10-4-2016

Comparing Two CBM Maze Selection Tools: Considering Scoring and Interpretive Metrics for Universal Screening

Jeremy W. Ford
Boise State University

Kristen N. Missall
University of Washington

John L. Hosp
University of Massachusetts Amherst

Jennifer L. Kuhle
Mississippi Bend Area Education Agency

Comparing Two CBM Maze Selection Tools: Considering Scoring and Interpretive Metrics for Universal Screening

Jeremy W. Ford, Ph.D.
Boise State University

Kristen N. Missall, Ph.D.
University of Washington

John L. Hosp, Ph.D.
University of Massachusetts Amherst

and

Jennifer L. Kuhle, Ph.D.
Mississippi Bend Area Education Agency

Author Note: This research was supported by the Iowa Department of Education through contract #66511 to The University of Iowa. The opinions expressed are expressed are those of the authors and do not represent the views of the Iowa Department of Education.

Correspondence concerning this article should be addressed to Jeremy W. Ford, Department of Early & Special Education, Education Building Room 204, Boise State University, Boise, ID 83725. Emails of authors: Jeremy W. Ford jwford@boisestate.edu; Kristen N. Missall, kmissall@uw.edu; John L. Hosp, johnhosp@umass.edu; Jennifer L. Kuhle, jkuhle@aea9.k12.ia.us

Abstract

Advances in maze selection curriculum-based measurement (CBM) have led to several published tools with technical information for interpretation (e.g., norms, benchmarks, cut-scores, classification accuracy) that have increased their usefulness for universal screening. A range of scoring practices have emerged for evaluating student performance on maze selection (e.g., correct restoration, incorrect restoration, correct restoration minus incorrect restoration, and correct restoration minus one-half incorrect restoration). However, lack of clear understanding about the intersection between scoring and interpretation has resulted in limited evidence about using maze selection for making universal screening decisions. In this study, 925 students in Grades 3-6 completed two CBMs for maze selection. Student performance on the two was compared across different scoring metrics. Limitations and practical implications are discussed.

Keywords: maze, maze selection, universal screening, resource allocation

Curriculum-based measurement (CBM) emphasizes direct measurement of a variety of academic skills, including early literacy (e.g., phonological awareness, letter naming, letter sounds), reading (e.g., reading rate, comprehension), mathematics (e.g., computation, concepts, applications), and written language (e.g., words spelled correctly, total words written, correct writing sequence). CBM can be used to assess the effectiveness of interventions designed to remediate or prevent academic skill difficulties (Deno, 1985). In reading, the metric of Words Correct per Minute (WCPM) collected through CBM-Reading (CBM-R) passages is commonly used as an indicator of overall reading proficiency (Pierce, McMaster, & Deno, 2010). WCPM is determined by an examiner calculating the number of correct words a student reads aloud from a passage in 1 min (Hosp, Hosp, & Howell, 2016).

Despite the empirical support and prevalence of CBM-R for measuring reading skills of students in Grades 1 through 6, there are three specific problems with its use. First, the requirement for individual student administration takes time away from classroom instruction and may take more time than administration guidelines suggest with the addition of transitions (Ford & Hosp, 2015). This may be undesirable as instruction should be obstructed minimally by assessment, with the latter conducted only to inform the former (Hosp & Ardoin, 2008). Second, the WCPM metric is less sensitive to student skill acquisition in Grades 4 and 5, as compared to other elementary grades (Hintze & Shapiro, 1997; Jenkins & Jewell, 1993; Needenriep, Hale, Skinner, Hawkins, & Winn, 2007). Third, while CBM-R is used commonly as a measure of reading comprehension there are cautions; CBM-R is not a direct measure of reading comprehension, but it correlates strongly with standardized measures of reading comprehension (Potter & Wamre, 1990; Skinner, Needenriep, Bradley-Klug, & Ziemann, 2002). However, concerns with the face validity of CBM-R for measuring reading comprehension (Parker, Hasbrouck, & Tindal, 1992; Wayman, Wallace, Wiley, Tichà, & Espin, 2007) suggest more direct measures of reading comprehension (e.g., maze selection) may be necessary.

To complete maze selection (see Hosp, Hosp, & Howell, 2016), students read a passage silently for 1-3 min. Often the first and last sentences of the passage remain intact while every 7th word is deleted throughout the rest of the passage. Where words are deleted, students are provided with three replacement words, including the answer and two distractors. Students are directed to circle the word that best fits the sentence. Performance on maze selection typically is measured by the total number of correct restorations (Silberglitt, Burns, Madyun, & Lail, 2006). Maze selection, as an example of CBM, uses standardized procedures for repeatedly assessing progress toward long-term goals (Hosp et al., 2016). When compared to read aloud measures, maze selection has demonstrated similar technical characteristics such as sensitivity to growth and standard errors of estimate (Fuchs & Fuchs, 1992). Investigation of alternate-form reliability has found correlation coefficients around .80 and reliable estimates of growth at the individual and aggregate levels (Shin, Deno, & Espin, 2000). Examination of criterion validity across a broad range of reading comprehension measures has found correlations ranging from .56 to .86 (Wayman et al., 2007). Maze selection can be group administered, which may make the measure more feasible than CBM-R for use in schools. That is, administering maze selection to an entire class takes approximately 5 min, whereas individually administering CBM-R to an entire class may take 40 to 50 min given a class of 25 students. Given less time for assessment, more time is available for instruction.

Although school psychologists and researchers tend to perceive maze selection as more aligned with reading comprehension than overall reading skill, challenges to such claims have been made (January & Ardoin, 2012). One challenge is that maze selection functions more as a measure of silent sentence reading fluency rather than comprehension because students can answer items accurately without considering prior information contained in the passage (Kendeou, Papadopoulos, & Spanoudis, 2012). Another challenge is related to the number of distractors accompanying a correct answer. Parker and colleagues (1992) noted that with only three choices per item, the likelihood of a student providing a correct restoration is nearly 33%. A third challenge is the quality of the distractors. One approach to developing maze passages includes one distractor as a "far distractor" which is often grammatically incorrect. The second distractor is a "near distractor" that is grammatically correct, but does not relate to the content of the passage. As such, the accuracy of a student's response is greatly influenced by the quality of the distractors (Parker et al., 1992). Due to these measurement issues, other metrics of scoring than correct restorations for maze selection have been explored.

Maze Selection Metrics

One example of another metric for measuring student performance on maze selection is subtracting the total number of errors from the total number of correct restorations (Pierce et al., 2010). Another example is adjusting the total number of correct restorations by subtracting one-half the number of errors made by the student (Brown-Chidsey, Davis, & Maya, 2003). Brown-Chidsey et al. (2003) used this procedure based on recommendations by Deno et al. (2002) for controlling for the effects of random guessing. Also, to control for random guessing, some researchers have advocated discontinuing counting student restorations after two (Deno et al., 2002) or three consecutive errors (Deno et al., 2002; Jenkins & Jewell, 1992).

Deno et al. (2002) examined terminating counting correct restorations after two or three consecutive errors. Differences in the correlations between maze selection performance and standardized reading test scores were not found. Deno et al. (2002) also examined the effects of each of these scoring metrics on identifying the false-negatives for determining students at-risk in Grades 2 through 6. Results indicated that termination after two consecutive errors

decreased the likelihood of false-negatives for students in Grade 2. In addition, Deno et al. (2002) raised concerns that both of these metrics might increase false-positives and create extra within-student variance for students in older grades when neglecting correct restorations (i.e., correct maze selections after two or three consecutive errors). Thus, differences in scoring procedures for maze selection may differentially identify students at risk in reading.

