF) but also administered a structured WIF probe that consisted of only decodable words (WIF-D) to investigate potential differences in their association with reading achievement.

Classification Accuracy of Universal Screeners

Findings from this study add to an existing body of research supporting the use of CBM-R as a universal screening assessment (Kilgus et al., 2014). In first grade, scores from CBM-R demonstrated the greatest overall classification accuracy as compared with each subskill mastery measure. Moreover, when sensitivity was examined at 90% and a specificity guideline of 80% was used, either CBM-R, WIF-HF, or WIF-D was appropriate; however, CBM-R identified the greatest number of first-grade students not at risk and overidentified the fewest number of students (i.e., false positives). Notably, NWF-whole, which is widely administered in first grade for universal screening, had the lowest classification accuracy when sensitivity was set at 90%. Thus, although the NWF-whole probes developed in the current study required unitization and measured a range of decoding skills, results were consistent with prior research suggesting that NWF does not accurately discriminate among at-risk readers in first grade (Clemens et al., 2011; Johnson, Jenkins, Petscher, & Catts, 2009; Vanderwood et al., 2008). On the basis of these results, it would appear that the subskill mastery measures (and particularly NWF-whole) have little utility when administered alone as universal screeners, as they are best at identifying students who are not at risk as opposed to accurately identifying those who are at risk.

Results of this study suggest that during the winter universal screening period, CBM-R is the single most accurate screening measure for first-grade students at risk for reading difficulties. This finding is consistent with previous research suggesting that when CBM-R is administered in the fall of first grade, it classifies at-risk readers better than NWF (Johnson et al., 2009). Furthermore, improvements in the classification accuracy of CBM-R by adding a subskill mastery measure varied based on the measure. That is, although improvements were relatively small, adding WIF-D to CBM-R resulted in the greatest increase in specificity and PPV (holding sensitivity at 90%) over CBM-R alone. CBM-R + NWF-whole produced even smaller improvements in classification accuracy, and administering CBM-R + WIF-HF did not offer additional accuracy in classifying students at risk for reading difficulty.

For second-grade students, findings from the present study support the use of scores from CBM-R and subskill mastery measures for classifying at-risk readers. WIF-D had the highest overall classification accuracy, and with sensitivity at 90%, WIF-D was most accurate at classifying students who were not at risk and overidentified fewer students than did CBM-R, WIF-HF, and NWF-whole. This finding is particularly interesting, given that WIF-D is not typically administered in second grade. However, when the classification accuracy of adding a subskill mastery measure to CBM-R was examined with sensitivity set at 90%, a slightly different pattern of findings was evident, when compared to the results of the statistical optimization. That is, although CBM-R + WIF-D produced a large increase in specificity and PPV, adding NWF-whole to CBM-R yielded the greatest increase in classification accuracy over CBM-R alone. It may be that NWF-whole better captured the range of students' decoding skills and, therefore, was an appropriate complement to CBM-R.

Limitations and Future Research Directions

Findings from this study should be interpreted with several limitations considered. First, NWF-whole administration procedures differed from those used in previous research, as well as from typical assessment practices, in that students were asked to say the pseudowords as units without the option to provide the individual sounds in each word. We chose to use such procedures in an attempt to ensure that the same skill (i.e., blending of sounds) was being measured across participants, as previous research suggests giving students the option to provide the individual sounds or the entire word results in variability in the con-

struct measured (Ritchey, 2008). It is possible that greater variability in lower achieving students' NWF-whole scores would have been observed if students were able to choose their decoding strategy. A second limitation regarding our methodology is that the CBM-R score used in this study was averaged across two probes (one at grade level, one below grade level) instead of taking the median score from three grade-level probes. Third, although the word lists and preprimer CBM-R probe used in this study were developed based on empirical evidence, previous research has not demonstrated the validity of these measures. Given these limitations, future research should continue evaluating the validity of scores yielded from measures used in this study for universal screening.

There are other limitations with our sample that may limit the generalizability of these findings to other populations. First, this study included students from a small sample of schools (i.e., two), and all measures were administered during the winter universal screening period. Thus, research investigating the validity of measures used in this study during other screening periods and in a larger sample of schools is warranted. Furthermore, previous research (e.g., Hosp et al., 2011) indicated that there may be potential bias in the decisions made with CBM scores based on, among other factors, the socioeconomic status (SES) or the race and ethnicity of students. In our sample, schools differed based on the percentage of students who received free and reduced-price meals, which is often a proxy for SES. However, given that we did not have individual SES data, we were not able to make comparisons based on students who received free and reduced-price meals versus those who did not, nor were we able to control for SES in our analyses. Furthermore, the racial and ethnic composition of our sample, although reflective of the area in which participants were recruited, lacked diversity. Therefore, future research should investigate whether findings differ based on students' SES or racial and ethnic background.

