Using Word Identification Fluency to Monitor First-Grade Reading Development

REBECCA O. ZUMETA, **DONALD L. COMPTON**, and **LYNN S. FUCHS**
REBECCA O. ZUMETA (Washington, DC CEC), Doctoral Graduate; DONALD L. COMPTON (Tennessee CEC), Associate Professor; and LYNN S. FUCHS (Tennessee CEC), Professor and Nicholas Hobbs Chair, Department of Special Education, Vanderbilt University, Nashville, Tennessee

Abstract

This study assessed the effects of sampling breadth on technical features of word identification fluency (WIF), a tool for screening and monitoring the reading development of first graders. From a potential pool of 704 first-grade students, the authors measured both a representative sample ($n = 284$) and 2 other subgroups: those with low reading achievement ($n = 202$) and those with high/average achievement ($n = 213$). Data were collected weekly on broadly and narrowly sampled WIF lists for 15 weeks and on criterion measures in the fall and spring. Broad lists were developed by sampling words from 500 high-frequency words, whereas narrow lists were created by sampling from the 133 words from Dolch preprimer, primer, and first-grade word lists. Overall, predictive validity for performance level, predictive validity for growth, and commonality analysis showed narrow sampling was better for screening the representative group and the high/average subgroup. Broad sampling was superior for screening the low-achieving subgroup and for progress monitoring across groups.

Reading skill is an integral component of academic success. When students fail to develop adequate literacy skills at a young age, they are more likely to struggle throughout their academic careers (Juel, 1988), and it is time-intensive and expensive to remediate their difficulties. Students who are poor readers in the early grades rarely catch up to their peers, even with extra help (Juel, 1988), putting them at risk for learning disabilities (LD) and a variety of negative educational outcomes.

Early intervention to prevent development of reading difficulties can be an effective way to ameliorate this problem (Torgesen, Wagner, Rashotte, Rose, et al., 1999), and screening and progress monitoring can identify students who require such intervention (Compton, Fuchs, Fuchs, & Bryant, 2006). Curriculum-based measurement (CBM) is an assessment tool that provides teachers with reliable and valid information about academic progress (Deno, 1985; Deno, 2003; L. S. Fuchs, Fuchs, & Compton, 2004). When administered periodically, scores can be used to screen students for academic risk; when administered at least monthly, scores can be used to model students' academic development. Originally developed during the 1970s and 1980s to evaluate the effectiveness of special education programs (Brownell, Sindelar, Kiely, & Danielson, 2010; Deno, 1985), CBM is currently used in response to

intervention (RTI) models to monitor students' progress in both general and special education (National Center on Response to Intervention, NCRTI, 2010).

SCREENING, PROGRESS MONITORING, AND THE IDENTIFICATION OF DISABILITY RISK

RTI is an alternative to the traditional discrepancy model used to identify LD. RTI emphasizes early identification and intervention through universal screening and progress monitoring over time to verify academic risk and response to high-quality instruction (Vaughn & Fuchs, 2003). LD is classified on the basis of a dual discrepancy in overall achievement and progress, compared to peers. That is, LD is evident when a student with low achievement also demonstrates inadequate progress over time, despite generally effective instruction. This evidence is gathered through universal screening and ongoing progress-monitoring assessments (NCRTI, 2010). Thus, RTI relies on the assumption that measures of academic achievement and progress are valid, reliable markers of students' risk status and indicators of development. The purpose of our study was to investigate psychometric properties of word identification fluency (WIF), a popular assessment used to screen and monitor first graders' reading development.

Monitoring First-Grade Reading Development With CBM

Given the demonstrated importance of early identification and intervention to prevent reading problems (Ardoin & Christ, 2008; Baker et al., 2008; Stage & Jacobsen, 2001), it is imperative that screening and progress monitoring assessments maximize appropriate and timely identification of risk and disability. As noted, CBM is a useful method for screening students at risk for reading problems (Compton et al., 2006; Compton et al., 2010) and for assessing whether interventions facilitate adequate academic progress over time (Deno, 1985; Deno, 2003; L. S. Fuchs, Deno, & Mirkin, 1984). CBMs are technically valid, standardized formative assessments (Deno, 2003) that require students to complete brief, frequent, parallel form tests that provide indicators of academic achievement. Teachers collect and graph performance data to analyze overall skill and growth to determine if progress toward end-of-year benchmarks is sufficient. CBM may be used to monitor progress for students in an entire school or classroom, for tracking an individual's progress toward end-of-year benchmarks or individualized education program goals, or for screening students at one time point to determine risk for academic failure (Deno, 2003).

The variety of uses for CBM is particularly important in light of the essential role high-quality screening and progress monitoring plays in the effective implementation of RTI (NCRTI, 2010). CBM measures are fluency based, whereas many norm-referenced measures of reading emphasize accuracy over fluency, despite evidence that fluency explains significant unique variance in reading ability (L. S. Fuchs, Fuchs, Hosp, & Jenkins, 2001; Meisinger, Bloom, & Hynd, 2010). *Fluency,* defined here as the accurate and fast performance of a task, may be an important component of assessment because it reveals automaticity. When students are fluent readers, they can devote greater cognitive resources to higher order tasks such as comprehension (LaBerge & Samuels, 1974).

Although a variety of standardized and norm-referenced tools may be used to screen students, given the scope of our study we limited our discussion to CBM and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002), all fluency-based measures used to screen and monitor reading progress at first grade (Manzo, 2005). The most popular and researched form of reading CBM, Passage Reading Fluency (PRF; also known as *oral reading fluency*), measures a student's speed and accuracy in reading words in connected text. Empirical evaluations of PRF's technical characteristics have

revealed that when PRF is used during the first half of first grade, floor effects occur (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009; L. S. Fuchs et al., 2004; Good, Simmons & Kame'enui, 2001). That is, many students have low scores, failing to distinguish among performers with varying levels of expertise and reducing predictive utility due to truncation of range (Crocker & Algina, 1986). Further, PRF may not be sensitive to growth when initial scores are very low because it fails to register improvement. This undermines PRF's utility as a progress-monitoring tool in the first half or more of first grade (L. S. Fuchs et al., 2004).

Good and colleagues (2001) developed DIBELS for assessing reading development in Grades K through 6. Included in these measures is Nonsense Word Fluency subtest (NWF) for use in the first half of first grade. NWF assesses skill at decoding consonant-vowel-consonant non-words. Although NWF is widely used (Manzo, 2005), we focused our study on WIF, another screening and progress monitoring tool intended for use with first graders. We chose WIF for several reasons. First, prior research has shown accurate isolated word reading is a significant indicator of first-grade reading competence (Foorman, Francis, Beeler, Winikates, & Fletcher, 1997). Second, fluency with reading isolated words has been shown to be useful for monitoring first graders' emerging reading. Finally, evidence suggests WIF has stronger predictive validity than NWF (Clemens, 2009; L. S. Fuchs et al., 2004; Healy, 2007; Meisinger et al., 2010).

We identified four studies that addressed the importance of WIF in screening for reading difficulty and monitoring development. Deno, Mirkin, and Chiang (1982) documented strong concurrent validity for WIF among 66 students in Grades 1 through 6. The other three studies not only demonstrated the validity of WIF, they also provided evidence of its superiority over NWF. Working exclusively with 151 at-risk first graders, L. S. Fuchs et al. (2004) demonstrated strong concurrent and predictive validity for WIF level and growth rate. The strength of these validity indicators exceeded that of NWF, suggesting that WIF should be the preferred assessment tool at first grade. Healy (2007) expanded upon the work of L. S. Fuchs and colleagues (2004) by focusing on a representative sample of first graders that included English language learners (ELLs). Overall, WIF demonstrated better predictive validity than NWF for English-speaking and ELL students. Similarly, Clemens (2009) compared WIF to NWF with 154 first graders and found that WIF performed better than NWF in almost all comparisons. Moreover, the technical features of WIF were strong across the entire first-grade year, circumventing the need to switch measures during the second half of the school year, as is recommended with NWF (Good et al., 2001). Taken together, these studies suggest that compared to NWF, WIF may be the more efficient and technically valid way of assessing first-grade reading development.

