Article

Hammill Institute on Disabilities

Assessment for Effective Intervention
2018, Vol. 43(3) 131–143
© Hammill Institute on Disabilities 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1534508417730823
aei.sagepub.com
SAGE

Classification Accuracy of Curriculum-Based Measures for Beginning Writers in First Grade

Pyung-Gang Jung, PhD[1] and Kristen L. McMaster, PhD[2]

Abstract

We examined the classification accuracy of Curriculum-Based Measurement in writing (CBM-W) Picture Word prompts scored for words written (WW), words spelled correctly (WSC), and correct word sequences (CWS). First graders (n = 133) were administered CBM-W prompts and the Test of Written Language–Third Edition (TOWL-3; Hammill & Larsen, 1996). Prompts scored for WSC showed acceptable levels of sensitivity (.947) and specificity (.587) with the TOWL-3 Contextual Language. Positive predictive values were low (approximately .20 to .30), and negative predictive values were high (mostly above .95). Overall classification accuracy, represented by the area under curve (AUC), ranged from .727 to .831. Further research regarding ways to improve classification accuracy of CBM-W and preliminary implications for practice are discussed.

Keywords

curriculum-based measures, beginning writers, classification accuracy, ROC

Proficient writing is important for students to be successful in school and later life. Most academic subjects require students to synthesize information or knowledge through written work. Postsecondary education and the workplace also call for written responses or products, such as to evaluate the qualifications of applicants (Graham, 2008). In spite of the importance of writing, recent reports from the National Assessment of Educational Progress (NAEP; 2012) indicate that a large percentage of students (74% of eighth graders and 73% of 12th graders) have not reached proficient writing levels. These results highlight the need to identify students who are at risk in writing and provide early intervention (Berninger, Nielson, Abbott, Wijsman, & Raskind, 2008).

To provide effective early writing intervention, it is essential to identify skills relevant to improve writing for young students. One widely accepted model of writing (Hayes & Flower, 1980) specifies three components: planning (generating, organizing, and goal setting), translating (text generation and transcription), and reviewing (evaluating and revising). This model reflects the process of skilled writers; however, it does not completely reflect the writing development process of young children (McCutchen, 2006). Thus, researchers have proposed a modified version of the Hayes and Flower (1980) model—the *Simple View of Writing*—to describe young students' writing development (Berninger & Amtmann, 2003).

According to the Simple View of Writing, writing consists of three components: transcription, text generation, and self-regulation. Each component is constrained by attention and memory (Berninger & Winn, 2006). Transcription skills have been shown to play a critical role in early writing development (Graham & Harris, 2000), because proficient transcription skills free up cognitive resources needed for higher-level writing skills (such as generating ideas, selecting appropriate words, constructing strong sentences, and so on). The Simple View of Writing has been supported by empirical studies examining the effects of transcription interventions to improve text generation skills (see McMaster, Kunkel, Shin, Jung, & Lembke, 2017).

To provide timely early writing intervention, it is essential to identify struggling writers as early as possible. Response to Intervention (RTI)—a preventative, multitiered system of support—prompts early identification and intervention for students at risk of academic failure (Fuchs & Fuchs, 2007). In general, RTI consists of (a) universal screening, (b) research-based instruction, (c) progress monitoring, and (d) increasingly intense levels of intervention. As part of universal screening, it is critical to *accurately* identify students at risk, because under- or over-identification

[1]Ewha Womans University, Seoul, Republic of Korea
[2]University of Minnesota, Minneapolis, USA

Corresponding Author:
Pyung-Gang Jung, Ewha Womans University, 534-3 Education Building A, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Republic of Korea.
Email: jungx165@gmail.com

can lead to either delaying remediation of difficulties through intervention or wasting educational resources. Use of screening measures with adequate technical characteristics might help schools accurately identify students at risk by minimizing such errors.

Curriculum-Based Measurement (CBM)

CBM is a set of brief measures to indicate students' overall academic performance in basic reading, mathematics, spelling, and written expression (Deno, 1985, 2003). For decades, CBM has been used for a variety of academic purposes due to its unique characteristics of standardized administration and scoring procedures, adequate technical features, possibilities of using multiple alternate forms, and time efficiency (Fuchs, 2004). These academic purposes include screening, progress monitoring, and evaluating teachers' instructional programs. CBM is unique in that it is a form of *general outcome measurement* (Fuchs & Deno, 1991)—it is designed to measure essential outcomes that students are expected to achieve by the end of a school year, rather than mastery of a series of subskills. Thus, CBM scores represent an index of a student's overall proficiency in academic areas, and is well suited for monitoring student progress and evaluating instructional effectiveness based on student performance.

CBM in Writing (CBM-W)

CBM-W research was initiated with CBM-W Story prompts, a task designed to capture students' overall writing proficiency by prompting a narrative or informational essay for 3 min in response to a given topic. CBM-W Story prompts scored using a variety of quantitative indices have shown moderate to high criterion validity for upper-grade elementary students (r = .41 to .88; Deno, Mirkin, & Marston, 1980; Videen, Deno, & Marston, 1982) but weak to moderate criterion validity with primary-grade students (r = .34 to .67; Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006; Jewell & Malecki, 2005; Parker, Tindal, & Hasbrouck, 1991), suggesting that Story prompts are less valid for assessing beginning writers than for students in higher grades. Thus, more recently, researchers have developed alternative CBM-W approaches for young students based on the Simple View of Writing (Berninger & Amtmann, 2003).

CBM-W tasks have been developed at the subword, word, sentence, and passage levels to measure students' beginning writing skills (see McMaster et al., 2011 for a review). Tasks include Letter Writing, Sound Spelling, Word Copying, Word Dictation, Real Word Spelling, Nonsense Word Spelling, and Letter prompts (designed to tap transcription skills; McMaster, Du, & Pétursdóttir, 2009; Lembke, Deno, & Hall, 2003; Ritchey, 2006); and Sentence Copying, Sentence Dictation, Sentence Writing, Picture Word, Picture Story, Picture Theme, and Photo prompts (designed to tap transcription and text generation skills, McMaster et al., 2009; McMaster et al., 2011; Coker & Ritchey, 2010; Lembke et al., 2003; Ritchey & Coker, 2013). Students' responses are scored quantitatively, including the number of words written (WW), words spelled correctly (WSC), correct letter or word sequences (CLS/CWS), and correct minus incorrect letter or word sequences (CILS/CIWS). CLS/CWS are defined as any adjacent, correctly spelled letters/words that are acceptable within the context of the sample to a native English speaker (Videen et al., 1982).