Pierce et al. (2010) examined differences among five maze selection scoring metrics: (1) correct restorations, (2) correct restorations minus incorrect restorations, (3) correct restorations minus one-half of incorrect restorations, (4) ceasing to score after three consecutive incorrect restorations, and (5) ceasing to score after two consecutive incorrect restorations. Results found all metrics to be inter-correlated highly and to demonstrate acceptable criterion-related validity with a criterion measure of standardized reading. However, Pierce et al. (2010) did not examine false-positives to validate the hypothesis put forth by Deno et al. (2002).

In addition, no peer-reviewed research has examined the relation between different published CBMs for maze selection and a criterion measure of reading. Ardoin and Christ (2009) compared student performance across published CBM-R tools and noted statistically significant differences in scores, but they did not extend analyses to compare scores on maze selection. Differences in student performance scores across CBM maze selection tools could result in inconsistencies in identifying students likely to be proficient on a criterion outcome measure. Such an identification would occur typically during universal screening.

Maze Selection for Universal Screening

Universal screening involves the systematic assessment of all students in a class, grade, school, or district to help determine whether students are meeting expectations. Universal screening data typically are collected three times during the school year (e.g., fall, winter, and spring) and can be collected in individual or group settings. Ikeda, Neessen, and Witt (2008) discussed universal screening within the context of resource allocation and identified two primary purposes: (a) to determine if a school's core is meeting the needs of the majority of students, and (b) to identify students who may require additional (i.e., supplemental or intensive) resources. Thus, variation in how different published maze selection tools measure student performance could have a direct impact on the allocation of a school's resources.

The literature on CBM and universal screening has mostly focused on CBM-R. In one such study, Rowe, Witmer, Cook, and daCruz (2014) found teachers supported the use of CBM-R for making screening decisions. However, a dearth of research exists examining maze selection as a tool for universal screening. In one study Ardoin et al. (2004) investigated whether maze selection added any additional value to CBM-R WCPM for predicting state test scores for students in Grade 3. Using stepwise regression, Ardoin et al. (2004) found that maze selection did not explain significant variance when entered after CBM-R WCPM. The study did not report the amount of unique variance accounted for by maze selection.

In a second study involving English Learners (ELs) and non-ELs in Grades 3 and 5, Wiley and Deno (2005) examined the relative predictive validity between CBM-R and maze selection. Results determined maze selection to be a stronger predictor than CBM-R for non-EL students in Grade 5 for the state reading performance test. However, for EL and non-EL students in Grade 3 and EL students in Grade 5, results found maze selection and CBM-R to be comparable for predicting performance. Although this study used a small sample, maze selection was supported as an appropriate tool for universal screening for Grade 5 students.

In a third study, over a 7-year period Silberglitt et al. (2006) examined the relation between CBM-R and a state test of reading for students in Grades 4 through 8. For students in Grades 7 and 8 maze selection was also included. Results found a declining correlation between CBM-R and state reading test scores as students' grade increased. For students in middle school, CBM-R and maze selection were both observed to have moderate correlations with the state reading test. Although Silberglitt et al. (2006) found a correlation between maze selection and state test scores, only data from students in Grade 7 and 8 were included. Thus, their results did not advance the research on using maze selection for universal screening with students in Grades 4 and 5. However, a fourth study conducted by Graney, Martinez, Missall, and Aricak (2010) addressed this area. Graney et al. (2010) compared the technical adequacy of CBM-R and maze selection to inform their use in a universal screening process for students in Grades 4 and 5. Evidence of short- and long-term alternate forms reliability, criterion validity, and predictive validity was adequate. Results supported both CBM-R and maze selection for use in universal screening for Grades 4 and 5.

In addition to the studies above that examined the use of maze selection as a universal screening measure, Wayman et al. (2007) described the available research on the technical adequacy of maze selection. Despite suggestions that maze selection is a more appropriate measure than CBM-R for universal screening for students in late elementary school (Jenkins & Jewell, 1993; Wayman et al., 2007), Wayman et al. (2007) suggested more empirical evidence is needed to support this claim. Mercer et al. (2012) recently extended research on the technical adequacy of maze selection by investigating the number of probes and assessment duration (1-3 min.) necessary for reliable decisions. However, differences in potential metrics for maze selection across publishers have not been examined.

To build knowledge about CBM maze selection scoring, interpretation and use, the present study examined student performance on two published CBM maze selection tools. We also examined the extent to which publisher-provided comparison scores (i.e., "cut scores" from aimsweb and "benchmarks" from DIBELS Next) for each maze selection tool recommended a similar percentage of students to receive Tier I, II, or III instruction. Last, we examined the application of whether different CBMs for maze selection identified different students as unlikely to perform proficiently on a criterion measure of reading. Specific research questions included:

1. Are there significant differences in possible scoring metrics (i.e., correct restorations, incorrect restorations, correct restorations minus incorrect restorations, and/or correct restorations minus one-half incorrect restorations) on CBM maze selection across publishers?
2. Using publisher-provided comparison scores for interpreting scores on CBM maze selection, is there a difference in the percentage of students recommended for Tier I, II, and III instruction?
3. Using publisher-provided comparison scores for interpreting scores on CBM maze selection, is there a significant difference between publishers in classification accuracy when identifying students as proficient or non-proficient on a statewide, high-stakes assessment for accountability?

Method

Participant Identification and Selection

This study was conducted using data from a larger statewide project. The purpose of the larger project was to determine reading proficiency scores on commonly administered reading CBMs predictive of reading proficiency levels on statewide assessments. To accomplish this goal, pre-reading and reading skills were assessed in a representative state sample of students in kindergarten through Grade 6 using various CBMs. As such, at least one elementary school per area education agency (nine in this state) was included. Because several elementary schools in this sample enrolled students in kindergarten through Grade 5, three middle schools were also included in order to adequately sample Grade 6. In sum, 14 schools from 13 school districts participated.

Students were sampled at the building, as opposed to the individual, level to encompass the state's full geography (e.g., urban/rural, each education region). Sampling at the building level also allowed for representation of state demographics for race/ethnicity, disability status, family income, and primary home language. A total of 1,685 students were included in the statewide project of which 93.5% were white, 13.1% had an Individual Education Plan (IEP), 43.6% received free or reduced lunch, and 4.1% were English learners. From this sample, a subsample of students in grades 3 through 6 (N = 925) were selected for the current study if they completed one grade-level maze selection passage from aimsweb (Shinn & Shinn, 2002) and one grade-level maze selection passage from DIBELS Next (Good et al., 2011). All students in a grade completed the same aimsweb passage and the same DIBELS Next passage. Demographics for both the overall project, and this study, were consistent with state demographics (see Table 1).

Measures

Aimsweb Maze and DIBELS Next Daze reading passages for spring universal screening were administered. For aimsweb passages, the maximum number of possible correct restorations was 30 for Grade 3, 45 for Grade 4, 45 for Grade 5, and 51 for Grade 6. For DIBELS Next passages, the maximum number of possible correct restorations was 51 for Grade 3, 49 for Grade 4, 57 for Grade 5, and 64 for Grade 6.

Aimsweb Maze. When completing aimsweb Maze students read silently grade-level passages of 150 to 400 words for 3 min. The first sentence of the passage is intact and thereafter, every 7th word is replaced with parentheses containing three words listed horizontally in text. These three words contain the passage's correct answer and two incorrect answers. Incorrect answers serve as distractors. One distractor, commonly known as a "near distractor," is a word of the same part of speech as the correct answer (e.g., noun, verb, adverb) that does not make sense or preserve meaning. The second distractor, commonly known as a "far distractor," is selected randomly from the story and is a word of a different part of speech that does not preserve meaning. After students complete the reading, the passage is scored by determining the number of correct restorations made by a student.

