12-10-2019

Technical Adequacy of Curriculum-Based Measures in Writing in Grades 1–3

Abigail A. Allen

Pyung-Gang Jung

Apryl L. Poch

Dana Brandes

Jaehyun Shin

See next page for additional authors

Authors

Abigail A. Allen, Pyung-Gang Jung, Apryl L. Poch, Dana Brandes, Jaehyun Shin, Erica S. Lembke, and Kristen L. McMaster

Technical Adequacy of Curriculum-Based Measures in Writing in Grades 1–3

Abigail A. Allen[a], Pyung-Gang Jung[b], Apryl L. Poch[c], Dana Brandes[d], Jaehyun Shin[e], Erica S. Lembke[f], and Kristen L. McMaster[d]

[a]Clemson University, Clemson, SC, USA;

[b]Ewha Womans University, Seoul, Republic of Korea;

[c]Duquesne

University, Pittsburgh, PA, USA;

[d]University of Minnesota, Minneapolis, MN, USA;

[e]Gyeongin National University, Incheon, Republic of Korea;

[f]University of Missouri, Columbia, MO, USA

To cite this article: Abigail A. Allen, Pyung-Gang Jung, Apryl L. Poch, Dana Brandes, Jaehyun Shin, Erica S. Lembke & Kristen L. McMaster (2020) Technical Adequacy of Curriculum- Based Measures in Writing in Grades 1–3, Reading & Writing Quarterly, 36:6, 563-587, DOI: https://doi.org/10.1080/10573569.2019.1689211

ABSTRACT

The purpose of this study was to investigate evidence of reliability, criterion validity, and grade-level differences of curriculum-based measures of writing (CBM-W) with 612 students in grades 1–3. Four scoring procedures (words written, words spelled correctly, correct word sequences, and correct minus incorrect word sequences) were used with two CBM-W tasks (picture–word and story prompt) during fall, winter, and spring of one academic year. A subsample of participants (*n* = 244) were given a criterion measure in spring of the academic year. Pearson's *r* coefficients were calculated to determine evidence of alternate form reliability and criterion validity, and a MANOVA was used to detect significant growth within and across grade levels. Results indicated that scores on both CBM-W tasks had adequate reliability and validity coefficients in grades 2–3 and mixed results in grade 1. Significant growth was detected within and across all grades at each time point on each task. Implications for research and practice are discussed.

Writing is a complex task involving multiple cognitive and linguistic factors and is critical to students' academic and vocational success. Students who struggle with writing during their education typically have fewer postsecondary and employment opportunities (Graham & Perin, 2007). Despite the importance of writing proficiency, U.S. students' writing performance is often unsatisfactory. Nearly 75% of students in fourth, eighth, and twelfth grades were not proficient in writing according to the most recent writing data from the National Assessment of Educational Progress (NAEP; National Center for Education Statistics, 2011). Writing difficulties or disabilities are often not identified until intermediate grades when required writing tasks become more complex (Berninger et al., 2002, 2006); however, converging empirical evidence supports the benefits of early identification and early intervention to students struggling with writing in early elementary grades (McMaster, Kunkel, Shin, Jung, & Lembke, 2018). It is essential to identify students who are struggling with written language early using reliable and valid assessments to prevent or lessen the effect of writing difficulties and disabilities.

Conceptual models of writing provide a way to understand what writing entails. The Simple View of Writing is one such framework designed around core lower and higher order skills. Developed by Juel, Griffith, and Gough (1986), the model examined the interaction of two variables, transcription and ideation. Transcription, the lower order skill, was focused on spelling and handwriting, drawing on subskills like alphabetic and orthographic knowledge. Ideation, the higher order skill, was focused on developing ideas to write about and develop within one's writing. In later years, Berninger et al. (2002) posited another version of the Simple View of Writing in an effort to better understand beginning writing. Their conceptual model was represented as a triangle with transcription and self-regulatory executive functions (e.g. self-evaluation, goal- setting, and self-reinforcement) as the base features and text generation (a reconceptualization of ideation as the oral representation of written language) as the "vertex" (Berninger et al., 2002, p. 292; Berninger & Amtmann, 2003, p. 349), with all elements constrained by memory (i.e., short-term, working, and long-term memory). We use the Simple View of Writing as a theoretical model for our work reported in this manuscript as it served as the theoretical foundation for our understanding and refining curriculum-based measures in early writing (CBM-W) that we report here.