McMaster et al. (2009) proposed that, for CBM-W writing tasks to be considered to have "sufficient" evidence of technical adequacy, they should have evidence of a minimum of r = .70 for reliability, and r = .50 for criterion validity. Tasks that have met these criteria include Letter Writing and Sound Spelling (for kindergartners, r = .82 to .94 for split-half and internal consistency reliability, r = .53 to .77 for criterion validity with norm-referenced assessments; Ritchey, 2006), Word Dictation and Sentence Dictation (for first to second graders, r = .75 to .95 for alternate-form reliability, r = .76 to .92 for criterion validity with atomistic variables; Lembke et al., 2003), and Picture Word and Sentence Copying (for first graders; r = .70 to .93 for alternate-form reliability, r = .51 to .61 with a norm-referenced assessment, teacher ratings, and a district rubric, McMaster et al., 2009; McMaster et al., 2011).

For this study, we selected Picture Word to examine its utility as a screening measure for early identification. Picture Word meets criteria for "sufficient" technical adequacy and also can be group administered (unlike Letter Writing, Sound Spelling, and Word and Sentence Dictation, which must be administered individually). Group administration saves time and human resources and thus is more efficient for the purpose of screening. Picture Word captures foundational sentence-level writing skills that students need to meet first-grade writing standards (e.g., Common Core State Standards; National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). Further, teachers have indicated a preference for Picture Word over Sentence Copying, citing that it represents a more authentic writing task (McMaster et al., 2009). Beyond the technical features of static scores, Picture Word has shown to be sensitive to growth over time (McMaster et al., 2011), indicating its potential for monitoring student progress. Evidence of the utility of Picture Word for screening would further support its use as a tool for identification and progress monitoring in a comprehensive RTI model.

Classification Accuracy of CBM-W

Criterion validity of CBM-W provides preliminary support for assessing students' writing skills, but correlational

evidence is not sufficient (VanDerHeyden, 2011). Rather, criterion validity studies help identify potential screening tools, and "classification studies are the *sine qua non* of screening research" (Jenkins, Hudson, & Johnson, 2007, p. 6). Research showing the extent to which CBM-W accurately classifies students as at risk or not at risk in writing would provide powerful evidence of its utility for screening beyond criterion validity evidence.

Several indices can be used to examine classification accuracy, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the curve (AUC; Swets, Dawes, & Monahan, 2000). Sensitivity and specificity represent the inherent properties of screening measures to distinguish between students at risk and those who are not at risk based on an established criterion. Sensitivity refers to the percentage of students who are truly at risk among students identified as at risk by a screening measure. Specificity refers to the percentage of students who are truly not at risk among those identified as not at risk by the screen. Thus, sensitivities and specificities allow practitioners to examine how accurately each measure discriminates between alternative states of risk. However, the information does not provide useful information in regard to interpreting results to make clinical decisions for individual students (VanDerHeyden, 2011; Zweig & Campbell, 1993).

PPV and NPV, which represent the efficiency of the measure, provide helpful information for practitioners. PPV refers to the probability that students identified as at risk by screening measures actually need additional intervention. NPV refers to the probability that students identified as not at risk by screening actually do not need additional intervention (VanDerHeyden, 2011). The AUC indicates the overall classification accuracy of the measure.

A small body of research has examined the classification accuracy of CBM-W for beginning writers. Ritchey and Coker (2013) examined the extent to which two CBM-W tasks (Picture Story and Story Starter) accurately identified risk status using quantitative and qualitative scoring procedures for second- and third-grade students with the Woodcock-Johnson Tests of Achievement–Third Edition (WJ3; Woodcock, McGrew, Schrank, & Mather, 2001/2007) or teacher ratings. Sensitivity varied by grades, ranging from .25 to 1.00; most specificity was between approximately .70 and .80 for both tasks. The AUC ranged from .57 to .85 for combined grades across writing tasks and scoring procedures. These findings provide preliminary classification accuracy evidence of CBM-W for early elementary students. In this study, however, over half of the sensitivity values did not meet the criterion of .90 recommended for screening measures by Jenkins et al. (2007), because the authors selected cut-off scores without holding sensitivity at a certain level to explore classification accuracy.

The authors addressed this issue in two additional studies by examining classification accuracy holding sensitivity at .90, and by examining classification accuracy of CBM-W and fluency-based reading tasks separately and in combination. Coker and Ritchey (2014) used four CBM-W tasks (Letter Writing, Sound Spelling, Word Spelling, and Sentence Writing) and three reading tasks for Kindergarten students. The overall accuracy (AUC) of individual writing tasks ranged from .60 to .73 with the Test of Early Written Language, 2nd Edition (TEWL-2; Hresko, Herron, & Peak, 1996) and .79 to .87 with teacher ratings. Overall accuracy was higher when they used combined tasks for screening than when they used individual writing tasks, indicating the value of adding reading measures to strengthen classification accuracy.

Ritchey and Coker (2014) conducted a similar study with first graders. They used three CBM-W tasks (Spelling, Sentence Writing, and Picture Story Writing) and three reading tasks for screening. Consistent with Coker and Ritchey's (2014) findings, in general, combined writing tasks showed higher overall accuracy than individual writing tasks, and combined reading and writing tasks boosted the overall accuracy level. The AUC for individual and combined writing tasks ranged from .72 to .82 with the WJ3 (Woodcock et al., 2001, 2007), and .87 to .91 with teacher ratings, providing evidence that these three CBM-W tasks can be used for screening.

The existing evidence indicates that CBM-W can accurately identify struggling beginning writers, and that combined writing tasks or additional measures can improve overall accuracy. In previous studies, classification accuracy characteristics of several CBM-W tasks were examined. However, no research has yet been conducted on the CBM-W Picture Word task, which has evidence of sufficient alternate-form reliability and criterion validity (McMaster et al., 2009), as well as evidence of sensitivity to growth (McMaster et al., 2011). Classification accuracy evidence would add support for the use of the Picture Word task as a flexible tool for identification and progress monitoring within an RTI framework. Thus, the present study extends the literature by examining the classification accuracy of the Picture Word task for first-grade students.

The purpose of this study was to examine classification accuracy of the CBM-W Picture Word task to determine its utility for screening. We examined classification accuracy characteristics of three different scoring procedures (WW, WSC, and CWS) in terms of sensitivity, specificity, PPV, NPV, and overall classification accuracy for first graders.

Method

To examine the classification accuracy of CBM-W, data were drawn from two previous studies (McMaster et al., 2009; McMaster et al., 2011). McMaster et al. (2009) administered

seven CBM-W tasks for beginning writers that prompted copying or generation of words, sentences, or stories in response to written or picture prompts. Among these tasks, Sentence Copying, Picture Word, Photo, and Story prompts showed evidence of reliability (rs > .70) and criterion validity (rs > .50) for some scoring metrics. McMaster et al. (2011) examined technical features of slopes from three of these tasks (Picture Word, Sentence Copying, Story prompts). Among the three, Picture Word appeared most sensitive to growth. Whereas the previous studies focused on reliability and criterion validity of CBM-W scores and their capacity to show growth, in this study, we focused on classification accuracy of Picture Word beyond the criterion validity evidence of the measure.