Graney et al. (2010) found alternate-form reliability estimates for aimsweb Maze passages to be .82 in Grades 4 and 5. In addition, aimsweb Maze has been shown to have (a) sufficient sensitivity to growth across the school year, and (b) strong correlations with standardized reading and language arts tests (Espin, Wallace, Lembke, Campbell, & Long, 2010; Graney et al., 2010).

DIBELS Next Daze. When completing DIBELS Next Daze students read silently grade-level passages of 350 to 550 words for 3 min. The first sentence of the passage is intact, and thereafter, every 7th word is replaced with an in-text box containing three words listed vertically. These three words contain the passage's correct answer and two incorrect answers which serve as distractors. In development, distractors were selected randomly from a pool of words found in the passage and were only used once (Good et al., 2011). After students complete the reading, the passage is scored by determining the number of a student's correct restorations and the number of incorrect restorations. An adjusted score is then created (i.e., correct restorations – ½ incorrect restorations) to compensate for guessing.

Single-form alternate-form reliability for DIBELS Next Daze has ranged from .66 to .81 (Good et al., 2011). Good et al. (2011) also reported three-form alternate-form reliability from .85 to .93. Predictive validity with DIBELS Next and the Group Reading Assessment and Diagnostic Evaluation (GRADE; Williams, 2001) Total Test score has ranged from .56 to .67 at the beginning of the year to .56 to .61 at the middle of the year (Good et al., 2011). Concurrent validity with the GRADE Total Test score has ranged from .64 to .68 (Good et al., 2011).

Iowa Assessments. Students from the sampled state completed the Iowa Assessments (formerly the Iowa Tests of Basic Skills, ITBS; Hoover, Dunbar, & Frisbie, 2001) as the annual, high-stakes assessment of achievement. Schools in the study commonly administered the Iowa Assessments annually starting in Grade 3; however, some chose to do so in earlier grades as well. The Reading test of the Iowa Assessments is comprised of Parts 1 & 2 (Form E; http://itp.education.uiowa.edu/ia/default.aspx) for Grades 1 through 6. A student is considered to be proficient on the Iowas if his/her score is equal to, or greater than, the 41st percentile.

Procedures

Administration of maze selection passages occurred as part of the larger reading assessment project. The state's Department of Education predetermined the two sets of published maze selection materials used in this study based on examination of technical adequacy, ease of training and administration, and overall cost.

Training and reliability. Prior to administration, assessors were trained on each CBM tool and met procedural fidelity before data collection with students. A fidelity checklist for conducting CBM (Hosp et al., 2016) was modified for our purposes as not all steps were necessary for maze selection administration. For example, the fidelity checklist contains a step for "follows the procedure for time allowed on each item" during administration. This step was not applicable given the nature of administering maze selection. Individuals responsible for administering the maze selection task obtained 95% or higher fidelity prior to administration during observed practice. Assessors were either members of the research team or were individuals hired for additional assistance for collecting data at larger schools (e.g., state department of education employees, graduate students in school psychology). School site staff were not responsible for administration.

Data collection. CBM data were collected from individual schools within 1 to 2 days. Teams of assessors were assembled depending on the number of students across the grade levels and the school-specific requirements and timelines. Prior to school visits, one project staff contacted the principal and arranged days, testing spaces, and coordinated testing schedules across classrooms. All project data were collected within a six-week timeframe in the spring of the academic year.

During administration, each student was given one maze selection passage from each publisher. Passage order was predetermined and varied systematically over the course of the study. Passages were typically group-administered in a student's classroom following standardized directions. During administration the individual administering the maze selection passages and the classroom teacher monitored the classroom to ensure students did their own work. For students who missed the group administration (e.g., due to a scheduling conflict), maze selection passages were administered individually. Iowa Assessments were administered by the school following their typical assessment schedule. CBM data were collected within a two-week window of students completing the Iowa Assessments.

Data analysis. To analyze data for this study, we first separated and reviewed maze selection data from the larger statewide database to ensure that each entry (i.e., participant) included both aimsweb and DIBELS Next maze selection passages. Then, we calculated correct restorations (CR), incorrect restorations (ICR), CR minus ICR, and CR minus ½ ICR for each passage, and obtained descriptive statistics for each metric. To answer our first research question, we conducted pairwise comparisons of the mean scores for CR, ICR, CR minus ICR, and CR minus ½ ICR for all grades.

To answer our second research question, we identified the recommended comparison scores for guiding interpretation for each publisher. Thus for aimsweb, the document, "Aimsweb Default Cut Scores Explained" (Pearson Education, 2011) was used to determine the CR score necessary for instructional recommendations (i.e., Tier 1, II, or III) for end-of-year expectations. For DIBELS Next, the DIBELS Next Technical Manual (Dynamic Measurement Group, Inc., 2013) was used for this purpose. In order to answer our third research question, students' results were classified in a dichotomous fashion (i.e., proficient or not proficient) on each CBM maze selection tool and the Iowas prior to calculating both general and specific prediction metrics.

Results

Prior to presenting our results, we first provide descriptive statistics of students' performance for both aimsweb and DIBELS Next. Then, to answer our first research question, we present paired sample statistics for each maze selection metric included in the analysis for both publishers. Next, we present the results of applying publisher-provided comparison scores to identify students' recommended instructional placement (i.e., Tier I, II, III) to answer our second research question.

To address our third research question, we show student performance on aimsweb and DIBELS Next as compared to Iowa Assessments proficiency to inform whether students' instructional recommendations were identified as a true positive, false positive, false negative, or true negative. Last, to examine potential differences in the number of students identified as being unlikely to meet expectations (i.e., determining if there was a difference in the number of students identified as at-risk for not performing at or above the 41st percentile on the Iowa Assessments), we present the following prediction metrics for each publisher's materials at each grade level: Overall correct classification accuracy, Kappa, sensitivity, specificity, positive predictive power, and negative predictive power – along with false negative and positive rates – for each grade level (see Hosp, 2012). These metrics were tested using a two-proportions test (Sprinthall, 2003) with an alpha of 0.002 to correct for multiple comparisons.

Descriptive Statistics

Descriptive statistics are presented in Table 2. Students in all grades obtained more CR on DIBELS Next passages than on Aimsweb. On both aimsweb and DIBELS Next the number of CR increased by grade level. CR differences between the publishers were greater in Grades 3 and 4 than in Grades 5 and 6.

In general, students obtained more ICR on aimsweb than on DIBELS Next. The number of ICR students obtained on aimsweb was similar in Grades 3 and 4, whereas students obtained fewer ICR on DIBELS Next in Grades 3 and 4. Students obtained fewer ICR in Grade 5 than in Grades 3 and 4 on Aimsweb, but on DIBELS Next students obtained more ICR in Grade 5 than in Grades 3 and 4. In Grade 6 students obtained more ICR on aimsweb than in Grade 5, whereas on DIBELS Next students obtained fewer. Also in Grade 6 students obtained the same number of ICR on aimsweb and DIBELS Next.

The number of CR-ICR and CR- ½ ICR students obtained was higher on DIBELS Next than aimsweb for all grades. This is likely a result of students obtaining more CR on DIBELS Next and, in general, more ICR on aimsweb across grades. Following the same pattern found with CR, differences between publishers for both CR-ICR and CR- ½ ICR were greater in Grades 3 and 4 and less in Grades 5 and 6.

Paired Sample Statistics

Paired sample statistics are shown in Table 3 comparing student performance on aimsweb and DIBELS Next passages. Pearson product-moment correlations were used to determine relations between aimsweb Maze and DIBELS Next Daze metrics. A dependent-samples t-test was used to determine if metrics were significantly different when comparing performance on aimsweb and DIBELS Next passages.