Implications for Practice

Results from the present study have important implications for the practice of universal screening in first and second grade to identify students at risk for reading disabilities. First, the findings support the use of CBM-R scores for universal screening in first grade to identify students who are underachieving in reading. Furthermore, the subskill mastery measures failed to accurately classify first-grade students who were at risk, bringing into question the necessity of administering WIF probes or NWF probes that require unitization if CBM-R screening data are available. Notably, despite publishers' recommendations that NWF should be administered during first grade, findings indicate that NWF-whole should not replace CBM-R nor should NWF-whole be administered with CBM-R to identify at-risk students, at least during the winter screening period. In second grade, the findings were less clear but suggest that administering either CBM-R or WIF-D for universal screening may be appropriate. Furthermore, if a school is interested in adding a subskill mastery measure to CBM-R for universal screening in either first or second grade, findings suggest that adding WIF-D in first grade or NWF-whole in second grade may provide the most accurate identification of students who are underachieving in reading. However, in first grade, differences between the classification accuracy of CBM-R alone and word list measures added to CBM-R were minimal (i.e., one to two additional students classified as at risk). Similarly, administering NWF-whole with CBM-R in second grade yielded approximately seven more students identified as being at risk. Therefore, educators must decide whether it is worth the time and resources to increase their screening efforts in order to have small improvements in classification accuracy.

CONCLUSIONS

The purpose of this study was to use Kane's (2013a, 2013b) argument-based approach to validation as a framework to evaluate the interpretations and use of universal

screeners in first and second grade. Specifically, we demonstrated that scores from the WIF-D, WIF-HF, and NWF-whole measures in this study adequately indicated performance in the larger domains of global reading achievement and early reading skill (extrapolation inferences). This study also evaluated how universal screening data are used for making decisions about students' risk status, focusing on whether administering word lists to emerging readers during universal screenings could either improve or supplant existing universal screening practices. The results of this study confirmed, once again, that CBM-R is a valid, strong estimate of students' global reading achievement and that CBM-R can classify at-risk readers in first and second grade accurately. Although findings indicated that including WIF-D (first grade) or NWF-whole (second grade) as supplements to CBM-R may provide small increases in the number of students identified as at risk, spending the additional time and resources required to screen all students may not be practical. It is also important to note that NWF and WIF probes are subskill mastery measures, which—by design—are not intended to be indicators of global reading achievement. Furthermore, if the purpose of universal screening within a Multi-Tiered System of Support framework is to identify students who may be at risk for learning disability in reading, using a general outcome measure (such as CBM-R) seems most appropriate. By using CBM-R to screen for at-risk readers, educators can quickly identify that a problem exists before following up with additional assessment to determine the underlying skill deficit causing reading difficulties.

REFERENCES

Ardoin, S. P., Witt, J. C., Suldo, S. M., Connell, J. E., Koenig, J. L., Resetar, J. L., . . . Williams, K. L. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psychology Review, 33*, 218–233.

Burke, M. D., & Hagan-Burke, S. (2007). Concurrent criterion-related validity of early literacy indicators for middle of first grade. *Assessment for Effective Intervention, 32*, 66–77. doi:10.1177/15345084070320020401

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities, 42*, 163–176. doi:10.1177/0022219408326219

Center on Response to Intervention. (2015). *Technical standard 1: Classification accuracy.* Retrieved from http://www.rti4success.org/resources/tools-charts/screening-tools-chart/screening-tools-chart-rating-system

Christ, T. J., Arañas, Y. A., Kember, J. M., Kiss, A. J., McCarthy-Trentman, A., Monaghen, B. D., . . . White, M. J. (2014). *Formative assessment system for teachers technical manual: EarlyReading, CBMReading, aReading, aMath, and earlyMath.* Minneapolis, MN: Formative Assessment System for Teachers.

Christ, T. J., & Nelson, P. M. (2014). Developing and evaluating screening systems: Practical and psychometric considerations. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 79–110). Washington, DC: American Psychological Association.