Although the two previous studies were conducted in different school years, data were merged for this study, because they were collected in the same district using the same measures, administration, and scoring procedures. In both studies, each student was administered the same Picture Word prompts and the Test of Written Language–Third Edition (TOWL-3; Hammill & Larsen, 1996) at approximately the same time of year. For this study, we used CBM-W scores collected in February of both study years, and TOWL-3 scores collected in May of both years.

Setting and Participants

Participants were from three elementary schools in a large Midwestern urban district. School 1 served 608 students in kindergarten through fourth grade; 43% were from culturally or linguistically diverse backgrounds, 19% received free or reduced lunch, 6% received special education, and 7% were English Language Learners (ELLs). School 2 served 281 students in kindergarten through fifth grades; 46.6% were White, 27.1% were African American, 19.6% were Hispanic, 2.5% were Asian, and 4.3% were American Indian. Approximately 48.8% received free or reduced lunch, 11.7% received special education service, and 18.2% were ELLs. School 3 served 932 students in kindergarten through eighth grades; 23% were White, 34.8% African American, 37.3% Hispanic, 3.3% Asian, and 1.6% was American Indian. Sixty-two percent received free or reduced lunch, 15.9% received special education, and 31.6% were ELLs.

McMaster et al. (2009) included 48 first-grade students from two classrooms in School 1, and McMaster et al. (2011) included 85 first-graders from five classrooms in Schools 2 and 3. Participating student demographic information by school is presented in Table 1. Demographic information was missing for one student. The mean age was 6.56 years (SD = 0.59, range = 4.87 years to 7.78 years).

Measures

CBM-W task. Picture Word prompts were designed to capture sentence-level writing performance by prompting students to generate sentences using the words provided. Each prompt consists of words with a picture above each word (e.g., apple, hat, teacher) in case of reading difficulties (see Figure 1). The researchers selected 45 high-frequency words from the Houghton Mifflin curriculum (Cooper & Pikulski, 2005), which was used in the district at the time of the first study. The researchers found pictures from Microsoft ClipArt or drew them by hand (McMaster et al., 2009). Before the task began, the examiner drew a picture on the board (e.g., tree) and wrote the name of the object underneath. Then, the examiner asked students to generate a sentence using the word and wrote sentence examples on the board. After this practice, the examiner instructed students to write as many sentences as possible with the words in Picture Word prompts. After 3 min, students stopped writing, raised their pencils in the air (to show they had stopped), and then circled the last letter they wrote.

Across the two previous studies, common scoring procedures were WW, WSC, and CWS. CIWS was scored in McMaster et al. (2009) and found to have low reliability coefficients, and thus was not used in subsequent work. Alternate-form reliabilities of CBM-W Picture Word prompts scored have ranged from r = .59 to .79 for WW, .61 to .76 for WSC, and .58 to .77 for CWS. Criterion validities ranged from r = .55 to .60 with teacher ratings, r = .37 to .54 with a district rubric, and r = .23 to .54 with the TOWL-3 (McMaster et al., 2009; McMaster et al., 2011).

Criterion measure. The Test of Written Language, Third Edition (TOWL-3; Hammill & Larsen, 1996) was selected for use in the original two studies to provide criterion validity evidence for CBM-W as a measure of overall writing proficiency. Because there is no agreed-on gold standard writing proficiency measure, the TOWL-3 was selected in consultation with an expert in writing research at the time of the previous studies. In addition, according to the manual, the TOWL-3 "can be used to identify students who perform significantly more poorly than their peers in writing and who as a result need special help" (Hammill & Larsen, 1996, p.6), indicating that the measure can be used for screening. Also, the TOWL-3 can be group administered, making it an efficient option for schools and research.

The TOWL-3 is a comprehensive test of written language designed for students 7 years to 17 years 11 months of age. The Spontaneous Writing subtest was group administered to all students in May. Students were shown a picture depicting a futuristic scene of astronauts, spaceships, and construction activity, and told to think of a story related to the picture. Students were encouraged to plan their story and write as much as they could for 15 min. Writing samples were scored based on three categories: Contextual Conventions (including capitalization, punctuation, and spelling), Contextual Language (including quality of vocabulary, sentence construction, and grammar), and Story

Table 1. Student Demographics.

	School 1 (*n* = 47)	School 2 (*n* = 37)	School 3 (*n* = 48)	Total (*N* = 132[a])
Variable	*n* (%)	*n* (%)	*n* (%)	*N* (%)
Age in years				
4 years	0 (0)	0 (0)	1 (2.1)	1 (0.7)
5 years	0 (0)	22 (59.5)	27 (56.3)	49 (37.1)
6 years	28 (58.3)	15 (40.5)	20 (41.6)	63 (47.8)
7 years	19 (39.6)	0 (0)		19 (14.4)
Sex				
Male	19 (39.6)	21 (56.8)	23 (47.9)	63 (47.7)
Female	28 (58.3)	16 (43.2)	25 (52.1)	69 (52.3)
Race				
American Indian	1 (2.1)	0 (0)	2 (4.2)	3 (2.3)
African American	5 (10.4)	7 (18.9)	17 (35.4)	29 (22.0)
Asian	4 (8.3)	1 (2.7)	1 (2.1)	6 (4.5)
Hispanic	2 (4.2)	8 (21.6)	14 (29.2)	24 (18.2)
White	35 (72.9)	21 (56.8)	14 (29.2)	70 (53.0)
FRL				
No FRL	35 (75)	20 (54.1)	17 (35.4)	72 (54.5)
Receives FRL	12 (25)	17 (45.9)	31 (64.6)	60 (45.5)
SPED				
No IEP	46 (95.8)	32 (86.5)	39 (81.3)	117 (88.6)
Has IEP	1 (2.1)	5 (13.5)	9 (18.7)	15 (11.4)
ELL status				
Non-ELL	42 (87.5)	31 (83.8)	36 (75)	109 (82.6)
ELL	5 (10.4)	6 (16.2)	12 (25)	23 (17.4)
Home language				
English	43 (89.6)	30 (81.1)	33 (68.8)	106 (80.3)
French	0 (0)	0 (0)	1 (2.1)	1 (0.7)
Spanish	0 (0)	6 (16.2)	14 (29.2)	20 (15.3)
Somali	0 (0)	1 (2.7)	0 (0)	1 (0.7)
Others	4 (10.4)	0 (0)	0 (0)	4 (3.0)

Note. FRL = free reduced lunch; SPED = special education status; IEP = individualized education program; ELL = English language learner.
[a]Demographic data were missing for one student.