Correct restorations. There were strong, positive, statistically significant correlations between aimsweb and DIBELS Next (r = .78 to .85, $p < .0005$).

Incorrect restorations. In Grades 3 and 4 there were moderate, positive, statistically significant correlations between aimsweb and DIBELS Next (r = .61 and .62, $p < .0005$). There were small, positive, statistically significant correlations between aimsweb and DIBELS Next in Grade 5 (r = .28, $p < .0005$) and Grade 6 (r = .22, p = .001).

Correct Restorations – Incorrect Restorations. There were strong, positive, statistically significant correlations between aimsweb and DIBELS Next (r = .76 to .82, $p < .0005$).

Correct Restorations – ½ Incorrect Restorations. There were strong, positive correlations between aimsweb and DIBELS Next which were statistically significant (r = .79 to .84, $p < .0005$).

Percentage at Each Recommended Tier

Given the observed differences between aimsweb and DIBELS Next on both the CR and CR- ½ ICR metrics, it was important to explore potential differences when applying publisher-provided comparison scores for making screening decisions. Figure 1 shows the percentages of students identified as being recommend to be provided with Tier I, Tier II or Tier III instruction according to publisher and across grade.

In Grades 3, 4, and 5 recommendations for providing Tier III instruction were comparable across aimsweb and DIBELS Next. In Grade 6, however, the percentage of students recommended for Tier III instruction was almost twice as high for aimsweb (23%) as it was for DIBELS Next (12%). In all grades the percentage of students recommended to be provided with Tier II instruction were higher for aimsweb (36% to 45%) than DIBELS Next (17% to 29%). Regarding recommendations for Tier I instruction, DIBELS Next was observed to have a higher percentage of students for all grades (47% to 66%) compared to aimsweb (31 to 39%).

General Prediction Metrics

We calculated two general predication metrics across grades for our study. See Table 4 for further results.

Overall Correct Classification. For all grades, DIBELS Next was observed to have a higher overall correct classification accuracy. However, using an alpha of 0.002 to adjust for multiple comparisons, no statistically significant differences were observed for Grades 3, 4, or 5. However, all grades were statistically significant using the traditional alpha level of 0.05 with Grade 4 approaching significance at the 0.01 level. Further, Grade 3 did approach significance at the 0.002 level (p = 0.003). In addition, a statistically significant difference was found in overall correct classification accuracy in Grade 6 ($p < .0005$).

Kappa. Similar to overall correct classification accuracy, differences in Kappa were in favor of DIBELS Next for all grades and were not statistically significant for Grades 3, 4, and 5 when comparing to an alpha of 0.002. Also, similar to overall correct classification accuracy, differences in Kappa approached statistical significance when using traditional alpha levels. Differences in Kappa for Grade 6 were found to be statistically significant ($p < .0005$).

Specific Prediction Metrics

We calculated a total of six specific prediction metrics across grades for our study. See Table 4 for further results. *Sensitivity*. Statistically significant differences in sensitivity were observed for Grades 3, 4, 5, and 6 ($p < .0005$) in favor of DIBELS Next.

Specificity. Statistically significant differences in specificity were observed for Grades 3, 5, and 6 when adjusting for multiple comparisons ($p < .0005$). Differences in Grade 4 approached statistical significance when compared to an alpha of 0.01 ($p = 0.015$). All differences were in favor of Aimsweb.

Positive Predictive Power. In Grades 3, 4, and 5 no statistically significant differences were observed for positive predictive power when adjusting for multiple comparisons. Differences approached statistical significance for Grade 5 in favor of aimsweb ($p = .029$) when compared to the traditional alpha of 0.05. In Grade 6 a statistically significant difference was observed in favor of DIBELS Next ($p < .0005$).

Negative Predictive Power. In Grades 3, 4, and 5 no statistically significant differences were observed for negative predictive power when adjusting for multiple comparisons. Differences approached statistical significance for Grade 3 when compared to an alpha of 0.01 ($p = 0.012$) in favor of DIBELS Next, while differences in Grade 4 were statistically significant at the traditional alpha level of 0.05 ($p = 0.03$) also in favor of DIBELS Next. In Grade 6 a statistically significant difference was observed in favor of DIBELS Next ($p < .0005$).

False Negative Rate. In Grades 3, 4, 5, and 6 the false negative rate was statistically significantly higher for aimsweb compared to DIBELS Next ($p < .0005$).

False Positive Rate. In Grades 3, 5, and 6 the false positive rate was statistically significantly higher for DIBELS Next compared to aimsweb ($p < .0005$). In Grade 4 the false positive rate was also higher for DIBELS Next (0.136) compared to aimsweb (0.062), a statistically significant difference when compared to an alpha of 0.01.

Discussion

Our study was designed to serve multiple purposes. One purpose was to examine potential differences between two different published CBM tools for maze selection on several metrics (CR, ICR, CR-ICR, CR- ½ ICR). In addition, publisher-provided comparison scores were used for aimsweb and DIBELS Next to determine the percentage of students recommended for Tier I, II, or III services. For aimsweb, these recommended "cut scores" are derived from normative data, whereas DIBELS Next provides empirically-based benchmarks which predict to future student performance. Last, differences in published guidelines for score interpretation for making screening decisions was examined.

With regard to our first purpose, student performance on aimsweb Maze and DIBELS Next Daze differed when examining a range of scoring metrics. Students in Grades 4 through 6 all obtained more CR on DIBELS Next than on aimsweb while students in Grades 3 and 4 obtained more ICR on aimsweb. Students in Grade 5 were observed to obtain more ICR on DIBELS Next than on aimsweb and students in Grade 6 obtained the same number of ICR on both measures. In addition, ICR was low for both aimsweb and DIBELS Next. This suggests that metrics that take into account ICR will do little to increase usefulness of scores. However, the inclusion of ICR will identify students who complete the maze selection task by guessing. This would suggest CR measures a slightly different construct than reading – perhaps one's ability to quickly scan words to determine whether they are appropriate given the context of the passage. One final difference was that students obtained more CR-ICR and more CR- ½ ICR on DIBELS Next than aimsweb.

For most metrics (i.e., CR, CR-ICR, and CR- ½ ICR), the correlations between aimsweb and DIBELS Next were strong and positive. However, there was a moderate, positive correlation for ICR in Grades 3 and 4. In Grades 5 and 6 there was only a small, positive correlation for ICR. This suggests that aimsweb and DIBELS Next are measuring slightly different constructs in their metric of ICR.

With regard to our second pu`rpose, it is important to note that for Grades 3, 4, and 5 the percentage of students recommended for Tier III instruction was comparable between aimsweb and DIBELS Next. In addition, aimsweb recommended a higher percentage of students for Tier II instruction and DIBELS Next recommended a higher percentage of students for Tier I only instruction. In terms of making screening decisions, this would suggest schools allocate a similar amount of resources to meet the needs of students identified as needing Tier III instruction whether using aimsweb Maze or DIBELS Next. However, the observed pattern of different recommendations for the percentages of students to be provided with Tier II or Tier I instruction across publishers is problematic. This observation is primarily problematic because aimsweb scoring guidelines would have resulted in providing more intensive resources (i.e., more than only Tier I instruction) to more students than DIBELS Next. Further, it is problematic that in Grade 6, a higher number of students were recommended to be provided with Tier III instruction using aimsweb. Given the limited nature of resources available in schools, it is not practical to provide resources to students who do not need them in order to meet expectations. Thus it is important to address our third purpose, examining differences in prediction across aimsweb and DIBELS Next.

In regard to general prediction metrics (i.e., overall correct classification accuracy and Kappa) the lack of statistical significance across aimsweb and DIBELS Next in Grades 3, 4, and 5 (p = .003 to .032) suggests no difference in using aimsweb or DIBELS Next, and their respective recommended standard scores for comparison, as a tool for assisting with screening decisions. However, statistical significance for both metrics was found in Grade 6. Further, in order to understand the observed differences between these two tools it is important to examine differences in specific (or directional) prediction metrics.