Purpose and Rationale of Present Study

Although research shows that WIF functions as a technically valid method, with superior predictions of reading development compared to NWF for monitoring first graders' reading progress, it remains an imperfect predictor of reading competence (Clemens, 2009; L. S. Fuchs et al., 2004; Healy, 2007). Thus, it is important to investigate how to improve WIF so it more accurately serves to predict reading problems and monitor student progress at first grade, when prevention efforts may have the greatest promise of success (Ardoin & Christ, 2008; Baker et al., 2008). Thus, the purpose of our study was to compare two approaches to WIF assessment: the established method that samples from a narrow group of high-frequency words versus an alternate approach that samples from a broader assortment of high-frequency words. The rationale behind broader sampling is that by increasing the range and difficulty of words, students' reading skills may be more clearly differentiated, especially at the higher end of the achievement continuum. Also, in the broadly sampled lists, words are presented in clusters representing decreasing frequency, such that children

may have greater success at reading words near the beginning of the test, thereby facilitating motivation. Additionally, the broader sampling may be more sensitive to growth because it may provide students benefiting from instruction greater opportunity to demonstrate increased competence. This difference in sampling and presentation may also increase the range of level estimates and growth rates. Although prior research has demonstrated the technical adequacy of WIF and other measures of word reading (Torgesen, Wagner, and Rashotte, 1999; Woodcock, 1998), an evaluation of sampling approaches used to construct word lists has not been conducted.

To accomplish our purpose, we monitored the progress of first graders from 14 urban public schools, identifying a representative sample of 284 students with high, average, and low initial reading achievement, as well as two other groups, or subsamples: students with low initial performance ($n = 202$) and students with high or average initial performance ($n = 213$). We tested students on criterion measures in the fall and spring of first grade, and we monitored their progress using narrowly and broadly sampled WIF lists for 15 weeks. We refer to the established sampling approach as WIF-N (narrow) and the alternate sampling approach as WIF-B (broad). We investigated the predictive validity of both WIF scores to contrast their utility for screening. We also investigated predictive validity of growth across the 15 weeks to compare progress-monitoring utility. Finally, we used commonality analysis to compare overall and unique variance explained by WIF level and growth rates.

METHOD

Participants

This research was conducted as part of an ongoing study investigating key measurement issues associated with RTI models, including who should enter the RTI process and how to operationalize inadequate response to instruction/intervention (Compton et al., 2006). As part of this project, we obtained Institutional Review Board approval and tested 704 first graders with parental consent from 56 general education classrooms in 14 schools in an urban school district at the beginning of the school year. Half the schools received Title I funding. Table 1 reports additional demographic details. From this sample, we identified a representative sample ($n = 284$) of students and two other samples: students with low initial achievement ($n = 202$), and students with average/high initial achievement ($n = 213$); the *low study-entry* (LSE) and *not-low study-entry* (not-LSE) *subsamples,* respectively. We describe procedures for identifying these groups in the following section.

Measures

Initial Screening to Determine Initial Reading Status—We assessed the initial pool of students ($n = 704$) on rapid letter naming (RLN), rapid letter sound naming (RLS), and two WIF-N probes. With RLN (D. Fuchs et al., 2001), students are presented with a list of 26 letters and instructed to read as many as they can in 1 min. Scores range from 0 to 26 or higher if the student reads the entire list in less than 1 min. (If a student reads the list in less than 1 min, reading rate is prorated.) With RLS, students are presented with 26 letters and asked to read as many letter sounds as they can in 1 min (L. S. Fuchs & Fuchs, 2001). Scores are calculated in the same manner as with RLN. Test–retest reliability for this measure is .92 at first grade (L. S. Fuchs & Fuchs, 2001). With WIF-N, students read two high-frequency sight word lists. WIF-N lists (L. S. Fuchs et al., 2004) were developed as part of a prior study and comprise words from the 133 high-frequency words from the Dolch preprimer, primer, and first-grade-level lists. To create these lists, 50 words were selected at random and appear on the page in arbitrary order. Thirty WIF-N alternate form lists were developed this way. The score is the number of words read correctly in 1 min. Alternate form reliability for WIF-N was .95 to .97.

Progress Monitoring With WIF-N and WIF-B—We used two kinds of WIF in this study. WIF-B was created by sampling from a larger corpus of words than WIF-N. We constructed these lists by sampling from the 500 most frequently written words from the list generated by Zeno, Ivens, Millard & Duvvuri, (1995). From that list, we divided words into 50 groups of 10 words. The 10 words in the first group were the most frequent words; in the second group, the next most frequent words; and so on. Then, we selected one word from each group at random (with replacement) so that there were 50 words on each WIF-B list. Next, we arranged the words on the page so the most common words appeared first, progressing to the least common words. We created 20 alternate forms using this procedure. Alternate form reliability for WIF-B lists used in this study was .95 to .97. We administered two alternate forms of WIF-N and two alternate forms of WIF-B each week for 15 school weeks (i.e., 10 of the 20 WIF-B lists were used twice). Lists were administered in a fixed order, and once all lists had been administered, examiners began with the first list again. Including school holidays, 10 to 12 weeks intervened between administrations of identical WIF-B lists. Given that these were lists of isolated words and that substantial time passed between administrations, it is unlikely that memory inflated scores. (There were enough WIF-N alternate forms so that each list was used only once.)

Norm-Referenced Measures Selected for Criterion Validity—We collected data on five norm-referenced reading measures during the fall and spring to assess criterion validity. We chose these measures for their strong technical characteristics and common use. In addition, these measures represent fundamental components of first-grade reading skills (i.e., decoding, sight word, text reading) with both accuracy and fluency-based assessments. Woodcock Reading Mastery Test-Revised (WRMT-R) Word Identification (WID; Woodcock, 1998) assesses sight word reading accuracy; students read as many words as they can from lists of words. The test is not timed, and the score is the total number of words the student reads correctly. Testing is discontinued once a student makes six consecutive errors. Split-half reliability is .97 (Woodcock, 1998).

WRMT-R Word Attack (WAT; Woodcock, 1998) assesses accuracy at reading pseudo words. There are 45 items, organized in order of difficulty. The test is untimed and discontinued after six consecutive errors. The score is the number of correct words read. At first grade, split-half reliability is .90, and test-retest reliability is .95 (Woodcock, 1998).

The Comprehensive Reading Assessment Battery (CRAB) Fluency (L. S. Fuchs, Fuchs, & Hamlett, 1989) assesses passage reading fluency. The assessment comprises 400-word folktales at a 1.5 grade level. Scores are the number of words attempted and read correctly at the end of 1 and 3 min. We used the fluency score at the end of 1 min in this study. Test-retest reliability was .93 to .96 (L. S. Fuchs et al., 1989).

The Test of Oral Word Reading Efficiency (TOWRE) Sight Word (Torgesen, Wagner, & Rashotte, 1999) measures sight word fluency. Students read as many sight words as they can in 45 s. Split-half reliability is .91 (Torgesen, Wagner, & Rashotte, 1999).

With the TOWRE Decoding (Torgesen, Wagner, & Rashotte, 1999), students read as many nonsense words as they can in 45 s. The score is the number of words read correctly. Split-half reliability is .90 (Torgesen, Wagner, & Rashotte, 1999).

Selecting the Representative Sample and Subsamples

For sampling purposes, we created a single factor score using the initial screening measures (letter naming fluency, letter sound fluency, and WIF-N). Three classes were formed on the factor score using latent class modeling (Mplus 6.1) representing low, average, and high study-entry. We used an empirical approach to form initial performance classes for two

reasons: (a) there were not local norms available for classifying children into classes at the start of first grade on the screening measures, and (b) an empirical approach is less prone to human bias in selecting class membership. We classified the total pool of 704 children into high-, average-, and low-performing groups, with 485 initially sampled for the longitudinal study. We oversampled low-performing children to increase the number of struggling readers in the study. Of the 485 children included in the longitudinal study, 220 were classified as low-study-entry (LSE), 173 as average-study-entry (ASE), and 92 as high-study-entry (HSE). In this study we use the 415 children who were assessed on the five pre- and posttest criterion measures. We identified a representative sample of 284 students: 71 (25%) LSE, 142 (50%) ASE, and 71 (25%) HSE.