Figure 1. Sample Picture Word prompt.

Construction (including quality of plot, prose, character development, interest, and other compositional elements). Alternate-form reliabilities of the Spontaneous Writing subtest for 7-year-olds have been reported to be $r = .60$ for Contextual Conventions, $r = .81$ for Contextual Language, and $r = .87$ for Story Construction, and validity correlation coefficients with the Writing Scale of the Comprehensive Scales of Student Abilities (Hammill & Hresko, 1994) were .46 to .48 for Contextual Conventions, .43 to .44 for Contextual Language, and .34 for Story Construction for elementary students (Hammill & Larsen, 1996).

Procedures

Test administration. In McMaster et al. (2009), CBM-W Picture Word prompts were administered in February to March and again in May 2006 by two graduate research assistants (GRAs). In McMaster et al. (2011), Picture Word prompts were administered weekly over 12 weeks by an advanced doctoral student in special education and five classroom teachers from February to May in 2007. For the current study, scores from CBM prompts administered in February were used. The doctoral student trained teachers how to administer the prompts during a 1-hr session. The trainer modeled how to administer the prompts in each classroom, and the classroom teachers administered the CBM-W Picture Word prompts weekly during the remaining weeks. The TOWL-3 was administered by a researcher in May 2006 and 2007 for each study.

Fidelity. In McMaster et al. (2009), graduate students administered CBM-W; fidelity of administration was not formally assessed. In McMaster et al. (2011), a graduate student observed teachers' administration of CBM-W tasks using a checklist comprising nine items that assessed teachers' implementation of directions, timing, and responding to student behaviors. Example items include "presents an example of Picture Word prompt on the board," "demonstrates how students should complete the entire Picture Word task with the sample copy," "demonstrates how to deal with spelling difficulties while taking test," and "starts/stops timer at the correct times." All five classroom teachers administered the writing tasks with 100% accuracy. According to notes from the observer, students followed teachers' directions and worked quietly.

Scoring and interrater agreement. Procedures of scoring and interrater agreement for the three CBM-W tasks and the TOWL-3 were conducted identically across both studies. Scorers included the same second author of the two studies, four graduate research assistants (GRAs; doctoral students in special education or school psychology), and one special education teacher. First, each scorer scored one set of writing samples, compared the scores, discussed differences, and resolved questions. Then, each scorer scored additional samples independently. To maintain at least 80% interrater agreement, an "expert" scorer (advanced special education doctoral student) compared each GRA's scores with hers. If agreement was less than 80%, she re-trained the GRA and re-scored any protocols scored by this GRA. To calculate agreement, one out of every eight samples was randomly selected and scored by the expert independently. Agreement was calculated using a point-by-point method, and the number of agreements was divided by agreements plus disagreements. Average agreement ranged from 89% to 100% across studies.

For the TOWL-3, interrater agreements were calculated using the same point-by-point method on a randomly selected 10% of writing samples. Interrater agreements were above 90% for the three categories of the TOWL-3 in McMaster et al. (2009) and 96% on Contextual Conventions, 90% on Contextual Language, and 79% on Story Construction in McMaster et al. (2011). Scoring and interrater agreement procedures are described in detail in the two previous studies.

Selecting cut-off scores. Literature addressing ways to identify students who do not respond to research-based instruction has used at-risk criteria ranging from below the 30th to 15th percentile (e.g., Mathes et al., 2005; Scanlon, Vellutino, Small, Fanuelle, & Sweeney, 2005; Simmons et al., 2008). For this study, the lowest 15th percentile based on the TOWL-3 norms was selected to identify students most at risk in writing. The cut-off scores corresponding to the lowest 15th percentile for ages 7.0 to 7.11 were 0 for Contextual Conventions, 4 for Contextual Language, and 1 for Story Construction (Hammill & Larsen, 1996).

Data Analysis

Before examining classification accuracy of CBM-W tasks, correlations among screening and criterion measures were examined. Criterion validity coefficients were calculated between CBM-W Picture Word prompts and the TOWL-3. Three categories (Contextual Conventions, Contextual Language, and Story Construction) and Total scores (the sum of the three categories) were used as criterion measures. Criterion validity correlations are reported in the "Results."

Among TOWL-3 measures, Contextual Conventions was excluded because it showed weak, nonsignificant criterion validity, and the Total score was excluded as it yielded a high base rate (further evidence for excluding the two scores are reported in the "Results"). Thus, only two TOWL-3 categories (Contextual Language and Story Construction) were included for further analyses. ROC curve analysis was used to explore the classification characteristics of CBM-W including sensitivity, specificity, and overall accuracy represented by the AUC. ROC curve

Table 2. Descriptive Statistics for CBM-W Picture Word Prompts and the TOWL-3.

Measure	*M*	*SD*	Minimum	Maximum	Skewness	Kurtosis
CBM-W Picture Word						
WW	15.26	7.36	0.00	35.67	0.48	0.06
WSC	12.32	7.22	0.00	33.00	0.63	0.08
CWS	9.89	7.03	0.00	29.00	0.71	−0.05
TOWL-3						
Contextual Conventions	1.78	2.07	0.00	11.00	1.90	4.77
Contextual Language	6.98	3.60	0.00	15.00	−0.37	−0.47
Story Construction	4.09	3.89	0.00	17.00	1.03	0.68
Total	12.94	7.69	0.00	41.00	0.70	1.20

Note. CBM-W = curriculum-based measures in writing; TOWL-3 = Test of Written Language–Third Edition; WW = words written; WSC = words spelled correctly; CWS = correct words sequences.

analysis provides all possible sensitivity and specificity in pairs over a range of outcomes of the measures and presents the pairs on an ROC curve. The horizontal axis of the ROC curve plot represents false positive rates, and the vertical axis of the plot represents true positive rates. A straight diagonal line in the plot would indicate that the screening measure identifies students as at risk or not at risk at a rate no better than chance. The ROC analysis also provides a single number indicating the overall classification accuracy of the measure represented by the AUC.

PPV and NPV were calculated based on the four categories that students were classified into corresponding to the results of the CBM-W and criterion measure: (a) *true positives* refer to those identified as at risk on both CBM-W and criterion measure, (b) *true negatives* are those identified as not at risk on both CBM-W and criterion measure, (c) *false positives* are those identified as at risk on CBM-W, but not on the criterion measure, and (d) *false negatives* are those identified as not at risk on the CBM-W, but at risk on the criterion measure. PPV is calculated as the number of true positives divided by the sum of true and false positives. NPV is calculated as the number of true negatives divided by the sum of true and false negatives.