In examining sensitivity for aimsweb and DIBELS Next, the latter was observed to be more sensitive across all grades. When calculating sensitivity, only students who perform as proficient on the outcome instrument (e.g., the Iowa Assessments) are considered. That is, the metric represents the proportion of students who were proficient on the outcome instrument that were also proficient on the screening instrument. On the other hand, aimsweb was observed to have greater specificity for Grades 3, 4, and 6. When calculating specificity, only students who do not perform as proficient on the outcome instrument are considered. Thus the metric represents the proportion of students who were not proficient on the outcome instrument that were also not proficient on the screening instrument.

While one may conclude these results mean DIBELS Next does a better job at predicting students being proficient, and aimsweb does a better job at predicting students being not proficient, this is not the case for the practice of making screening decisions. This is due to the nature of the metrics in that they first consider performance on the outcome instrument. When school psychologists and other educators make screening decisions, student performance on the outcome instrument is unknown. This reality means the tools being used to predict must, in fact, predict. That is, the tools being used for making screening decisions must first consider students' performance on the screening instrument and use that performance to predict the likeliness of students meeting performance expectations on the outcome instrument. The specific prediction metrics, false positive rate (FPR) and false negative rate (FNR) are used to accomplish this task.

FPR is calculated as 1 – Specificity and is the proportion of students identified by the screening measure as proficient but who are observed to be not proficient on the outcome measure. FNR is calculated as 1 – Sensitivity and is the proportion of students identified by the screening measure to be not proficient but who are observed to be proficient on the outcome measure. Hosp (2012) highlighted that an inaccurate understanding of what is being predicted can cause confusion over this terminology. For example, when predicting a negative outcome such as at-risk status the metric FPR refers to students who are predicted to be at-risk for a negative outcome based on performance on a screening measure, but whom are observed to meet expectations on the outcome. We were interested in predicting a positive outcome (i.e., obtaining a proficient score on the Iowa Assessments from maze selection performance). As such the metric FPR reflects the proportion of students incorrectly predicted to meet expectations on the outcome.

Across all grades, DIBELS Next was observed to have a lower FNR than aimsweb. In fact, the FNR for aimsweb in Grades 3, 4, and 5 ranged from 0.551 to 0.586, meaning over half of the students identified as likely to not be proficient on the Iowa Assessments were, in fact, observed to be proficient. In Grade 6, the FNR was even higher for aimsweb at 0.768. In Grades 3 and 5 aimsweb was observed to have lower FPR than DIBELS Next, however in Grade 6 aimsweb was observed to have a higher FPR. Such an observation illustrates the importance of balancing true and false positive rates (Swets, 1992). That is, school psychologists and other educators must determine if the use of a comparison score that will result in a high true positive rate is helpful if it will simultaneously result in a high false

positive rate. In our study, a specific question to ask is whether it is worth misidentifying 55% of your students as needing additional intervention if it means you also fail to identify less than 15% of your students that do need intervention (Grades 3 and 5). This is not a simple question to answer as the resources available to schools vary and, as a result, schools vary in their ability to provide additional intervention to students potentially at-risk for not meeting expectations. We discuss this further below when highlighting practical implications of our results.

It is unclear why differences between aimsweb Maze and DIBELS Next Daze were observed. One possible reason may be variations in item presentation. That is, aimsweb Maze presents items horizontally whereas DIBELS Next Daze does so vertically. This difference might have affected student responses systematically. A second reason for differences in student performance may be an artifact of text complexity. That is, while sets of passages from each suite are equated internally, passages may not be equated across suites. A third reason could be the differences in how distractors were selected given publishers' different methods. Our results suggest the process by which distractors are selected may result in ICR measuring slightly different constructs on aimsweb than DIBELS Next. Nevertheless, we believe our results have practical implications regarding the use of maze selection for universal screening. We discuss these, but first we note study limitations.

Limitations

One limitation is the use of only one maze selection passage per publisher per student. When examining the number of CBM-R passages needed for universal screening, Mercer et al. (2012) reported the number of necessary passages varies by grade level. Regarding decision-making for individual students, having multiple passages of performance is important to account for variability across passages (i.e., students perform differently on different passages). Our study addresses this limitation to some extent as the variability of passages is reduced by our large sample size. In addition, universal screening using maze selection often uses only one passage (Good et al., 2011; Shinn & Shinn, 2002). Thus, exploring prediction accuracy with only one passage (as opposed to median WCPM from three CBM-R passages) has practical implications.

A second limitation is our participants were sampled from one Midwestern state. Although efforts were taken to ensure the sample was reflective of state demographics, it may not generalize to national demographics. This is true especially in terms of race/ethnicity given state and sample proportions of students from minority groups were lower than national proportions. Thus, despite a relatively large sample size there were not enough participants to examine differences in performance across subgroups of students (e.g., race/ethnicity, socioeconomic status, language proficiency, and disability).

A third limitation of our study may be the use of only one criterion measure. While it is important to consider the likelihood that students will perform to a desired level on states' tests of accountability in the universal screening process, it might be informative to consider additional criterion measures of reading when using metrics other than CR with aimsweb data, for example. In addition, given differences in student performance on aimsweb and DIBELS Next maze selection passages in this study, research examining differences in their relations to other criterion measures of reading is likely prudent.

Practical Implications

Despite the limitations of this study, we believe there are at least three key practical implications of our findings. One practical implication is that school psychologists must be aware that score guidelines may not be suitable for use with all measures. Historically, researchers have published growth and cut-score guidelines by grade for CBM measures (cf. Fuchs & Fuchs, 2002; Hasbrouck & Tindal, 2006; Shapiro, 2011). Further, schools have often taken the approach of developing local norms for benchmarks to assist with decision-making. These approaches have been useful in the past for discussing expectations for students, however current models using classification accuracy to determine empirically-derived benchmarks (e.g., DIBELS Next) are more sophisticated and measure-specific than other approaches.

Indeed, mounting evidence (e.g., Espin et al., 2010; Good et al., 2001; Hintze & Silberglitt, 2005; Stage & Jacobson, 2001; Wiley & Deno, 2005) suggests the use of local data can lead to increased diagnostic efficiency compared to publisher benchmarks. Further, Leblanc, Dufore, and McDougal (2012) have provided a primer for how school psychologists can use Excel to develop local comparison scores. However these studies included appropriate

calculation of various prediction metrics, far different than the common approach of schools developing local norms by determining the lowest performing 5-15% of students and targeting intervention with this subgroup. Thus, it is no longer appropriate to consider blanket expectations unless there are no empirically-derived benchmarks available for a publisher's materials.

A second practical implication of our findings is that results support using more than one type of metric for determining students' level of risk. Although we only considered student performance on maze selection using one screening tool, research has demonstrated a framework using multiple tools (i.e., CBM-R and maze selection) increases greater classification accuracy rates than the use of either tool individually (e.g., Decker, Hixson, Shaw, & Johnson, 2014). Indeed, in the area of reading it has become common practice for schools to make decisions during universal screening using both CBM-R and maze selection data (Good et al., 2011).