Clemens, N. H., Hilt-Panahon, A., Shapiro, E. S., & Yoon, M. (2012). Tracing student responsiveness to intervention with early literacy skills indicators: Do they reflect growth toward text reading outcomes? *Reading Psychology, 33*, 47–77. doi:10.1080/02702711.2011.630608

Clemens, N. H., Shapiro, E. S., & Thoemmes, F. (2011). Improving the efficacy of first grade reading screening: An investigation of word identification fluency with other early literacy indicators. *School Psychology Quarterly, 26*, 231–244. doi:10.1037/a0025173

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

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. doi:10.1037/a0018448

Cummings, K. D., Dewey, E. N., Latimer, R. J., & Good, R. H., III. (2011). Pathways to word reading and decoding: The roles of automaticity and accuracy. *School Psychology Review, 40*, 284–295.

Deno, S. L., Reschly, A. L., Lembke, E. S., Magnusson, D., Callender, S. A., Windram, H., & Stachel, N. (2009). Developing a school-wide progress-monitoring system. *Psychology in the Schools, 46*, 44–55. doi:10.1002/pits.20353

Fien, H., Baker, S. K., Smolkowski, K., Mercier Smith, J. L., Kame'enui, E. J., & Beck, C. T. (2008). Using nonsense word fluency to predict reading proficiency in kindergarten through second grade for English learners and native English speakers. *School Psychology Review, 37*, 391–408.

Fien, H., Park, Y., Baker, S. K., Smith, J. L. M., Stoolmiller, M., & Kame'enui, E. J. (2010). An examination of the relation of nonsense word fluency initial status and gains to reading outcomes for beginning readers. *School Psychology Review, 39*, 631–653.

Fry, E. (1980). The new instant word list. *Reading Teacher, 34*, 284–289.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57*, 488–501.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children, 71*, 7–21. doi: 10.1177/001440290407100101

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*, 239–256. doi:10.1207/s1532799xssr0503_3

Good, R. H., III, Baker, S. K., & Peyton, J. A. (2009). Making sense of nonsense word fluency: Determining adequate progress in early first-grade reading. *Reading & Writing Quarterly, 25*, 33–56. doi: 10.1080/10573560802491224

Good, R. H., & Kaminski, R. A. (Eds.). (2007). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement. Retrieved from http://dibels.uoregon.edu

Harn, B. A., Stoolmiller, M., & Chard, D. J. (2008). Measuring the dimensions of alphabetic principle on the reading development of first graders: The role of automaticity and unitization. *Journal of Learning Disabilities, 41*, 143–157. doi:10.1177/0022219407313585

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2001). *Iowa test of basic skills, form A*. Rolling Meadows, IL: Riverside.

Hosp, J. L., Hosp, M. A., & Dole, J. K. (2011). Potential bias in predictive validity of universal screening measures across disaggregation subgroups. *School Psychology Review, 40*, 108–131.

Jamgochian, E. M., Park, B. J., Nese, J. F. T., Lai, C. F., Sáez, L., Anderson, D., . . . , Tindal, G. (2010). *Technical adequacy of the easyCBM grade 2 reading measures* (Technical Report No. 1004). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

January, S.-A. A., & Ardoin, S. P. (2015). Technical adequacy and acceptability of curriculum-based measurement and the Measures of Academic Progress. *Assessment for Effective Intervention, 41*, 3–15. doi: 10.1177/1534508415579095

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a Response to Intervention framework. *School Psychology Review, 36*, 582–600.

Johns, J. L. (1971). The Dolch basic word list—Then and now. *Journal of Reading Behavior, 3*, 35–40.

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice, 24*, 174–185. doi:10.1111/j.1540-5826.2009.00291.x

Kane, M. T. (2013a). The argument-based approach to validation. *School Psychology Review, 42*, 448–457.

Kane, M. T. (2013b). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. doi:10.1111/jedm.12000

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of School Psychology, 52*, 377–405. doi:10.1016/j.jsp.2014.06.002

Lai, C. F., Nese, J. F. T., Jamgochian, E. M., Kamata, A., Anderson, D., Park, B. J., . . . Tindal, G. (2010). *Technical adequacy of the easyCBM primary level reading measures (Grades K–1), 2009–2010 version* (Technical Report No. 1003). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Menon, S., & Hiebert, E. H. (1999). *Literature anthologies: The task for first-grade readers* (Ciera Report No. 1–009). Ann Arbor, MI: Center for Improvement of Early Reading Achievement.