Results

Descriptive Statistics

Table 2 shows descriptive statistics including means, *SD*s, skewness, and kurtosis of scores on CBM-W Picture Word and TOWL-3 for all students ($n = 133$). Minimum scores on all measures included 0, indicating possible floor effects; however, skewness and kurtosis values alleviated this concern (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009), as most values were within the range of −1.96 and 1.96, except for Contextual Conventions (1.90 for skewness and 4.77 for kurtosis). Picture Word and Contextual Language showed normal distributions. The remaining measures—Story Construction and TOWL-3 Total scores—showed positively skewed but not extreme distributions. Distribution graphs are presented in Figure 2.

Students at Risk in Writing Based on TOWL-3 Scores

Using below the 15th percentile as the cut-off for "at-risk" status on each TOWL 3 category and Total score, 41 students (33%) were identified as at risk on Contextual Conventions, 22 (18%) on Contextual Language, and 17 (20%) on Story Construction. Sixty-two students (50%) were not identified as at risk in writing on any of the three categories. One hundred thirteen students (92%) were identified as at risk based on the Total score. The base rate on the Total score was unusually high, which we suspected was because children in this sample were on the lower edge of the normative age range for the TOWL-3. Given concerns that this high base rate would lead to inaccurate classification information, especially PPV and NPV, the Total score was excluded from classification accuracy analyses.

Depending on the identification outcomes from CBM-W and TOWL-3, the four categories including true positives, true negatives, false positives, and false negatives were examined and summarized in Table 3 for each scoring procedure of CBM-W Picture Word used as a screening measure and the TOWL-3 used as a criterion measure.

Criterion Validity

Criterion validity coefficients among CBM-W Picture Word and the TOWL-3 scores are displayed in Table 4. Picture Word data were statistically significantly correlated with the TOWL-3, ranging from $r = 0.27$ to 0.47 ($ps < .01$), except for with Contextual Conventions ($r = 0.53$ with WW, 0.07 with WSC, and 0.13 with CWS; all $ps \geq .05$). Given that criterion validity evidence is necessary (but not sufficient) to identify potential screening measures (Jenkins et al., 2007), only the two categories with

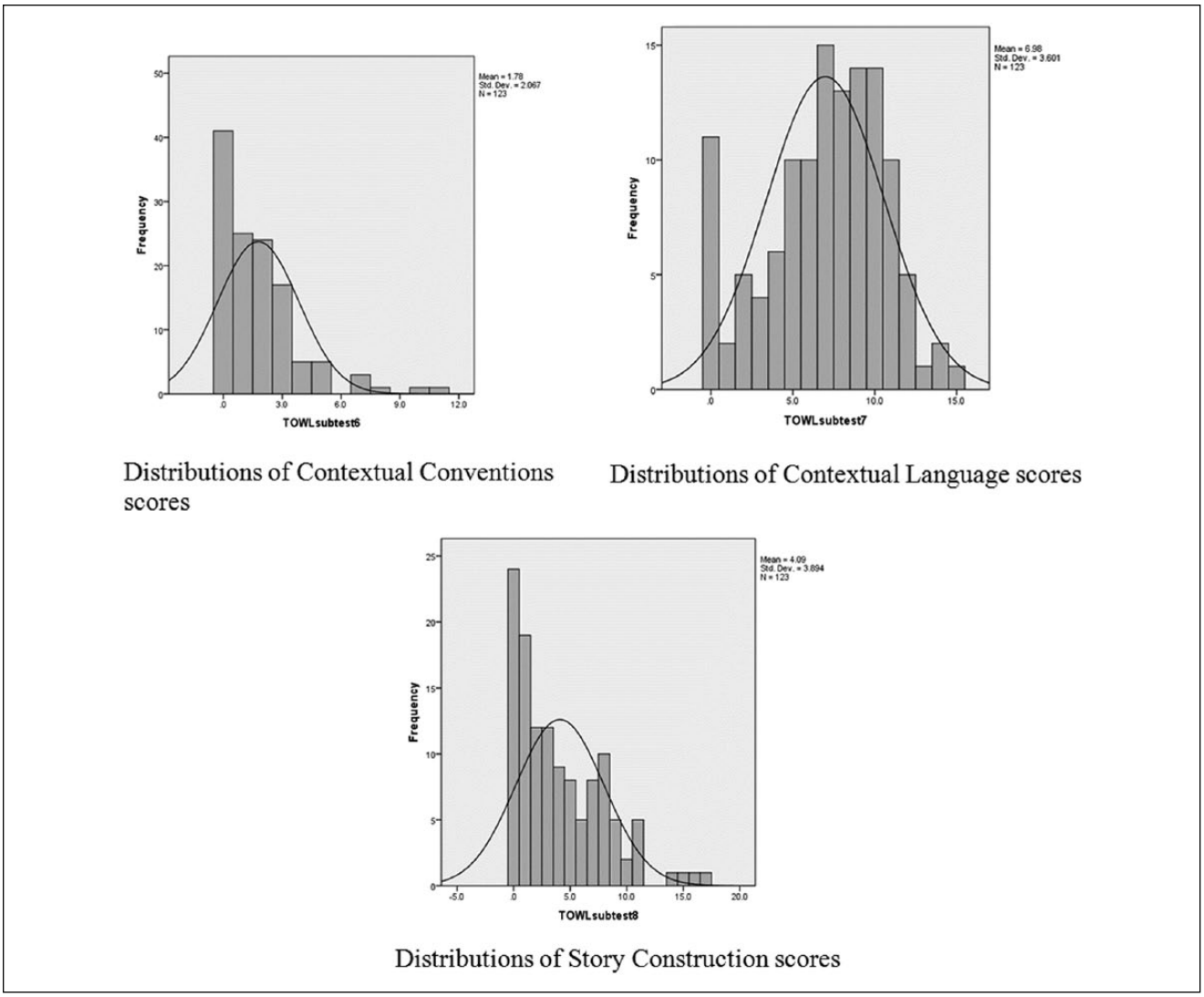

Figure 2. Distributions of TOWL-3 scores by subtest.
Note. TOWL-3 = Test of Written Language–Third Edition.

Table 3. Numbers of Students Identified as At Risk or Not At Risk on CBM-W and TOWL-3 Measures.

	CBM-W Picture Word					
	WW		WSC		CWS	
TOWL-3	At risk	Not at risk	At risk	Not at risk	At risk	Not at risk
Contextual Language						
At risk	18	2	18	1	18	1
Not at risk	50	42	38	54	68	24
Story Construction						
At risk	18	2	18	1	18	1
Not at risk	50	42	56	36	68	24

Note. CBM-W = curriculum-based measures in writing; TOWL-3 = Test of Written Language–Third Edition; WW = words written; WSC = words spelled correctly; CWS = correct words sequences.