Then, we considered two subsamples. The first comprised all 202 LSE students available in the database of 415 students. (We included the 71 LSE students in the representative sample, as well as the remaining 131 LSE students in the original database of 415, for a subsample total of 202 LSE students.) The other subsample combined the remaining 213 students with ASE and HSE achievement from the representative sample (i.e., not-LSE). The mean age in the representative sample was 6.54 years, with a standard deviation of 0.34. The mean age was 6.55 years ($SD = 0.41$) in the LSE subsample, and 6.54 years ($SD = 0.34$) in the not-LSE subsample. Table 1 provides additional demographic information for the three groups, on which separate analyses were conducted.

Assessment Procedures

We used data collected from initial fall testing sessions (late September and early October) to designate students' high, average, or low study-entry status. Fall criterion measures were also collected in October and November. We collected progress-monitoring data over 15 school weeks from late November to mid-March, but spanning 17 calendar weeks due to school holidays. If students were absent, research assistants made up the progress-monitoring session within 1 week of the target data collection date whenever possible. All students read the same four word lists each week. Two word lists were alternate forms of the WIF-N format; two were alternate forms of the WIF-B format. We collected spring criterion measures in April and May, 2 to 6 weeks after completion of progress-monitoring data collection. Assessments were individually administered by 12 graduate-level research assistants or full-time project staff with master's degrees trained to collect data with 100% accuracy.

ANALYSIS AND RESULTS

WIF Level of Performance

Table 2 provides descriptive information on the performance of each subsample on criterion and fall WIF measures. To determine students' fall level of WIF performance, we averaged their Week 1 WIF-N scores and averaged their Week 1 WIF-B scores, respectively. The purpose of averaging scores was to reduce measurement error associated with relying on data from a single probe. We used the same procedure with Week 17 WIF-N and WIF-B data to determine students' spring WIF level performance.

WIF Rate of Improvement

We used individual growth curve modeling to identify average rates of improvement on the WIF-N progress-monitoring data and on the WIF-B progress-monitoring data. As with level scores, we then averaged each student's score on two WIF-N probes collected during the same test session to determine the weekly WIF-N progress monitoring score. We used the same procedure to derive the weekly WIF-B score. Next, we tested linear and quadratic models and found the quadratic model to be a better fit across samples/subsamples and

progress-monitoring measures. (Model comparison tests comparing linear and quadratic models: Representative sample WIF-N X^2,[4, $N = 284$] = 68.18, $p < .001$ and WIF-B X^2[4, $N = 284$] = 63.35, $p < .001$; LSE subsample WIF-N X^2[4, $N = 202$] = 89.77, $p < .001$ and WIF-B X^2[4, $N = 202$] = 184.02, $p < .001$; Not-LSE subsample WIF-N X^2[4, $N = 213$] = 43.29, $p < .001$ and WIF-B X^2[4, $N = 213$] = 43.37, $p < .001$). When a quadratic model is centered at a given point, the linear slope from the model represents the growth rate for that time period (Schatschneider, Wagner, & Crawford, 2008). This value is known as the *instantaneous growth rate* for that time point. In cases where a quadratic model fits the data better than a linear model, reporting instantaneous growth at different time points is a more accurate way to characterize change over time than a single linear slope (Raudenbush & Bryk, 2002). We determined instantaneous growth rates by taking the value of the linear slope within models centered at three time points: Weeks 1, 9, and 17. The mean instantaneous growth rates for WIF-N and WIF-B at Weeks 1, 9, and 17 derived from these models are reported for the representative sample, and the LSE and not-LSE subsamples, in Table 3. Overall, growth rate magnitude followed the expected pattern, with representative sample growth rates generally larger than LSE growth rates, and not-LSE growth rates larger than representative and LSE growth rates.

Comparisons of WIF-N and WIF-B

In this study, we investigated two approaches to sampling words and constructing WIF lists for the purposes of screening and monitoring the progress of first-grade readers. We wanted to determine whether a broad approach to sampling words and arranging those words in order of frequency would affect the validity of WIF level and growth estimates, compared to the traditional narrower approach to sampling and formatting. We compared predictive validity for WIF-N versus WIF-B for the representative sample as well as the LSE and not-LSE subsamples. Pairs of correlations between criterion and WIF measures were compared using Walker and Lev's (1953) formula.

Predictive Validity—Table 4 reports predictive validity comparisons between WIF-N and WIF-B. Fall WIF level was used to predict students' spring performance on criterion measures. Additionally, instantaneous growth rates at Weeks 1, 9, and 17 were used to predict spring performance on criterion measures for the representative sample and both subsamples.

For the representative sample, WIF-N fall level was a significantly stronger predictor than WIF-B level with respect to students' spring performance on the WRMT-R WID ($p < .01$) and TOWRE Sight Word ($p < .01$) subtests. There were no significant differences with respect to other criterion measures, however. The pattern differed for instantaneous growth rates. At Weeks 1, 9, and 17, the WIF-B instantaneous growth rate was a significantly stronger predictor of performance on spring WRMT-R WID ($p < .01$ across weeks). WIF-B instantaneous growth rates at Weeks 1,9, and 17 were also significantly more strongly correlated than WIF-N with respect to students' spring performance on WRMT-R WAT, CRAB Fluency, TOWRE Sight Word, and TOWRE Decoding assessments (see Table 4).

For the LSE subsample, fall WIF-B level was a significantly stronger predictor of spring WRMT-R WID ($p < .01$) performance than WIF-N level. There were no significant differences on other criterion measures. With respect to spring WRMT-R WID performance, however, WIF-B instantaneous growth rates at Weeks 9 and 17 were significantly stronger predictors than WIF-N ($p < .01$ across weeks). This pattern was also true for prediction of spring WRMT-R WAT, CRAB Fluency, TOWRE Sight Word, and TOWRE Decoding performance ($p < .01$ across Weeks 9 and 17).

For the not-LSE subsample, significant differences tended to favor WIF-B, with the exception of fall level, where WIF-N was a stronger predictor of spring performance on TOWRE Sight Word ($p < .01$). There were no other significant differences between level estimates, however. WIF-B instantaneous growth rates at Weeks 1 and 9 were significantly stronger predictors of spring performance on WRMT-R WID and WAT, and TOWRE Decoding assessments than WIF-N growth rates (see Table 3). Additionally, WIF-B growth rates at Weeks 1, 9, and 17 were significantly better predictors of students' spring performance on CRAB Fluency and TOWRE Sight Word assessments than corresponding WIF-N growth rates (see Table 4).

Commonality Analysis

Finally, we applied a commonality analysis to determine the overall and unique variance the two versions of WIF accounted for when modeled against the five criterion spring reading achievement measures. This allowed us to determine the relative contributions of WIF-N and WIF-B level and growth rate, respectively, across models. Results of these analyses are shown in Table 5 for the representative sample, Table 6 for the LSE sub-sample, and Table 7 for the not-LSE subsample. WIF level measures were significant predictors of nearly all outcome measures across the representative sample and both subsamples. By block-entering predictors into regression models, we determined the unique variance accounted for by WIF-N and WIF-B for each subsample and outcome measure. With respect to WIF level comparisons, both the broad and narrow sampling formats explained significant unique variance in virtually all models. WIF-N tended to contribute more unique variance than WIF-B when the representative sample was analyzed, with the exception of the model where TOWRE Decoding was the dependent variable. In this case, unique variance contributed was nearly identical. The unique variance accounted for by WIF-N was fairly large across criterion measures, ranging from 3.3% to 14.9%. There was more variability in both sub-samples, however. For the LSE subsample, WIF-B explained more unique variance than WIF-N in three of five models. For the not-LSE subsample, WIF-N predicted more unique variance than WIF-B in three of five models. In nearly all cases, unique variance accounted for by WIF-N and WIF-B in both subsamples was significant, but differences in the amount of unique variance accounted for by predictors was not as large as in the representative sample.