Oslund, E. L., Hagan-Burke, S., Taylor, A. B., Simmons, D. C., Simmons, L., Kwok, O.-M., . . . Coyne, M. D. (2012). Predicting kindergarteners' response to early reading intervention: An examination of progress-monitoring measures. *Reading Psychology, 33*, 78–103. doi:10.1080/02702711.2012.630611

Pearson. (2012). *AIMSweb test of early literacy administration and scoring guide.* Bloomington, MN: Pearson Education. Retrieved from http://www.aimsweb.com/wp-content/uploads/TEL_Admin_Scoring-Guide_2.0.pdf

Pratt, K., Martin, M., White, M. J., Christ, T. J., Ardoin, S. P., & Eckert, T. L. (2011). *Development of FAIP-R passage sets: Level 1* (Report No. 3). Minneapolis, MN: Department of Educational Psychology, University of Minnesota.

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*, 427–469. doi:10.1016/j.jsp.2009.07.001

Ritchey, K. D. (2008). Assessing letter sound knowledge: A comparison of letter sound fluency and nonsense word fluency. *Exceptional Children, 74*, 487–506.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245–251. doi:10.1037/0033-2909.87.2.245

Vanderwood, M. L., Linklater, D., & Healy, K. (2008). Predictive accuracy of nonsense word fluency for English language learners. *School Psychology Review, 37*, 5–17.

Vellutino, F. R., Scanlon, D. M., Small, S., & Fanuele, D. P. (2006). Response to intervention as a vehicle for distinguishing between children with and without reading disabilities: Evidence for the role of kindergarten and first-grade interventions. *Journal of Learning Disabilities, 39*, 157–169. doi:10.1177/00222194060390020401

White, M. J., Martin, M., Monaghen, B., Ardoin, S. P., Christ, T. J., & Eckert, T. L. (2011). *Early literacy measures development* (Report No. 6). Minneapolis, MN: Department of Educational Psychology, University of Minnesota.

Zumeta, R. O., Compton, D. L., & Fuchs, L. S. (2012). Using word identification fluency to monitor first-grade reading development. *Exceptional Children, 78*, 201–220.

Date Received: March 27, 2015
Date Accepted: September 1, 2015
Editor: Amy Reschly ■

Stacy-Ann A. January, PhD, is an assistant professor in the Department of Psychology at the University of South Carolina. She earned her doctoral degree from the School Psychology program at the University of Georgia prior to completing a 2-year postdoctoral research fellowship at the University of Nebraska–Lincoln. Dr. January's research interests include (a) investigating the technical characteristics and decision-making utility of universal screening and progress-monitoring assessments and (b) evaluating interventions that target academic or behavioral skills.

Scott P. Ardoin, PhD, is Co-Director of the Center for Autism and Behavioral Education Research and a professor in the Department of Educational Psychology at the University of Georgia (UGA). His recent work employs eye-tracking technology to (a) measure changes in the reading behavior of students as a function of interventions and (b) examine the extent to which test and question formats alter students' reading behavior and test-taking strategies. His research is published in numerous refereed journals; he is a recipient of the APA Division 16 Lightner Witmer Award and a recipient of the UGA Creative Research Medal.

Theodore J. Christ, PhD, is a professor of school psychology in the Department of Educational Psychology and is the Director for the Center of Applied Research and Educational Improvement (CAREI) and Co-Director of the Research Institute for Problem Solving, which all reside at the University of Minnesota. It was his work in those roles that established him as the Founder and Chief Scientific Officer of FastBridge Learning (fastbridge.org). Dr. Christ is engaged to innovate techniques and technology to serve professional educators who seek to improve educational outcomes. He is interested and engaged with the content, methodology, analytics, technology, software, and people who pursue the same.

Tanya L. Eckert, PhD, is an associate professor of psychology at Syracuse University. Her research interests include examining procedures for assessing academic and behavioral problems, developing classroom-based interventions, and measuring the acceptability of assessment and intervention procedures. Dr. Eckert is Senior Associate Editor of *School Psychology Review*.

Mary Jane White, PhD, is a research associate at the University of Minnesota and serves as project coordinator for a variety of Dr. Theodore J. Christ's federally funded assessment grants. She also supports development of assessments for FastBridge Learning. She earned her degree in educational psychology with an emphasis in reading comprehension.