A third practical implication is that schools need to be aware of the relation between their universal screening process and subsequent intervention resource allocation. If schools use inappropriate comparison scores (e.g., general cut-scores or local norms), students may not be identified accurately for intervention. With resources in schools often in short supply, such a scenario should be avoided. Perhaps the most important reason for doing so is that provision of resources to students not in need reduces what is available for students who do have need. Swets (1992) suggests that a "strict criterion" be used when there is a low probability of an individual possessing the trait being predicted, resulting in few "positives" being identified as a result of performance on the screening tool. In the case that there is a high probability of an individual possessing the trait being predicted, Swets (1992) suggests a "lenient criterion." As such, many "positives" will be identified. Determining the likeliness that a student possesses the trait in question during universal screening (i.e., being a proficient reader) must include consideration of local factors (curriculum, instruction, etc.). This may contribute to the promising findings noted by those using local data to arrive at comparison scores which predict to future outcomes (e.g., Leblanc et al. 2012).

In addition, issues with balancing true and false positives also highlights the benefits of multiple gate screening practices (see Gilbert, Compton, Fuchs, & Fuchs, 2012). Such practices acknowledge students will be misidentified as needing additional intervention. However, by providing such instruction, and monitoring students' response, it is possible for school psychologists and other educators to identify the students who were misidentified in the screening process. Even so, the fact remains that for some period of time (i.e., several weeks) additional resources will be allocated to, potentially a significant number of, students who were not in need of such intervention. Not only does this represent a possible barrier to providing instruction to the students misidentified, it also results in a reduction of the resources available to those students who truly need them. Thus the issue with determining the degree of "strictness" for comparison scores when making screening decisions remains.

We believe such an issue is best left to the professional judgement of school psychologists and other educators applying their expertise at the local level. However, such judgement must be used in the context of data-based decision-making. Deno (2016) discusses the benefits of using a statistical approach for identifying students in need of additional intervention instead of professional judgement. Such an approach is reflected in the use of cut-scores in CBM tools for making screening decisions, and one we agree with. Further, Deno (2016) states, "Carrying out system-level directives does not change the fact that we are accepting the values and the decision frameworks designed by others and, in so doing, we (are) complicit in making the decisions" (p. 15). This means it is paramount for school psychologists and other educators to be aware of the nature of the cut-scores they are using (i.e., normative-based or empirically derived) and, more importantly, be cognizant of the potential repercussions of using such scores for instructional decision-making. By considering such things, school psychologists can potentially facilitate conversations with other educators about how to allocate school resources. For example, the problem-solving team at school A (which has a well-resourced intervention system in place) may suggest using a lenient criterion for establishing a cut-score and risk identifying a significant number of false negatives. Doing so would greatly increase the likeliness that all students needing assistance would receive additional intervention, however many students could also be provided with assistance that they do not need to meet expectations (an unnecessary allocation of resources). However, the school psychologist can point out the availability of the school's well-resourced intervention system to the problem-solving team as a means to implement additional intervention and determine whether it is really needed for specific students (i.e., multiple gating procedures). The school psychologist can further point out that while additional resources will need to be allocated in the short term, it may not ultimately be necessary to continue to do so and the temporary use of additional resources is an appropriate instructional decision given the resources available to them.

However, the school psychologist at school B may find it appropriate to have a different discussion with the problem-solving team if the school's intervention system is not as well resourced. Specifically, the school psychologist may suggest the problem-solving team use a stricter criterion for establishing a cut-score. Doing so would mean increasing the likeliness that only the students receiving the limited resources available truly needed them to meet expectations, at the risk of a few students in need of assistance not receiving intervention. Given the resources available at school B this may also be an appropriate instructional decision.

Of course, any cut-score used should make sure to identify the students most in need. This may be conceptualized as students needing the allocation of Tier III resources. As our study's results suggest using aimsweb Maze or DIBELS Next Daze for screening decisions will identify this group of students similarly. However, questions about who to allocate Tier II resources to are more complex and require knowledge of local resources, and thus, professional judgment. The professional judgment of school psychologist's can potentially be exceptionally valuable for discussing issues related to cut-score determination when schools are in the process of selecting a specific CBM publisher or in attempting to develop local, empirically-derived cut-scores.

Conclusion

Universal screening involving maze selection can be useful for identifying students in need of additional resources to meet desired reading outcomes. However, a range of procedural choices and implications must be considered. In particular, school psychologists must be aware that scoring practices are not all equal and often result in different conclusions. These conclusions could result in identifying students unnecessarily for targeted resources (e.g., intervention, time with personnel, specific materials). Results here suggest that school psychologists should encourage the use of empirically-derived benchmarks in schools practices or, possibly, undertake the development of local benchmarks that involve calculation of predication metrics. Further, school psychologists must be mindful to consider more than one source of data when making decisions about students' instructional needs.

References

Ardoin, S., & Christ, T. (2009). Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, Aimsweb, and an experimental passage set. *School Psychology Review, 38*, 266-283.

Ardoin, S., Witt, J., Suldo, S., Connell, J., Koenig, J., Resetar, J., et al. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probe when conducting universal screening. *School Psychology Review, 33*, 218-233.

Brown-Chidsey, R., Davis, L., & Maya, C. (2003). Sources of variance in curriculum-based measures of silent reading. *Psychology in the Schools, 40*, 363-377.

Decker, D., Hixson, M., Shaw, A., & Johnson, G. (2014). Classification accuracy of oral reading fluency and maze in predicting performance on large-scale assessments. *Psychology in the Schools*, 51, 625-635.

Deno, S. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232.

Deno, S. L. (2016). Data-based decision-making. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.). *Handbook of Response to Intervention* (2nd ed.). New York: Springer.

Deno, S., Anderson, A., Calender, S., Lembke, E., Zorka, H., & Casey, A. (2002, February). *Developing a school-wide model for progress monitoring: A case example and empirical analysis.* Symposium at the annual meeting of the National Association of School Psychologists, Chicago, IL.

Espin, C., Wallace, T., Lembke, E., Campbell, H., & Long, J. (2010). Creating a progress-monitoring system in reading for middle-school students: Tracking progress toward meeting high-stakes standards. *Learning Disabilities Research & Practice, 25*, 60-75.

Ford, J., & Hosp. J. (2015). Classroom management for universal screening. In W. G. Scarlett (Ed.), *Classroom management: An A-to-Z Guide*. Thousand Oaks, CA: Sage Publishing.

Fuchs, L., & Fuchs, D. (2002). Curriculum-based measurement: Describing competence, enhancing outcomes, evaluating treatment effects, and identifying treatment nonresponders. *Peabody Journal of Education, 77*, 64-84.

Fuchs, L., & Fuchs, D. (1992). Identifying a measure for monitoring students' reading progress. *School Psychology Review, 21*, 45-58.

Gilbert, J. K., Compton, D. L., Fuchs, D., & Fuchs, L. S. (2012). Early screening for risk of reading disabilities recommendations for a four-step screening system. *Assessment for effective intervention*, *38*, 6-14.

Good, R., Simmons, D., & Kame'enui, E. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, *5*(3), 257-288.

Good, R., Kaminski, R., Dewey, E., Wallin, J., Powell-Smith, K., & Latimer, R. (2013). *DIBELS Next Technical Manual.* Eugene, OR: Dynamic Measurement Group.

Good, R., Kaminski, R., Dewey, E., Wallin, J., Powell-Smith, K., & Latimer, R. (2011). *DIBELS Next technical manual.* Eugene, OR: Dynamic Measurement Group.

Graney, S., Martínez, R., Missall, K., & Aricak, O. (2010). Universal screening of reading in late elementary school R-CBM versus CBM Maze. *Remedial and Special Education*, *31*, 368-377.

Hasbrouck, J., & Tindal, G. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher*, *59*, 636-644.

Hintze, J., & Shapiro, E. (1997). Curriculum-based measurement and literature-based reading: Is curriculum based measurement meeting the needs of changing reading curricula? *Journal of School Psychology*, *35*, 351-375.

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review*, *34*(3), 372.