With respect to instantaneous growth rates, WIF-B accounted for more unique variance than WIF-N across time points, samples, and dependent measures. As Tables 5, 6, and 7 indicate, models including the Week 9 instantaneous growth rate accounted for the largest amount of variance across dependent measures, compared to models with instantaneous growth rates at Weeks 1 and 17. This pattern was particularly evident for the LSE subsample (Table 6), where WIF-B accounted for 46.1% to 65.1% of the variance across outcome measures. With respect to unique variance explained, Tables 5, 6, and 7 show that WIF-B contributed 16.3% to 48.0% unique variance across outcome measures and samples at Week 9 (versus 0.1% to 13.0% unique variance for WIF-N). Similar patterns persisted at Weeks 1 and 17, but percentage of unique variance explained by each WIF list tended to be smaller than at Week 9.

We also observed negative commonalities in several models where we evaluated instantaneous growth rates for the not-LSE subsample and representative sample. Negative commonalities suggest a suppression effect (Nimon, Nathans, & Henson, 2010), which occurs when a predictor variable is not significantly correlated with the outcome variable, but is correlated with one of the other predictors. As can be seen by comparing Table 4 to Tables 5, 6, and 7, suppression occurred when the WIF-N growth rate did not yield a significant validity coefficient when correlated with the criterion variable of interest. (WIF-N growth rates and WIF-B growth rates are significantly correlated with one another,

however.) Suppressor variables can enhance prediction of a dependent variable because they remove some of the irrelevant variance of other variables, allowing them to better predict outcome (Nimon et at., 2010). Omission of suppressor variables can also lead to underestimating the effect of a significant predictor (Nimon et al., 2010). In all cases, WIF-N growth rate was the suppressor variable (not WIF-B).

Finally, we compared the unique variance explained in models with the overall strongest level and growth predictors: WIF-N level and WIF-B growth rate at Week 9 (see Table 8). With respect to the representative sample, the overall proportion of variance explained across outcome measures ranged from 59.9% to 86.8%, and all predictors were significant. WIF-N level uniquely explained 33.0% to 49.1% of variance across outcome measures, and the WIF-B Week 9 instantaneous growth rate uniquely explained 1.7% to 4.5% of variance.

The total proportion of variance explained by models for the LSE subsample ranged from 51.4% to 81.3% across outcome measures. Again, all predictors were significant. WIF-N level uniquely explained 7.7% to 20.8% of variance across outcome measures, and the WIF-B Week 9 instantaneous growth rate uniquely explained 6.6% to 18.4% of variance. For the not-LSE sub-sample, the total proportion of variance explained ranged from 42.6% to 81.4% across outcomes. WIF-N level uniquely explained 30.5% to 57.6% of variance across dependent measures. The WIF-B Week 9 instantaneous growth rate uniquely explained 1.2% to 4.9% of variance across outcome measures. Despite level explaining a greater proportion of variance, results suggest both WIF-N level and WIF-B growth rate explain significant unique variance in outcome.

DISCUSSION

Our results have important implications for screening and progress monitoring across the achievement continuum at first grade, including students with low initial achievement who may be at risk for reading disabilities. WIF level findings can help inform screening, whereas findings related to WIF growth rates should inform progress monitoring. For the representative sample, WIF-N level outpredicted WIF-B level in two of five comparisons. For the LSE subsample, WIF-B level outperformed WIF-N level in one of five comparisons. For the not-LSE subsample, one comparison favored WIF-N level. Taken together, these findings indicate that WIF-N level showed some advantage over WIF-B level, although most comparisons revealed similar predictive ability across the two measures. The lack of difference could be due in part to the fact that WIF-B was designed to enhance detection of change over time, which level estimates do not; they are intended for screening at a single time point to identify academic risk, and, for this purpose, our analyses suggest WIF-N has some advantage for screening representative and not-LSE samples.

The picture changed dramatically, however, when we considered the predictive validity of WIF growth rates. Unlike level estimates, which are used for screening, growth rates have important implications for progress monitoring. Correlations between growth rates and criterion measures were more modest than correlations between criterion measures and WIF levels, and several were non-significant (see Table 4)., Comparisons between correlations clearly favored WIF-B. Across comparisons and subsamples, instantaneous growth rates of WIF-B were generally stronger than WIF-N at predicting students' performance on spring criterion measures. This suggests that, overall, WIF-B is the more advantageous progress-monitoring tool, a finding consistent with our hypothesis that broadly sampled word lists, presented in order of frequency, may be more sensitive to change over time than more narrowly sampled lists.

Our commonality analyses provide the basis for extending insights about the contributions of WIF-N and WIF-B for level and growth. Across outcome measures, WIF-N level explained more unique variance than WIF-B level for students in the representative sample. This is consistent with findings from our predictive validity analyses and provides further evidence to suggest that when testing a representative sample, WIF-N should be the preferred screening tool. However, differences were less distinct when subsamples were considered. With respect to the LSE subsample, WIF-B explained more unique variance than WIF-N in three of the five outcome measures considered. The reverse was true for the not-LSE subsample, where WIF-N level explained more unique variance against three of five outcome measures. These patterns suggest WIF-B is more advantageous for screening LSE students, and WIF-N should be preferred when screening not-LSE or representative samples. By contrast, for progress monitoring, WIF-B was superior across subsamples, time points, and outcome measures.

This work contributes to the literature on WIF in several important ways. First, it provides corroborating evidence to support the overall strong validity coefficients for WIF-N level, which have been reported in prior work (Clemens, 2009; L. S. Fuchs et al., 2004; Healy, 2007). It also complements findings by Baker et al. (2008), suggesting that PRF level and growth estimates were valid predictors of performance on standardized reading tests when used with 34 students in late first through third grades. Our findings also extend the work of Baker et al. by suggesting WIF may be used to effectively screen and monitor progress across the achievement continuum throughout first grade (not just during the second half of the year, as with PRF).

In addition, our results suggest that WIF growth terms are quadratic, not linear. Importantly, our use of growth modeling to evaluate WIF progress extends work by Stage and Jacobsen (2001). When modeling 74 fourth graders' PRF progress, they reported linear growth with a relatively high-achieving sample, however. At the same time, our finding of significant quadratic growth for WIF stands in contrast to prior research (Clemens, 2009; L. S. Fuchs et al., 2004; Healy, 2007), which reported significant models with linear growth terms for WIF-N (and did not include WIF-B). At the same time, our analyses are consistent with and extend previous research by Ardoin and Christ (2008), who observed inconsistent oral reading fluency (ORF) growth with 86 second graders. The quadratic term finding is important because it suggests word reading growth may not be constant. Practically speaking, this implies it may be useful for teachers to set different goals for student growth on WIF depending on when data are collected, rather than expect constant growth throughout the year. This hypothesis warrants further investigation.

Our findings further suggest that growth on broadly sampled word lists (WIF-B) may be a better predictor of student outcome than growth on narrowly sampled lists (WIF-N). This pattern was consistent across the representative sample and both subsamples. Further, our use of a representative sample and a not-LSE subsample to evaluate screening and progress monitoring extends prior work by L. S. Fuchs et al. (2004), who reported screening and progress-monitoring data on at-risk students alone. It also extends work by Clemens (2009), who reported progress-monitoring data for at-risk students, and by Healy (2007), who reported on samples of English only and ELL students (but did not describe representativeness of their achievement). We did not include ELLs in our analysis, so research is needed to verify WIF-B findings with that population.

Our findings should also be considered in light of prior work by Schatschneider et al. (2008), who found ORF growth estimates did not explain unique variance in first graders' end-of-year reading achievement, suggesting that end of first-grade performance may be more important than initial performance when predicting future achievement. Importantly,

these authors used ORF scores across the entire year rather than weekly WIF to monitor progress. Given the noted problems with using ORF to monitor progress in early first grade (i.e., floor effects, lack of sensitivity to growth), WIF may be a more appropriate tool with which to determine the relative importance of level and growth throughout first grade in the identification of response to intervention. Additionally, Schatschneider et al. collected data at four time points and reported the instantaneous growth rate from the final time point. As they noted, more frequent data collection may yield a more reliable slope parameter, as demonstrated in the present study with weekly progress monitoring throughout first grade using WIF (also see Compton et al., 2010; L. S. Fuchs et al., 2004). Although our findings support the importance of using WIF to model first-grade reading development, the work of Schatschneider et al. suggests it may also be worthwhile to determine significance of WIF growth predictions beyond first grade.