Table 4. Criterion Validity of CBM-W Picture Word Prompts and the TOWL-3.

	CBM-W Picture Word			TOWL-3		
	WW	WSC	CWS	Contextual conventions	Contextual language	Story construction
CBM-W Picture Word						
WW						
WSC	**0.94****					
CWS	**0.84****	**0.92****				
TOWL-3						
Contextual Conventions	0.53	0.07	0.13			
Contextual Language	**0.45****	**0.47****	**0.47****	**0.26****		
Story Construction	**0.28****	**0.27****	**0.27****	**0.26****	**0.61****	
Total	**0.38****	**0.40****	**0.42****	**0.54****	**0.86****	**0.86****

Note. Significant correlation coefficients are in boldface. CBM-W = curriculum-based measures in writing; TOWL-3 = Test of Written Language–Third Edition; WW = words written; WSC = words spelled correctly; CWS = correct words sequences.
*$p < .05$. **$p < .01$.

Table 5. Classification Accuracy of CBM-W Picture Word Prompts for the TOWL-3.

Criterion measure	Screening measure	Cut	Sensitivity	Specificity	PPV	NPV	AUC	*p* value	95% CI	
									Low	High
Contextual Language	WW	16.2	.900	.457	.265	.955	.790	<.001	.689	.891
	WSC	11.5	.947	.587	.321	.982	.822	<.001	.727	.917
	CWS	15.2	.947	.261	.209	.960	.831	<.001	.717	.945
Story Construction	WW	16.2	.900	.457	.265	.955	.768	<.001	.662	.874
	WSC	14.2	.947	.391	.243	.973	.763	<.001	.652	.873
	CWS	15.2	.947	.261	.209	.960	.727	0.002	.594	.860

Note. CBM-W = curriculum-based measures in writing; TOWL-3 = Test of Written Language–Third Edition; PPV = positive predictive value; NPV = negative predictive value; AUC = area under curve; CI = confidence interval; WW = words written; WSC = words spelled correctly; CWS = correct words sequences.

evidence of criterion validity (Contextual Language and Story Construction) were selected for further classification accuracy analyses.

Classification Accuracy of CBM-W

Classification results from Picture Word prompts were compared with dichotomous outcomes derived from two categories (Contextual Language and Story Construction): Students who scored below the 15th percentile on these categories were considered to be at risk in writing; otherwise, they were considered not at risk.

Table 5 displays classification accuracy results for CBM-W Picture Word with Contextual Language and Story Construction. When sensitivity was held at or above .90, specificity was higher for WW (.457 for both categories) and WSC (.587 for Contextual Language and .391 for Story Construction) than for CWS (.261 for both categories). Sensitivity of CBM-W Picture Word ranged from .900 to .947 across the two categories. PPV—the capacity of CBM-W Picture Word prompts to identify students who are actually at risk based on the TOWL-3—was low for Contextual Language and Story Construction (.209 to .321). NPV was high for Contextual Language and Story Construction (mostly at or above .950). AUC ranged from .790 to .831 for Contextual Language, and .727 to .768 for Story Construction across all scoring metrics.

Discussion

In this study, we examined classification accuracy of CBM-W Picture Word scored for WW, WSC, and CWS to determine utility for screening first graders for risk in writing.

Criterion Validity

CBM-W Picture Word scores were statistically significantly and moderately correlated with Contextual Language and Story Construction categories of the TOWL-3, particularly

between CWS and Contextual Language ($r = .47$, $p < .01$), which measures quality of vocabulary, sentence construction, and grammar. This finding supports the use of Picture Word prompts as an indicator of students' sentence-level writing competency (McMaster et al., 2009). Further, Picture Word was moderately correlated with Total TOWL-3 scores ($r = .37$ to $.42$ across scoring procedures, $ps < .01$), reflecting that Picture Word prompts appear to capture overall writing proficiency. These findings are consistent with previous CBM-W research (McMaster et al., 2011).

Classification Accuracy

Sensitivity and specificity. CBM-W Picture Word prompts produced a range of specificity—from .261 to .587—when sensitivity was set at or above .90. In an RTI framework, sensitivity of .90 or above represents the optimal criterion for screening (Compton, Fuchs, Fuchs, & Bryant, 2006). There is little consensus regarding the acceptable criterion for specificity, but Catts et al. (2009) suggested that specificity of 50% or more is acceptable. Using this guideline, Picture Word prompts scored for WSC met the criterion, showing a specificity level of .587 when sensitivity was above .90 in relation to Contextual Language. In other words, Picture Word scored for WSC accurately identified approximately 95% of students truly at risk, and 59% of students truly not at risk. Picture Word scored for WW also showed promising evidence; WW with Contextual Language and Story Construction showed a specificity level of .457. Overall, CBM-W Picture Word showed promise to accurately identify students at risk in writing when holding the sensitivity level of .90 or above, which meets the purpose of screening to identify all or nearly all students who truly need intervention (Johnson, Jenkins, & Petscher, 2010).

Positive and negative predictive power. PPV and NPV provide more information to interpret screening results for individual students beyond the measure's accuracy itself (Zweig & Campbell, 1993). CBM-W Picture Word showed low PPV (.209 to .321) and high NPV (.955 to .982). For example, based on Contextual Language, if a student is identified as at risk in writing on Picture Word, there is approximately a 20% to 30% possibility that the student really is at risk. If a student is identified as not at risk in writing on Picture Word, there is about a 95% to 98% possibility that the student really is not at risk. The high NPV rates indicate that Picture Word can accurately identify students who may not need additional intervention in writing.

Classification accuracy findings in this study are consistent with findings for the Sentence Writing task for first graders examined by Ritchey and Coker (2014). Ritchey and Coker showed acceptable sensitivity and specificity levels for qualitative scores on Sentence Writing with teacher ratings, but not with two subtests of the WJ3 (Woodcock et al., 2001, 2007). The current study provided evidence of acceptable sensitivity and specificity levels of Picture Word scored for WSC and nearly acceptable levels of Picture Word scored for WW in relation to the TOWL-3. Thus, there is now preliminary evidence that two sentence-level CBM-W tasks, Sentence Writing and Picture Word, may be considered for use as screening measures for first graders.