Hoover, H., Dunbar, S., & Frisbie, D. (2001). *Iowa tests of basic skills (ITBS) forms A, B, and C.* Rolling Meadows, IL: Riverside Publishing Company.

Hosp, J. (2012). Using assessment data to make decisions about teaching and learning. In K. Harris, T. Urdan, & S. Graham (Eds.). *The APA handbook of educational psychology*. Washington, DC: American Psychological Association.

Hosp, J., & Ardoin, S. (2008). Assessment for instructional planning. *Assessment for Effective Intervention*, *33*, 69-77.

Hosp, M., Hosp, J., & Howell, K. (2016). *The ABCs of CBM: A practical guide to curriculum-based measurement* (2nd ed.). New York: Guilford Press.

Ikeda, M., Neesen, E., & Witt, J. (2008). Best practices in universal screening. In A. Thomas, & J. Grimes (Eds.), *Best practices in school psychology* (pp. 103-114). Bethesda, MD: National Association of School Psychologists.

January, S., & Ardoin, S. (2012). The impact of context and word type on students' maze task accuracy. *School Psychology Review*, *41*, 262-271.

Jenkins, J., & Jewell, M. (1992). An examination of the concurrent validity of the Basic Academic Skills Samples (BASS). *Assessment for Effective Intervention*, *17*, 273-288.

Jenkins, J., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children*, *59*, 421-432.

Kendeou, P., Papadopoulos T., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction, 22*, 354-367.

Leblanc, M., Dufore, E., & McDougal, J. (2012). Using General Outcome Measures to Predict Student Performance on State-Mandated Assessments: An Applied Approach for Establishing Predictive Cutscores. *Journal of Applied School Psychology*, *28*(1), 1-13.

Mercer, S., Dufrene, B., Zoder-Martell, K., Harpole, L., Mitchell, R., & Blaze, J. (2012). Generalizability theory analysis of CBM maze reliability in third-through fifth-grade students. *Assessment for Effective Intervention*, *37*, 183-190.

Neddenriep, C., Hale, A., Skinner, C., Hawkins, R., & Winn, B. (2007). A preliminary investigation of the concurrent validity of reading comprehension rate: A direct of reading comprehension. *Psychology in the Schools, 44*, 373-388.

Parker, R., Hasbrouck, J., & Tindal, G. (1992). The maze as a classroom-based reading measure: Construction methods, reliability, and validity. *The Journal of Special Education*, *26*, 195-218.

Pearson (2011). Aimsweb Default Cut Scores Explained. *State Prediction User's Guide*.

Pierce, R., McMaster, K., & Deno, S. (2010). The effects of using different procedures to score maze measures. *Learning Disabilities Research & Practice, 25*, 151-160.

Potter, M., & Wamre, H. (1990). Curriculum-based measurement and developmental reading models: Opportunities for cross validation. *Exceptional Children, 57*, 16–25.

Rowe, S. S., Witmer, S., Cook, E., & daCruz, K. (2014). Teachers' attitudes about using curriculum-based measurement in reading (CBM-R) for universal screening and progress monitoring. *Journal of Applied School Psychology*, *30*, 305-337.

Skinner, C., Neddenriep, C., Bradley-Klug, K., & Ziemann, J. (2002). Advances in curriculum-based measurement: Alternative rate measures for assessing reading skills in pre-and advanced readers. *Behavior Analyst Today*, *3*, 270-281.

Shapiro, E. S. (2011). *Academic skills problems: Direct assessment and intervention*. Guilford Press.

Shin, J., Deno, S., & Espin, C. (2000). Technical adequacy of CBM-SR task for curriculum-based measurement of reading growth. *Journal of Special Education, 34*, 164-173.

Shinn, M., & Shinn, M. (2002). *Aimsweb® training workbook.* Eden Prairie, MN: Edformation.

Silberglitt, B., Burns, M., Madyun, N., & Lail, K. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in the Schools, 43*, 527-535.

Sprinthall, R. (2003). *Basic statistical analysis* (7th ed.). New York: Pearson.

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*(3), 407.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47*, 522-532.

Wayman, M., Wallace, T., Wiley, H., Tichá, R., & Espin, C. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education, 41*, 85-120.

Williams, K. (2001). *GRADE: Group reading assessment and diagnostic evaluation*. Circle Pines, MN: American Guidance Service.

Wiley, H., & Deno, S. (2005). Read aloud and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education, 26*, 207-214.

Woodcock, R., McGrew, K., & Mather, N. (2001). *Woodcock-Johnson tests of achievement*. Itasca, IL: Riverside.

Table 1

Comparison of Demographics Across State, Project Sample, and Study Sample

Group	State		Project Sample		Study Sample	
	n	%	*n*	%	*n*	%
Total	487,559	100.0%	1,685	100.0%	925	100%
Female	234,737	48.1%	853	50.1%	479	51.8%
American Indian	2,784	0.6%	22	1.3%	11	1.0%
Asian	10,543	2.2%	15	1.0%	9	1.0%
Black	28,317	5.8%	50	3.0%	30	3.0%
Hispanic	33,974	7.0%	122	7.2%	42	5.0%
White	411,941	84.5%	1575	83.5%	800	86.6%
Multiracial	n/a	n/a	37	2.2%	23	3.0%
Economic Disadvantage	165,830	34.0%	734	43.6%	366	39.6%
Limited English Proficiency	20,334	4.2%	69	4.1%	32	3.5%
Students with Identified Disabilities	67,065	13.8%	220	13.1%	125	14.0%

Table 2

Descriptive Statistics for aimsweb Maze and DIBELS Next Daze for Correct Restorations, Incorrect Restorations, Correct Restorations-Incorrect Restorations, and Correct Restorations-1/2 Incorrect Restorations

Metric		Grade 3 (N = 228)		Grade 4 (N = 219)		Grade 5 (N = 252)		Grade 6 (N = 226)	
		aims- web	DIBELS Next	aims- web	DIBELS Next	aims- web	DIBELS Next	aims- web	DIBELS Next
CR	Mean	12.52	20.61	16.48	23.76	21.94	24.39	23.74	25.36
	SD	5.86	7.89	7.13	8.85	7.86	8.30	8.15	8.52
	SE	0.39	0.52	0.48	0.60	0.50	0.50	0.54	0.57
ICR	Mean	2.15	1.48	2.20	1.28	1.24	2.13	1.58	1.58
	SD	2.68	2.52	2.69	3.44	1.72	2.27	1.45	1.35
	SE	0.18	0.17	0.18	0.23	0.11	0.11	0.10	0.23
CR-ICR	Mean	10.37	19.13	14.28	22.48	20.71	22.25	22.19	23.79
	SD	6.71	8.99	8.22	10.18	8.47	8.55	8.56	8.77
	SE	0.44	0.60	0.56	0.69	0.53	0.53	0.57	0.58
CR-1/2ICR	Mean	11.47	19.88	15.40	23.15	21.33	23.32	22.98	24.58
	SD	6.12	8.36	7.54	9.31	8.13	8.19	8.33	8.62
	SE	0.41	0.55	0.51	0.63	0.41	0.52	0.55	0.57

Note. DIBELS = Dynamic Indicators of Basic Early Literacy Skills, Next; CR = correct restorations; ICR = incorrect restorations; SD = standard deviation; SE = standard error.