Given the importance of validated instructional methods and targeted efforts toward prevention, it is imperative that measurement tools accurately identify students who require extra help in a timely manner. The data presented here suggest that WIF-B measures should be used to monitor student progress on a regular, perhaps weekly basis. By contrast, when screening representative student samples, WIF-N may be preferred.

Our findings also suggest that WIF growth explains significant variance in first graders' end-of-year reading achievement across subsamples and outcome measures. This pattern was particularly evident in the LSE subsample, which may indicate that WIF-B growth rate could be used to differentiate subsets of students who start first grade with low achievement (i.e., to distinguish students with persistent low achievement from those who profit from instruction). This possibility warrants further investigation.

Despite the promise of these findings, certain limitations are worth noting. First, we collected information about the number of students who received special education, but did not collect data on eligibility categories or subject areas under which they qualified. This may limit generalizability because students may have received language or reading services that impacted their performance or progress on our assessments. Also, data were collected for 17 calendar weeks (15 school weeks), whereas a typical school year spans 40 to 42 calendar weeks. It would be useful to know how growth rates change over an entire school year. Further, fall data were not collected until several weeks into the year, so students would have had the opportunity to benefit from some reading instruction. Finally, it would be useful to know if further differentiating sampling procedures increases sensitivity of WIF-B to student growth. With these caveats in mind, results suggest that practitioners should use WIF-B for progress monitoring, but rely on WIF-N for screening, particularly when assessing students across achievement levels.

Acknowledgments

References

Ardoin SP, Christ TJ. Evaluating curriculum-based measurement slope estimates using data from triannual universal screenings. School Psychology Review. 2008; 37:109–125.

Baker SK, Smolkowski K, Katz R, Fien H, Seelety JR, Kamen'enui EJ, Beck CT. Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. School Psychology Review. 2008; 37:18–37.

Brownell MT, Sindelar PT, Kiely MT, Danielson LC. Special education teacher quality and preparation: Exposing foundations, constructing a new model. Exceptional Children. 2010; 76:357–377.

Catts HW, Petscher Y, Schatschneider C, Bridges MS, Mendoza K. Floor effects associated with universal screening and their impact on the early identification of reading disabilities. Journal of Learning Disabilities. 2009; 42:163–176.10.1177/0022219408326219 [PubMed: 19098274]

Clemens, NH. unpublished doctoral dissertation. Lehigh University; Bethlehem, PA: 2009. Toward consensus on first grade CBM measures.

Compton DL, Fuchs D, Fuchs LS, Bouton B, Gilbert JK, Barquero LA, Crouch RC. Selecting at-risk readers in first grade for early intervention: Eliminating false positives and exploring the promise of a two-stage screening process. Journal of Educational Psychology. 2010; 102:327–340.10.1037/a0018448 [PubMed: 20689725]

Compton DL, Fuchs LS, Fuchs D, 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; 98:394–409.10.1037/0022-0663.98.2.394

Crocker, L.; Algina, J. Introduction to classical and modern test theory. Belmont, CA: Wadsworth; 1986.

Deno SL. Curriculum-based measurement: The emerging alternative. Exceptional Children. 1985; 59:219–232. [PubMed: 2934262]

Deno SL. Developments in curriculum-based measurement. The Journal of Special Education. 2003; 37:184–192.10.1177/00224669030370030801

Deno SL, Mirkin PK, Chiang B. Identifying valid measures of reading. Exceptional Children. 1982; 49:36–45. [PubMed: 7140791]

Foorman BR, Francis DJ, Beeler T, Winikates D, Fletcher JM. Early interventions for children with reading problems: Study designs and preliminary findings. Learning Disabilities: A Multidisciptinary Journal. 1997; 8:63–71.

Fuchs D, Fuchs LS, Thompson A, Al Otaiba S, Yen L, Yang NJ, O'Connor R. Is reading important in reading readiness programs? A randomized field trial with teachers as program implemented. Journal of Educational Psychology. 2001; 93:251–267.10.1037/0022-0663.93.2.251

Fuchs LS, Deno SL, Mirkin PK. The effects of curriculum-based measurement evaluation on pedagogy, student achievement, and student awareness of learning. American Educational Research Journal. 1984; 21(2):449–460.

Fuchs, LS.; Fuchs, D. Progress monitoring with letter-sound fluency: Technical data. L. S. Fuchs; 328 Peabody, Vanderbilt University, Nashville, TN 37203: 2001.

Fuchs LS, Fuchs D, Compton DL. Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. Exceptional Children. 2004; 71:7–21.

Fuchs LS, Fuchs D, Hamlett CL. Monitoring reading growth using student recalls: Effects of two teacher feedback systems. Journal of Educational Research. 1989; 83:103–111.

Fuchs LS, Fuchs DF, Hosp MK, Jenkins JR. Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. Scientific Studies of Reading. 2001; 5:239–256.

Good, RH.; Kaminski, RA., editors. Dynamic indicators of basic early literacy skills. 6. Eugene, OR: Institute for the Development of Educational Achievement; 2002. Retrieved from http://dibels.uoregon.edu/

Good RH III, Simmons DC, Kame'enui EJ. The importance of decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. Scientific Studies of Reading. 2001; 5:257–288.10.1207/S1532799XSSR0503_4

Healy KD. Word identification fluency and nonsense word fluency as predictors of reading fluency in first grade. Dissertation Abstracts International. 2007; 69(01):A.

Juel C. Learning to read and write: A longitudinal study of 54 children from first through fourth grades. Journal of Educational Psychology. 1988; 80:437–447.10.1037/0022-0663.80.4.437

LaBerge D, Samuels SJ. Toward a theory of automatic information processing in reading. Cognitive Psychology. 1974; 6:293–323.10.1016/0010-0285(74)90015-2

Manzo KK. National clout of DIBELS test draws scrutiny. Education Week. 2005; 25:11–12.

Meisinger EB, Bloom JS, Hynd GW. Reading fluency: Implications for the assessment of children with reading disabilities. Annals of Dyslexia. 2010; 60:1–17.10.1007/s11881-009-0031-z [PubMed: 20033795]

MPlus (Version 6.1) [Computer Software]. Los Angeles, CA: Muthen & Muthen;

National Center on Response to Intervention. Essential components of RTI: A closer look at response to intervention. 2010. Retrieved from http://www.cldinternational.org/Articles/rtiessentialcomponents.pdf

Nimon, K.; Nathans, LL.; Henson, RK. Commonality analysis: A promising strategy for understanding regression effects. Paper presented at the annual meeting of the Southwest Educational Research Association; New Orleans, LA. 2010 Feb.

Raudenbush, SW.; Bryk, AS. Hierarchical linear models: Application and data analysis methods. Thousand Oaks, CA: Sage; 2002.

Schatschneider C, Wagner RK, Crawford EC. The importance of measuring growth in response to intervention models: Testing a core assumption. Learning and Individual Differences. 2008; 18:308–315.10.1016/j.lindif.2008.04.005 [PubMed: 22224065]

Stage SA, Jacobsen MD. Predicting student success on a state-mandated performance-based assessment using oral reading fluency. School Psychology Review. 2001; 30:407–419.

Torgesen, JK.; Wagner, RK.; Rashotte, CA. Test of word reading efficiency. Austin, TX: Pro-Ed; 1999.

Torgesen JK, Wagner RK, Rashotte CA, Rose E, Lindamood P, Conway T, Garvin C. Preventing reading failure in your children with phonological processing disabilities: Group and individual responses to instruction. Journal of Educational Psychology. 1999; 91:579–593.10.1037/0022-0663.91.4.579

Vaughn S, Fuchs LS. Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. Learning Disabilities Research & Practice. 2003; 18(3):137–146.10.1111/1540-5826.00070

Walker, HM.; Lev, J. Statistical inference. New York, NY: Holt & Co; 1953.

Woodcock, RW. Woodcock Reading Mastery Test—Revised. Circle Pines, MN: American Guidance Service; 1998.

Zeno, SM.; Ivens, SH.; Millard, RT.; Duvvuri, R. The educator's word frequency guide. New York, NY: Touchstone Science Associates; 1995.