Results of this study underscore the importance of examining classification accuracy beyond criterion validity (cf. Jenkins et al., 2007). Although Picture Word scores showed statistically significant correlations with TOWL-3 scores ($r = .30$ to $.49$), varying levels of specificity were found depending on scoring metrics. Among the three metrics (WW, WSC, and CWS), Picture Word tended to have nearly acceptable to acceptable levels of specificity for WW and WSC, but not for CWS, which is useful information because WW and WSC are more efficient to score than CWS. This finding also suggests that the Picture Word task may function differently depending on how it is scored. Researchers and practitioners should be aware of sensitivity and specificity levels of various scoring metrics in addition to criterion validity when they use CBM-W to identify students at risk in writing. Also, given that CWS appears more sensitive to growth than WW and WSC in previous research (McMaster et al., 2011), these findings suggest that different scoring procedures might be more appropriate for different purposes (e.g., WSC may be used for screening, but CWS might be more appropriate for monitoring progress).

Overall classification accuracy. Overall classification accuracy of CBM-W Picture Word, indicated by AUC, mostly ranged from .73 to .83, indicating the potential of Picture Word to distinguish between students at risk and not at risk in writing. AUC values, however, do not indicate that educators should solely rely on Picture Word for screening. AUC represents overall classification accuracy by condensing all information under the ROC curve. Although AUC is "the most common global measure" to present classification accuracy (Zweig & Campbell, 1993, p. 568), important information may be lost when expressing it in a single number. In addition, previous studies indicated that AUC values were strengthened when writing measures were combined with reading measures (Coker & Ritchey, 2014; Ritchey & Coker, 2014).

Limitations and Future Research

Although this study provides preliminary classification accuracy evidence of CBM-W Picture Word, several factors limit interpretation of the findings. First, findings of this study may not generalize to the broader population of beginning writers given the small number of participants from three urban schools, and so caution should be used in

applying these findings to other groups. Further research is needed with larger, more nationally representative samples. In addition, findings are limited to the two TOWL-3 categories (Contextual Language and Story Construction) that had acceptable criterion validity and base rates for the purposes of this study. Further research should determine whether findings generalize to other criterion measures, particularly given the lack of a "gold standard" writing measure. Despite this limitation, findings from this study contribute to the literature by showing the potential utility of CBM-W Picture Word—a measure with evidence of reliability, validity, and sensitivity to growth—for universal screening within multitiered systems of support.

Second, and related to the above point, the TOWL-3 may not be the best measure to identify students at risk in writing for first graders. As mentioned earlier, data for this study were combined from two previous studies (McMaster et al., 2009; McMaster et al., 2011), which both used the TOWL-3 for the purpose of establishing criterion validity evidence, because it was determined to be the best available criterion measure when the original studies were conducted. However, weak to moderate criterion validity coefficients (r = .34 to .48) raise concerns about using this measure for the purpose of screening in schools and research.

Although the age of students in our sample (M = 6.56 years) was below the normative age range for the measure (7 to 11 years), students' mean performance on each TOWL-3 category was close to the 50th percentile of the normative sample for 7-year-olds. The cut-off raw score for the 15th percentile on Story Construction was 1.0, indicating that the measure might not be sensitive enough to distinguish students at risk in writing from those not at risk. In addition, the high base rate (92%) for the TOWL-3 Total scores further indicates that the Total scores failed to discriminate between students struggling in writing and those not struggling, suggesting that the assessment may have been too difficult for first graders. A different criterion measure normed for younger students may have discriminated better among children at risk and not at risk.

The limitations above and findings in this study provide directions for future research. First, researchers should replicate this study to examine classification accuracy of CBM-W Picture Word using additional criterion measures and scoring procedures, because the accuracy of screening measures such as sensitivity and specificity varies depending on the types of criterion measures used (VanDerHeyden, 2011). Future researchers should also examine the use of other types of CBM-W for screening, such as Sentence Copying or Story Prompts for early elementary students, because these measures have also shown promising technical features in previous studies and can be group administered (McMaster et al., 2009; McMaster et al., 2011; Ritchey & Coker, 2013). Given that this study found promising evidence that a sentence-level writing task can be used for screening first graders, researchers should conduct longitudinal studies to determine whether sentence-level writing tasks identify struggling writers accurately. In addition, future researchers should examine and identify appropriate scoring procedures with accuracy and efficiency of screening that can be applied to specific grades.

Second, additional research is needed to examine ways to improve accuracy of CBM-W Picture Word, such as by adding scores from additional measures. Previous studies provided promising evidence of combining reading and writing measures for screening for kindergarten and first grade students (Coker & Ritchey, 2014; Ritchey & Coker, 2014). Continued research is needed to determine the best measures (single or in combination) for determining students' level of risk in writing. At the same time, researchers should continue to explore ways to optimize both accuracy *and* efficiency given limited resources and time available in schools.

Implications for Practice

Results of this study suggest that practitioners may use CBM-W Picture Word as part of screening for first-grade students at risk in writing. In addition, WSC showed evidence of acceptable sensitivity and specificity levels compared with WW and CWS. This finding is encouraging given that WSC is more efficient to calculate than CWS. However, three things should be considered to apply in practice. First, to obtain similar results to those in this study, practitioners must determine the extent to which the sample characteristics, as well as the focus and content of criterion measures used in this study, are relevant to their specific educational settings. Sample characteristics, types of criterion measures, decision-making rules used, and interventions that are in place all influence sensitivity and specificity levels (VanDerHeyden, 2011). Second, practitioners should consider the cost of incorrect classification. High sensitivity leads to high proportions of false positives—in other words, rates of students identified at risk who are actually not at risk. Increasing false positives is problematic because it leads to wasting educational resources and decreasing intervention intensity for those who really need the intervention, assuming limited resources for supplemental and special education services (Johnson et al., 2010). Third, practitioners should note that we used CBM-W data administered in February of first grade; results would likely be different at other times of the year.

Practitioners may also use CBM-W Picture Word prompts as an indicator of students' overall performance in writing. The results of criterion validity showed moderate correlations with Contextual Language and Story Construction categories and the Total scores of the TOWL-3, which is promising evidence considering the complex and multidimensional process of writing (Berninger &

Swanson, 1994). In addition, because first graders are in the early developmental stages of writing, practitioners should not rely on a single CBM-W score to determine students' current level of writing or make decisions for further diagnosis or intervention. Rather, they should consider using CBM-W scores as one of multiple sources of information to make sound instructional decisions.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by Grant H324H030003 awarded to the Institute on Community Integration and the Department of Educational Psychology, College of Education and Human Development, at the University of Minnesota, by the Office of Special Education Programs in the U.S. Department of Education.

References

Berninger, V. W., & Amtmann, D. (2003). Preventing written expression disabilities through early and continuing assessment and intervention for handwriting and/or spelling problems: Research into practice. In H. Swanson, K. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 323–344). New York, NY: The Guilford Press.