Table 3

Paired Samples Correlations and T-Tests

Grade	Metric	Correlation	Sig.	Mean Difference	*SD*	*SE*	95% CI Lower	95% CI Upper	*t*	*df*	Sig.
3	CR	.781	<.0005	-8.09	4.94	.33	-8.74	-7.45	-24.740	227	<.0005
	ICR	.608	<.0005	.67	2.31	.15	.37	.97	4.395	227	<.0005
	CR-ICR	.788	<.0005	-8.76	5.55	.37	-9.49	-8.04	-23.850	227	<.0005
	CR-½ ICR	.791	<.0005	-8.41	5.13	.34	-9.08	-7.74	-24.737	227	<.0005
4	CR	.831	<.0005	-7.28	4.93	.33	-7.94	-6.62	-21.837	218	<.0005
	ICR	.615	<.0005	.920	2.77	.19	.55	1.29	4.929	218	<.0005
	CR-ICR	.810	<.0005	-8.20	5.96	.40	-8.91	-7.41	-20.348	218	<.0005
	CR-½ ICR	.826	<.0005	-7.75	5.25	.36	-8.45	-7.05	-21.852	218	<.0005
5	CR	.800	<.0005	-2.44	5.02	.32	-3.07	-1.82	-7.726	251	<.0005
	ICR	.277	<.0005	-.90	2.72	.17	-1.23	-.55	-5.242	251	<.0005
	CR-ICR	.760	<.0005	-1.55	5.90	.37	-2.28	-.82	-4.162	251	<.0005
	CR-½ ICR	.788	<.0005	-2.00	5.31	.34	-2.66	-1.34	-5.968	251	<.0005
6	CR	.848	<.0005	-1.59	4.61	.31	-2.19	-.99	-5.183	225	<.0005
	ICR	.216	.001	.10	1.76	.12	-.22	.24	.076	225	.940
	CR-ICR	.818	<.0005	-1.60	5.24	.35	-2.28	-.91	-4.586	225	<.0005
	CR-½ ICR	.837	<.0005	-1.59	4.85	.33	-2.23	-.96	-4.934	225	<.0005

Note. CR = correct restorations; ICR = incorrect restorations; SD = standard deviation; SE = standard error mean; CI – confidence interval; df = degrees of freedom; Sig. = statistical significance value.

Table 4

Comparison of Prediction Metrics for aimsweb and DIBELS Next Maze Selection CBM Tools

	Grade 3			Grade 4			Grade 5			Grade 6		
Prediction Metric	aims-web	DIBELS Next	*p*	aims-web	DIBELS Next	*p*	aims-web	DIBELS Next	*p*	aims-web	DIBELS Next	*p*
General												
OCC	0.579	0.711	0.003	0.630	0.740	0.013	0.583	0.675	0.032	**0.292**	**0.721**	**<.0005**
Kappa	0.557	0.694	0.003	0.611	0.723	0.013	0.530	0.656	0.004	**0.251**	**0.706**	**<.0005**
Specific												**<.0005**
Sens.	**0.448**	**0.727**	**<.0005**	**0.449**	**0.667**	**< .001**	**0.414**	**0.615**	**<.0005**	**0.232**	**0.803**	**<.0005**
Spec.	**0.851**	**0.676**	**<.0005**	0.938	0.864	0.015	**0.928**	**0.795**	**<.0005**	**0.393**	**0.583**	**<.0005**
PPP	0.863	0.824	0.250	0.925	0.893	0.246	0.921	0.860	0.029	**0.393**	**0.765**	**<.0005**
NPP	0.426	0.544	0.012	0.500	0.603	0.030	0.438	0.504	0.139	**0.232**	**0.636**	**<.0005**
FNR	**0.552**	**0.273**	**<.0005**	**0.551**	**0.333**	**<.0005**	**0.586**	**0.385**	**<.0005**	**0.768**	**0.197**	**<.0005**
FPR	**0.149**	**0.324**	**<.0005**	0.062	0.136	0.010	**0.072**	**0.205**	**<.0005**	**0.607**	**0.417**	**<.0005**

Note. OCC = Overall Correct Classification Accuracy; Sens. = Sensitivity; Spec. = Specificity; PPP = Positive Predictive Power; NPP = Negative Predictive Power; FNR = False Negative Rate; FPR = False Positive Rate.

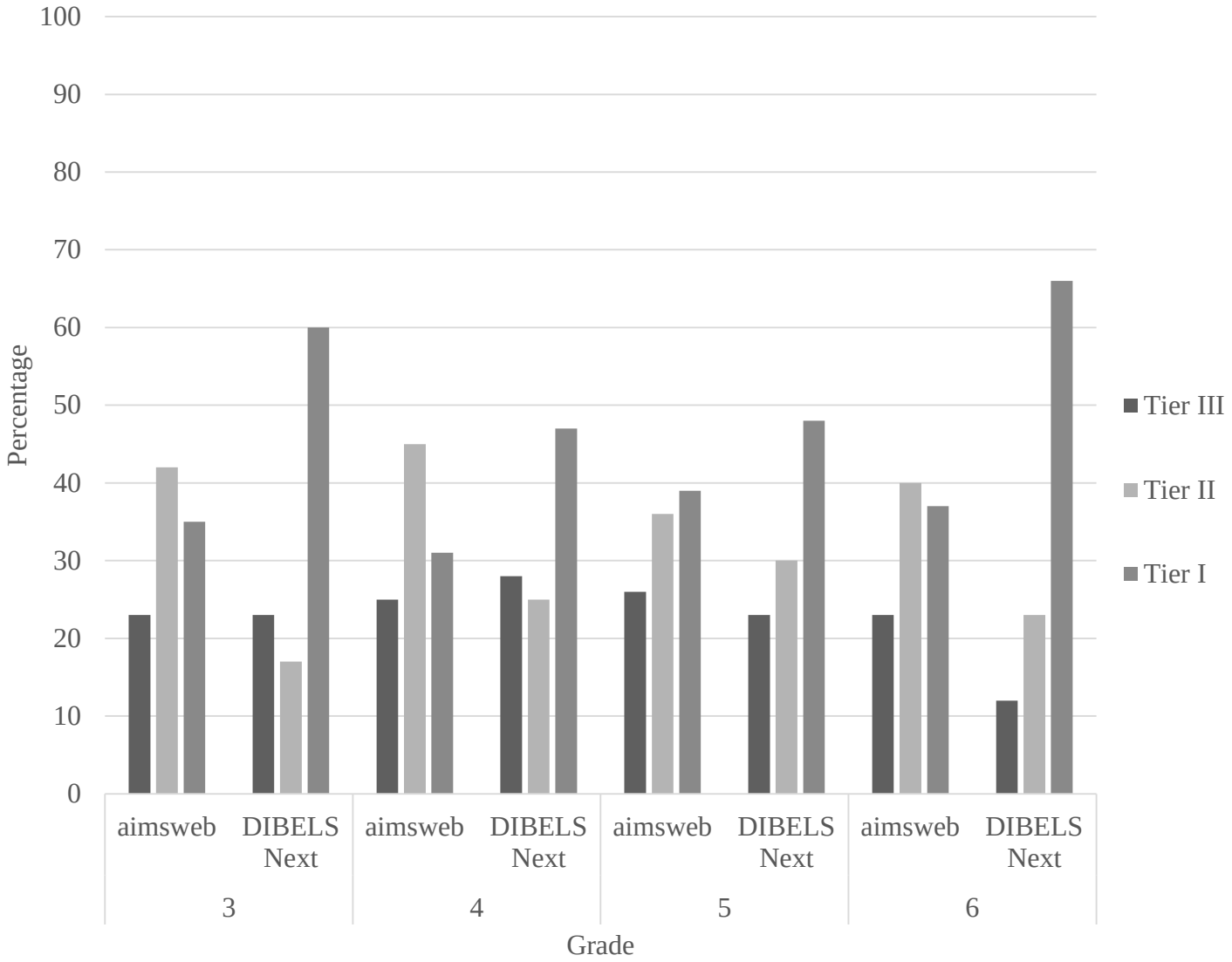

Figure 1. Percentage of Students at Comparison Scores across Publishers by Recommended Tier.