TABLE 1

Demographic Information

Demographic Variable	Representative (n = 284) %(n)		LSE (n = 202) %(n)		Not-LSE (n = 213) %(n)	
Gender						
Male	45.4	(129)	53.5	(108)	47.4	(101)
Female	47.9	(136)	43.1	(87)	51.2	(109)
Missing	6.7	(19)	3.5	(7)	1.4	(3)
Ethnicity						
Caucasian	34.9	(99)	32.2	(65)	45.1	(96)
African American	47.5	(135)	57.4	(116)	31.5	(67)
Hispanic	3.5	(10)	1.0	(2)	8.0	(17)
Asian	2.5	(7)	1.0	(2)	5.6	(12)
Kurdish	1.8	(5)	2.5	(5)	3.3	(7)
Biracial	2.5	(7)	3.5	(7)	1.4	(3)
Other	1.8	(5)	1.5	(3)	1.4	(30)
Missing	5.6	(16)	1.0	(2)	3.8	(8)
Subsidized Lunch	50.4	(143)	56.4	(114)	42.7	(91)
Missing	7.0	(20)	11.9	(24)	6.6	(14)
IEP	8.5	(24)	10.4	(21)	8.9	(19)
Missing	1.8	(5)	2.0	(4)	2.8	(6)
Retained	8.1	(23)	10.4	(21)	2.8	(6)
Missing	1.8	(5)	0.5	(1)	0.5	(1)
ELL	0.0	(0)	0.0	(0)	0.0	(0)
Missing	1.4	(4)	0.5	(1)	0.5	(1)

Note. LSE = low study-entry; Not-LSE = not-low study-entry; IEP = individualized education program; ELL = English language learners.

TABLE 2

Means and Standard Deviations for the Representative Sample and Subsamples, Fall Testing Sessions

Criterion Variable	*M*	*(SD)*	*M*	*(SD)*
WRMT-RWID				
Representative Sample[a]	33.93	(14.96)	43.94	(14.02)
LSE	18.57	(10.66)	29.15	(12.87)
Not-LSE	38.96	(12.74)	49.00	(10.89)
WRMT-R WAT				
Representative Sample	12.55	(12.55)	17.08	(9.98)
LSE	5.12	(5.22)	8.93	(7.62)
Not-LSE	15.25	(8.66)	20.32	(8.85)
CRAB Fluency				
Representative Sample	32.63	(30.59)	64.21	(39.61)
LSE	8.93	(8.28)	27.38	(20.60)
Not-LSE	41.99	(30.84)	77.12	(36.18)
TOWRE Sight Word Efficiency				
Representative Sample	29.04	(15.52)	43.07	(16.18)
LSE	13.86	(7.58)	26.28	(12.59)
Not-LSE	34.03	(14.26)	48.81	(13.06)
TOWRE Decoding Efficiency				
Representative Sample	10.67	(8.80)	15.56	(10.7)
LSE	3.59	(3.94)	6.73	(6.61)
Not-LSE	13.15	(8.64)	18.65	(10.11)
CBM Level				
WIF-N				
Representative Sample	43.91	(25.32)	58.74	(27.80)
LSE	17.74	(13.45)	31.12	(20.30)
Not-LSE	52.84	(21.95)	68.12	(23.10)
WIF-B				
Representative Sample	26.82	(20.70)	41.00	(26.11)
LSE	10.09	(6.46)	17.60	(13.52)
Not-LSE	32.53	(20.85)	48.79	(24.53)

Note: WRMT-R = Woodcock Reading Mastery Test-Revised; WID = Word Identification subtest; WAT = Word Attack subtest; CRAB = Comprehensive Reading Assessment Battery; TOWRE = Test of Oral Word Reading Efficiency; CBM = curriculum-based measurement; WIF-N = word identification fluency-narrow sampling; WIF-B = word identification fluency-broad sampling. LSE = low study-entry; Not-LSE = not-low study-entry.

[a] Representative sample, *n* = 284; LSE subsample, *n* = 202; not-LSE subsample, *n* = 213.

TABLE 3

CBM Instantaneous Growth Rates

CBM Measures	Week 1		Week 9		Week 17	
	X	(SE)	X	(SE)	X	(SE)
WIF-N						
Representative Sample[a]	0.55	(0.11)	0.82	(0.05)	1.07	(0.10)
LSE	0.63	(0.10)	0.79	(0.05)	0.93	(0.10)
Not-LSE	0.61	(0.13)	0.86	(0.06)	1.09	(0.12)
WIF-B						
Representative Sample[a]	0.59	(0.08)	0.90	(0.04)	1.17	(0.08)
LSE	0.17	(0.06)	0.52	(0.04)	0.84	(0.07)
Not-LSE	0.76	(0.10)	1.05	(0.04)	1.31	(0.09)

Note. Instantaneous growth rates are reported rather than linear slopes to reflect significant quadratic terms. CBM = curriculum-based measurement; WIF-N = word identification fluency, narrow sampling; WIF-B = word identification fluency, broad sampling; LSE = low study-entry; Not-LSE = not-low study-entry.

[a] Representative sample, *n* = 284; LSE subsample, *n* = 202; not-LSE subsample, *n* = 213.

TABLE 4

Predictive Validity for WIF-N versus WIF-B Level and Growth

	CBM Measure			
Criterion Variable	**WIF-N**	**WIF-B**	*t* value[a]	*p*-value
Fall CBM Level				
WRMT-RWID				
Representative Sample	.86[**]	.81[**]	3.71	.00
LSE	.78[**]	.83[**]	−2.99	.00
Not-LSE	.81[**]	.80[**]	0.53	.60
WRMT-WAT				
Representative Sample	.76[**]	.73[**]	1.75	.08
LSE	.63[**]	.65[**]	−0.88	.38
Not-LSE	.64[**]	.65[**]	−0.40	.69
CRAB Fluency				
Representative Sample	.89[**]	.88[**]	0.89	.37
LSE	.81[**]	.79[**]	1.16	.25
Not-LSE	.87[**]	.86[**]	0.65	.52
TOWRE Sight Word				
Representative Sample	.91[**]	.82[**]	8.09	.00
LSE	.82[**]	.81[**]	0.60	.55
Not-LSE	.87[**]	.81[**]	3.65	.00
TOWRE Decoding				
Representative Sample	.80[**]	.80[**]	0.0	1.00
LSE	.65[**]	.65[**]	0.0	1.00
Not-LSE	.72[**]	.75[**]	−1.36	.17
CBM Instantaneous Growth Rates[b]				
WRMT-R WID				
Representative Sample				
Growth Rate: Week 1	−.01	.30[**]	−4.85	.00
Growth Rate: Week 9	.05	.55[**]	−10.80	.00
Growth Rate: Week 17	.06	.25[**]	−3.11	.00
LSE				
Growth Rate: Week 1	.24[**]	.33[**]	−1.28	.20
Growth Rate: Week 9	.53[**]	.70[**]	−3.94	.00
Growth Rate: Week 17	.28[**]	.50[**]	−3.49	.00
Not-LSE				
Growth Rate: Week 1	−.13	.21[**]	−4.49	.00
Growth Rate: Week 9	−.12	.37[**]	−8.36	.00