Berninger, V. W., Nielsen, K. H., Abbott, R. D., Wijsman, E., & Raskind, W. (2008). Writing problems in developmental dyslexia: Under-recognized and under-treated. *Journal of School Psychology*, *46*, 1–21. doi:10.1016/j.jsp.2006.11.008

Berninger, V. W., & Swanson, H. L. (1994). Modifying Hayes and Flower's model of skilled writing to explain beginning and developing writing. In E. C. Butterfield & J. Carlson (Eds.), *Children's writing: Toward a process theory of the development of skilled writing* (pp. 57–81). London, England: JAI Press.

Berninger, V. W., & Winn, W. (2006). Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 96–114). New York, NY: The Guilford Press.

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

Coker, D. L., & Ritchey, K. D. (2010). Curriculum-based measurement of writing in kindergarten and first grade: An investigation of production and qualitative scores. *Exceptional Children*, *76*, 175–193. doi:10.1177/001440291007600203

Coker, D. L., & Ritchey, K. D. (2014). Universal screening for writing risk in kindergarten. *Assessment for Effective Intervention*, *39*, 245-256. doi: 10.1177/1534508413502389

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, *98*, 394–409. doi:10.1037/0022-0663.98.2.394

Cooper, D., & Pikulski, J. (2005). *Here we go!* Boston, MA: Houghton Mifflin.

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

Deno, S. L. (2003). Developments in curriculum-based measurement. *Journal of Special Education*, *37*, 184–192. doi:10.1177/00224669030370030801

Deno, S. L., Mirkin, P., & Marston, D. (1980). *Relationships among simple measures of written expression and performance on standardized achievement tests* (Vol. IRLD-RR-22). Minneapolis: Institute for Research on Learning Disabilities, University of Minnesota.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review*, *33*, 188–192.

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

Fuchs, L. S., & Fuchs, D. (2007). A model for implementing responsiveness to intervention. *Teaching Exceptional Children*, *39*, 14–20. doi:10.1177/004005990703900503

Gansle, K. A., VanDerHeyden, A. M., Noell, G. H., Resetar, J. L., & Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students. *School Psychology Review*, *35*, 435–450.

Graham, S. (2008). *Effective writing instruction for all students written for renaissance learning*. Retrieved from http://doc.renlearn.com/KMNet/R004250923GJCF33.pdf

Graham, S., & Harris, K. R. (2000). The role of self-regulation and transcription skills in writing and writing development. *Educational Psychologist*, *35*, 3–12. doi:10.1207/S15326985EP3501_2

Hammill, D. D., & Hresko, W. P. (1994). *Comprehensive scales of student abilities*. Austin, TX: PRO-ED.

Hammill, D. D., & Larsen, S. C. (1996). *Test of Written Language–Third edition*. Austin, TX: PRO-ED.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3-30). Hillsdale, NJ: Lawrence Erlbaum.

Hresko, W. P., Herron, S. R., & Peak, P. K. (1996). *Test of Early Written Language* (2nd ed.). Austin, TX: PRO-ED.

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

Jewell, J., & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review*, *34*, 27–44.

Johnson, E. S., Jenkins, J. R., & Petscher, Y. (2010). Improving the accuracy of a direct route screening process. *Assessment for Effective Intervention*, *35*, 131–140. doi:10.1177/1534508409348375

Lembke, E., Deno, S. L., & Hall, K. (2003). Identifying an indicator of growth in early writing proficiency for elementary

school students. *Assessment for Effective Intervention*, *28*, 23–35. doi:10.1177/073724770302800304

Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Reading Research Quarterly*, *40*, 148–182. doi:10.1598/RRQ.40.2.2

McCutchen, D. (2006). Cognitive factors in the development of children's writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 115–130). New York, NY: The Guilford Press.

McMaster, K. L., Du, X., & Pétursdóttir, A. (2009). Technical features of curriculum-based measures for beginning writers. *Journal of Learning Disabilities*, *42*, 41-60. doi: 10.1177/0022219408326212

McMaster, K. L., Du, X., Yeo, S., Deno, S. L., Parker, D., & Ellis, T. (2011). Curriculum-based measures of beginning writing: Technical features of the slope. *Exceptional Children*, *77*, 185-206. doi: 10.1177/001440291107700203

McMaster, K. L., Kunkel, A., Shin, J., Jung, P., & Lembke, E. (2017). Early writing interventions: A best-evidence synthesis. *Journal of Learning Disabilities*. Advance online publication. doi: 10.1177/0022219417708169

National Assessment of Educational Progress. (2012). *The Nation's Report Card: Writing 2011*. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards: English Language arts standards, Grade 1*. Washington, DC: Author.

Parker, R., Tindal, G., & Hasbrouck, J. (1991). Countable indices of writing quality: Their suitability for screening-eligibility decisions. *Exceptionality*, *2*, 1–17. doi:10.1080/09362839109524763

Ritchey, K. D. (2006). Learning to write: Progress-monitoring tools for beginning and at-risk writers. *Teaching Exceptional Children*, *39*, 22–26. doi:10.1177/004005990603900204

Ritchey, K. D., & Coker, D. L. (2013). An investigation of the validity and utility of two curriculum-based measurement writing tasks. *Reading & Writing Quarterly*, *29*, 89–119. doi:10.1080/10573569.2013.741957

Ritchey, K. D., & Coker, D. L. (2014). Identifying writing difficulties in first grade: An investigation of writing and reading measures. *Learning Disabilities Research & Practice*, *29*, 54–65. doi:10.1111/ldrp.12030

Scanlon, D. M., Vellutino, F. R., Small, S. G., Fanuelle, D. P., & Sweeney, J. M. (2005). Severe reading difficulties—Can they be prevented? A comparison of prevention and intervention approaches. *Exceptionality*, *13*, 209–227. doi:10.1207/s15327035ex1304_3

Simmons, D. C., Coyne, M. D., Kwok, O., McDonagh, S., Harn, B. A., & Kame'enui, E. J. (2008). Indexing response to intervention: A longitudinal study of reading risk from kindergarten through third Grade. *Journal of Learning Disabilities*, *41*, 158–173. doi:10.1177/0022219407313587

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26. doi:10.1111/1529-1006.001

VanDerHeyden, A. M. (2011). Technical adequacy of response to intervention decisions. *Exceptional Children*, *77*, 335–350. doi:10.1177/001440291107700305

Videen, J., Deno, S. L., & Marston, D. (1982). *Correct word sequences: A valid indicator of proficiency in written expression* (Vol. IRLD-RR-84). Minneapolis: Institute for Research on Learning Disabilities, University of Minnesota.

Woodcock, R. W., McGrew, F. A., Schrank, F. A., & Mather, N. (2007). *Woodcock Johnson III Tests of Achievement Normative Update.* Rolling Meadows, IL: Riverside. (Original work published 2001).

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*, 561–577.