	CBM Measure			
Criterion Variable	**WIF-N**	**WIF-B**	***t* value[a]**	***p*-value**
Growth Rate: Week 17	.03	.13	−1.37	.17
WRMT-RWAT				
Representative Sample				
Growth Rate: Week 1	.03	.30[**]	−4.20	.00
Growth Rate: Week 9	.08	.51[**]	−8.80	.00
Growth Rate: Week 17	.05	.21[**]	−2.59	.01
LSE				
Growth Rate: Week 1	.25[**]	.36[**]	−1.59	.11
Growth Rate: Week 9	.46[**]	.68[**]	−4.94	.00
Growth Rate: Week 17	.21[**]	.46[**]	−3.87	.00
Not-LSE				
Growth Rate: Week 1	−.05	.20[**]	3.24	.00
Growth Rate: Week 9	−.01	.34[**]	−5.59	.00
Growth Rate: Week 17	.04	.11	−0.96	.34
CRAB Fluency				
Representative Sample				
Growth Rate: Week 1	.02	.30[**]	−4.36	.00
Growth Rate: Week 9	.06	.58[**]	−15.65	.00
Growth Rate: Week 17	.04	.27[**]	−3.79	.00
LSE				
Growth Rate: Week 1	.32[**]	.39[**]	−1.03	.30
Growth Rate: Week 9	.58[**]	.81[**]	−3.94	.00
Growth Rate: Week 17	.26[**]	.57[**]	−5.16	.00
Not-LSE				
Growth Rate: Week 1	−.07	.21[**]	−3.65	.00
Growth Rate: Week 9	−.05	.43[**]	−8.28	.00
Growth Rate: Week 17	.03	.18[*]	−2.07	.04
TOWRE Sight Word				
Representative Sample				
Growth Rate: Week 1	.05	.25[**]	−4.93	.00
Growth Rate: Week 9	.11	.39[**]	−12.85	.00
Growth Rate: Week 17	.06	.13[*]	−3.81	.00
LSE				
Growth Rate: Week 1	.28[**]	.37[**]	−1.31	.19
Growth Rate: Week 9	.56[**]	.78[**]	−5.76	.00
Growth Rate: Week 17	.28[**]	.55[**]	−4.22	.00
Not-LSE				

| | CBM Measure | | | |
Criterion Variable	WIF-N	WIF-B	*t* value[a]	*p*-value
Growth Rate: Week 1	−.04	.28**	−4.25	.00
Growth Rate: Week 9	−.03	.51**	−9.95	.00
Growth Rate: Week 17	.01	.18*	−2.35	.02
TOWRE Decoding				
Representative Sample				
Growth Rate: Week 1	.05	.34**	−4.57	.00
Growth Rate: Week 9	.12	.56**	−9.30	.00
Growth Rate: Week 17	.07	.22**	−2.43	.02
LSE				
Growth Rate: Week 1	.25**	.40**	−2.20	.03
Growth Rate: Week 9	.47**	.69**	−5.00	.00
Growth Rate: Week 17	.23**	.44**	−3.22	.00
Not-LSE				
Growth Rate: Week 1	−.02	.25**	−3.54	.00
Growth Rate: Week 9	.04	.42**	−6.27	.00
Growth Rate: Week 17	.06	.13	−0.96	.34

Note. WIF-N = word identification fluency-narrow sampling; WIF-B = word identification fluency-broad sampling; WRMT-R = Woodcock Reading Mastery Test-Revised; WID = Word Identification subtest; WAT = Word Attack subtest; LSE = low study-entry; Not-LSE = not-low study-entry; CRAB = Comprehensive Reading Assessment Battery; TOWRE = Test of Oral Word Reading Efficiency; CBM = curriculum-based measurement.

[a] Degrees of freedom for the representative sample, low study-entry, and not-low study-entry subsamples were t(281), (199), and (201), respectively.

[b] Data were collected over 17 weeks to accommodate school holidays.

* $p<.05$.

** $p<.01$.

TABLE 5

Overall, Unique, and Common Variance Explained by WIF-N and WIF-B for the Representative Sample ($n = 284$)

WIF Measure	WRMT-R WID R^2	WRMT-R WAT R^2	CRAB R^2	TOWRE SW R^2	TOWRE-D R^2
Fall WIF Level					
WIF-N	74.7***	57.9***	81.6***	82.4***	63.3***
WIF-B	65.1***	52.9***	77.9***	67.6***	63.7***
WIF-N + WIF-B	75.2***	58.9***	84.3***	82.5***	67.0***
WIF-N Unique	10.1***	6.1***	6.4***	14.9***	3.3***
WIF-B Unique	0.6*	1.1**	2.7***	0.0	3.7***
Common	64.6	51.7	75.2	67.6	60.0
Week 1 Growth Rate					
WIF-N	0.0	0.1	0.0	0.2	0.2
WIF-B	8.8***	8.7***	9.1***	12.7***	11.6***
WIF-N + WIF-B	10.5***	9.4***	10.1***	13.5***	12.2***
WIF-N Unique	1.7*	0.7	1.0	0.8	0.6
WIF-B Unique	10.5***	9.4***	10.1***	13.3***	11.9***
Common	−1.7[a]	−0.7[a]	−1.0[a]	−0.6[a]	−0.3[a]
Week 9 Growth Rate					
WIF-N	0.3	0.6	0.4	1.2	1.5
WIF-B	30.2***	26.2***	33.2***	42.2***	31.7***
WIF-N + WIF-B	37.6***	30.9***	41.1***	49.3***	35.5***
WIF-N Unique	7.4***	4.7***	7.9***	7.0***	3.9***
WIF-B Unique	37.3***	30.3***	40.7***	48.0***	34.0***
Common	−7.1[a]	−4.1[a]	−7.5[a]	−5.7[a]	2.4[a]
Week 17 Growth Rate					
WIF-N	0.4	0.2	0.2	0.3	0.4
WIF-B	6.1***	4.5**	7.2***	8.2***	4.8**
WIF-N + WIF-B	6.4***	4.8**	8.0***	8.8***	4.9**
WIF-N Unique	0.3	0.3	0.8	0.6	0.1

WIF Measure	WRMT-R WID R^2	WRMT-R WAT R^2	CRAB R^2	TOWRE SW R^2	TOWRE-D R^2
WIF-B Unique	6.0 ***	4.6 ***	7.8 ***	8.5 ***	4.5 ***
Common	0.1	−0.1[a]	−0.6[a]	−0.3[a]	0.3

Note. WIF-N = word identification fluency-narrow sampling; WIF-B = word identification fluency-broad sampling; WRMT-R = Woodcock Reading Mastery Test-Revised; WID = Word Identification subtest; WAT = Word Attack subtest; CRAB = Comprehensive Reading Assessment Battery; TOWRE-SW = Test of Oral Word Reading Efficiency-Sight Word; TOWRE-D = Test of Oral Word Reading Efficiency-Decoding.

[a] = suppression effect.

*
p<.05;

**
p<.01;

***
p<.001.

TABLE 6

Overall, Unique, and Common Variance Explained by WIF-N and WIF-Bfor the LSE Subsample ($n = 202$)

WIF Measure	WRMT-R WID R^2	WRMT-R WAT R^2	CRAB R^2	TOWRE SW R^2	TOWRE D R^2
Fall WIF Level					
WIF-N	60.7 ***	39.1 ***	65.0 ***	67.1 ***	42.0 ***
WIF-B	68.4 ***	42.2 ***	63.0 ***	66.1 ***	42.6 ***
WIF-N + WIF-B	68.8 ***	42.8 ***	67.1 ***	69.7 ***	44.3 ***
WIF-N Unique	0.4	0.6	4.0 ***	2.6 ***	1.7 *
WIF-B Unique	8.1 ***	3.7 ***	2.0 ***	3.6 ***	2.3 **
Common	60.3	38.5	61.1	63.5	40.3
Week 1 Growth Rate					
WIF-N	5.7 **	6.1 ***	9.9 ***	7.9 ***	6.0 ***
WTF-B	11.0 ***	13.1 ***	14.9 ***	13.9 ***	15.6 ***
WIF-N + WIF-B	12.0 ***	12.0 ***	17 c ***	15.6 ***	16.2 ***
WIF-N Unique	1.0	0.9	2.5 *	1.6	0.6
WIF-B Unique	6.3 ***	7.9 ***	7.5 ***	7.6 ***	10.2 ***
Common	4.7	3.2	7.5	6.4	5.4
Week 9 Growth Rate					
WIF-N	28.0 ***	21.6 ***	33.9 ***	31.8 ***	22.5 ***
WIF-B	48.6 ***	46.1 ***	65.1 ***	60.9 ***	47.3 ***
WIF-N + WIF-B	49.7 ***	46.2 ***	65.8 ***	61.5 ***	47.5 ***
WIF-N Unique	1.1 *	0.1	0.7 *	0.7	0.2
WIF-B Unique	21.7 ***	24.7 ***	31.9 ***	29.7 ***	25.0 ***
Common	26.9	21.4	33.2	31.1	22.3
Week 17 Growth Rate					
WIF-N	7.9 ***	4.6 **	6.9 ***	7.6 ***	5.1 **
WTF-B	24.5 ***	20.7 ***	32.7	30.6 ***	19.4 ***
WIF-N + WIF-B	24.7 ***	20.7 ***	32.7	30.6 ***	19.4 ***

WIF Measure	WRMT-R WID R^2	WRMT-R WAT R^2	CRAB R^2	TOWRE SW R^2	TOWRE D R^2
WIF-N Unique	0.2	0.0	0.0	0.0	0.0
WIF-B Unique	16.8***	16.1***	25.7***	23.0***	144***
Common	7.7	4.6	7.0	7.6	5.0

Note. WIF-N = word identification fluency-narrow sampling; WIF-B = word identification fluency-broad sampling; LSE = low study-entry subsample; WRMT-R = Woodcock Reading Mastery Test-Revised; WID = Word Identification subtest; WAT = Word Attack subtest; CRAB = Comprehensive Reading Assessment Battery; TOWRE-SW = Test of Oral Word Reading Efficiency-Sight Word; TOWRE-D = Test of Oral Word Reading Efficiency-Decoding.

*
$p < .05$.

**
$p < .01$.

***
$p < .001$.

TABLE 7

Overall, Unique, and Common Variance Explained by WIF-N and WIF-B for the Not-LSE Subsample (n = 213)

WIF Measure	WRMT-R WID R^2	WRMT-R WAT R^2	CRAB R^2	TOWRE-SW R^2	TOWRE D R^2
Fall WIF Level					
WIF-N	66.2***	40.9***	74.8***	76.4***	51.6***
WIF-B	63.9***	42.2***	74.1***	65.4***	56.7***
WIF-N + WIF-B	69.2***	44.2***	79.1***	77.1***	58.0***
WIF-N Unique	5.3***	2.0**	5.0***	11.7***	1.3*
WIF-B Unique	3.0***	3.3***	4.4***	0.6*	6.5***
Common	60.9	38.9	69.7	64.8	50.2
Week 1 Growth Rate					
WIF-N	1.8	0.2	0.5	0.1	0.1
WIF-B	4.2**	4.1**	4.3**	7.9***	6.4**
WIF-N + WIF-B	8.9***	5.6**	6.5**	9.8***	7.8***
WIF-N Unique	4.7**	1.6	2.2*	1.9*	1.4
WIF-B Unique	7.1***	5.4**	6.0***	9.7***	7.7***
Common	2.9	−1.3[a]	−1.7[a]	−1.8[a]	−1.3[a]
Week 9 Growth Rate					
WIF-N	1.4	0.0	0.2	0.1	0.1
WIF-B	13.8***	11.7***	18.4***	26.1***	17.9***
WIF-N + WIF-B	26.8***	16.3***	28.1***	37.2***	22.2***
WIF-N Unique	13.0***	4.6**	9.7***	11.1***	4.3**
WIF-B Unique	25.4***	16.3***	27.9***	37.2***	22.0***
Common	−11.6[a]	−4.6[a]	−9.5[a]	−11.1[a]	−4.1[a]
Week 17 Growth Rate					
WIF-N	0.1	0.2	0.1	0.0	0.4
WIF-B	1.6	1.1	3.1*	3.1*	1.6
WIF-N + WIF-B	1.7	1.1	3.4*	3.6*	1.6
WIF-N Unique	0.1	0.0	0.3	0.5	0.0

WIF Measure	WRMT-R WID R^2	WRMT-R WAT R^2	CRAB R^2	TOWRE-SW R^2	TOWRE D R^2
WIF-B Unique	1.7	0.9	3.4**	3.6**	1.2
Common	−0.1[a]	0.2	−0.3[a]	−0.5[a]	0.4

Note. WIF-N = word identification fluency-narrow sampling; WIF-B = word identification fluency-broad; Not-LSE = not-low study-entry subsample; WRMT-R = Woodcock Reading Mastery Test-Revised; WID = Word Identification subtest; WAT = Word Attack subtest; CRAB = Comprehensive Reading Assessment Battery; TOWRE SW = Test of Oral Word Reading Efficiency-Sight Word; TOWRE D = Test of Oral Word Reading Efficiency-Decoding.

[a] = suppression effect.

*
$p < .05.$

**
$p < .01.$

***
$p < .001.$

TABLE 8

Commonality Analysis Showing Proportion of Variance Explained by WIF-N Level and WIF-B Week 9 Growth Rate (G2)

Model	WID B (SE)	WAT B (SE)	CRAB B (SE)	TOWRE-SW B (SE)	TOWRE-D B (SE)
Representative Sample: $B_0 + B1_{WIF-N} + B2_{WIF-BG2}$					
Intercept (B_0)	21.84 (.85) ***	3.09 (.78) ***	−.51 (1.99)	15.68 (73) ***	−.27 (.76)
WIF-N (B_1)	.43 (.02) ***	.26 (.02) ***	1.27 (.05) ***	.49 (.02) ***	.29 (.02) ***
WIF-B G2 (B_2)	3.22 (.74) ***	2.50 (.68) ***	9.22 (1.72) ***	6.10 (.63) ***	3.49 (.66) ***
Model R^2 %	76.2 ***	59.9 ***	81.6 ***	86.8 ***	66.6 ***
Unique R^2 % (WIF-N level)	44.7 ***	33.0 ***	49.1 ***	43.9 ***	33.8 ***
Unique R^2 % (WIF-B G2)	1.7 ***	2.0 ***	1.7 ***	4.5 ***	3.4 ***
Common R^2 %	28.7	24.2	33.2	37.8	28.4
Low Study-Entry: $B_0 + B1_{WIF-N} + B2_{WIF-BG2}$					
Intercept (B_0)	15.50 (.87) ***	2.25 (.63) ***	4.02(1.07) ***	12.02 (.68) ***	.69 (.53)
WIF-N (B_1)	.55 (.05) ***	.20 (.04) ***	.76 (.06) ***	.51 (.04) ***	.18 (.03) ***
WIF-B G2 (B_2)	7.49 (1.20) ***	6.01 (.86) ***	19.03 (1.47) ***	10.12 (.93) ***	5.52 (73) ***
Model R^2 %	67.3 ***	51.4 ***	81.3 ***	79.6 ***	55.3 ***
Unique R^2 % (WIF-N level)	20.8 ***	7.7 ***	15.1 ***	12.5 ***	13.3 ***
Unique R^2 % (WIF-B G2)	6.6 ***	12.3 ***	16.2 ***	18.4 ***	8.2 ***
Common R^2 %	42.0	33.8	48.8	48.4	34.0
Not-Low Study-Entry: $B_0 + B1_{WIF-N} + B2_{WIF-BG2}$					
Intercept (B_0)	26.68(1.19) ***	5.67(1.29) ***	−2.07 (3.29)	18.95 (1.07) ***	−.50(1.28)
WIF-N (B_1)	.38 (.02) ***	.24 (.02) ***	1.31 (.06) ***	.47 (.02) ***	.29 (.02) ***
WIF-B G2 (B_2)	1.98(73) **	1.95 (.79) *	8.55 (2.02) ***	4.83 (.66) ***	3.28 (79) ***
Model R^2 %	67.4 ***	42.6 ***	76.8 ***	81.4 ***	55.4 ***
Unique R^2 % (WIF-N level)	52.3 ***	30.5 ***	57.6 ***	54.8 ***	36.5 ***
Unique R^{2}% (WIF-B G2)	1.2 **	1.7 *	2.0 ***	4.9 ***	3.8 ***
Common R^2 %	12.6	10.0	16.4	21.1	14.1

Note. WIF-N = word identification fluency-narrow sampling; WIF-B = word identification fluency-broad sampling; WRMT-R = Woodcock Reading Mastery Test-Revised; WID = Word Identification subtest; WAT = Word Attack subtest; CRAB = Comprehensive Reading Assessment Battery; TOWRE-SW = Test of Oral Word Reading Efficiency-Sight Word; TOWRE-D = Test of Oral Word Reading Efficiency-Decoding.

*
$p<.05.$

**
$p<.01.$

***
$p<.001.$