Previous research has shown that using writing measures with evidence of reliability and validity can have a positive effect on student performance in spelling and writing when teachers use the measures for screening, monitoring progress, and making instructional changes (Fuchs, Fuchs, & Hamlett, 1989; Jung, McMaster, & delMas, 2017). In addition to evidence of reliability and validity, writing measures should be sensitive to student growth, particularly that of students with disabilities or learning problems, as they tend to grow at a slower rate and traditional assessment methods may not adequately capture their learning (Deno, 2003). Additionally, writing assessments must be simple and easy to administer and score, and assessment results should be easy to understand and clear enough to facilitate communication among professionals in schools. A measure that has evidence producing reliable, valid, and usable data is CBM-W (Deno, 1985; Deno, 2003). Researchers have explored CBM-W as an alternative to traditional methods of writing assessment, such as holistic ratings and rubrics, which tend to be less reliable and not sensitive to growth for students who struggle with writing (McMaster & Espin, 2007).

Curriculum-based measurement in writing

Curriculum-based measurement is designed to be a *global indicator*, meaning it is an indicator of overall proficiency in an academic area (Fuchs & Deno, 1991). Curriculum-based measurement in writing allows educators to directly observe and score students' performance on standard writing tasks to screen for risk, assess growth, and adjust instruction when progress is insufficient to meet important benchmarks (Deno, 2003). Research into the reliability and validity evidence of CBM-W in the early elementary grades has emerged over the last 15 years (McMaster, Ritchey, & Lembke, 2011). Below, we present writing tasks used in CBM research, and then describe scoring methods for each task.

Curriculum-based measures of writing tasks

The original CBM-W tasks studied by Deno, Mirkin, and Marston (1980) involved passage-level writing (henceforth referred to as “story prompts”). Story prompts align with the Simple View of Writing in terms of text generation, or developing and translating ideas into sentences and para- graphs, and transcription, or the spelling and handwriting skills required to put words onto paper. Although self-regulation is not traditionally assessed using CBM-W, it is inherently part of planning and executing a sentence or story and can be scaffolded by using different types of CBM-W prompts discussed below. The act of holding sounds, words, and information in the mind while crafting sentences is an essential part of the writing task at any level and it may be necessary to consider the influence of self-regulation on struggling writers’ performance on assessments (Graham, Harris, & McKeown, 2013; Swanson & Zheng, 2013).

Initial CBM-W research in story prompts involved giving participants 3–5 min to write a story based on a topic sentence (e.g. “Write about your best day of school”). Deno and colleagues found that story prompts demonstrated evidence of criterion validity (r 2≥ 0.70) for students in third through sixth grade (Deno, Marston, & Mirkin, 1982; Deno et al., 1980). Later studies found that scores obtained from story prompts (McMaster, Du, & Petursdottir, 2009; McMaster & Campbell, 2008; Ritchey & Coker, 2013) generally showed evidence of adequate reliability (r 2 ≥ 0.70) and criterion validity (r 2≥ 0.50) were sensitive to grade-level differences (Ritchey & Coker, 2013) and to growth (McMaster et al., 2009) in first through third grade.

While evidence supports the use of story prompts to measure writing in early elementary grades, research into sentence-writing CBM-W tasks was conducted to more precisely capture beginning writers’ skills. Curriculum-based measures of writing sentence writing involves writing single sentences in response to pictures (henceforth referred to as “picture–word prompts”). Like story prompts, picture–word prompts align with the text-generation and transcription components of the Simple View of Writing. The act of writing single sentences in a picture–word task instead of cohesive paragraphs in story prompts enables researchers and educators to scaffold or support young writers with lower text production skills and weaker self-regulation as participants are concentrating on producing smaller pieces of text. In terms of technical adequacy, picture– word has demonstrated evidence of adequate reliability (*r* 2≥ 0.70) and criterion validity (*r* 2≥ 0.60) in kindergarten through third grade, suggesting they could be used as indicators of early writing performance (Coker & Ritchey, 2010; Lembke, Deno, & Hall, 2003; McMaster, Du, et al. 2011; McMaster et al., 2009; Ritchey & Coker, 2014). Taken together, research evidence suggests that story prompts and picture–word tasks demonstrate evidence of technical adequacy to capture early writing performance in grades K-3. The available scoring methods for these tasks provide additional information about how to best assess early writing.

Scoring procedures

Scoring procedures typically used with CBM-W are words written (WW), words spelled correctly (WSC), correct word sequences (CWS; two adjacent words in a sentence that are spelled and used correctly in context; Videen, Marston, & Deno, 1982), and correct minus incorrect word sequences (C-IWS; Espin, Scierka, Skare, & Halverson, 1999). The WSC, CWS, and C-IWS scoring procedures align with the transcription component of the Simple View of Writing in that they capture some aspect of spelling skill. Words written, CWS, and C-IWS align with text-generation; they measure how much text has been written (WW) and whether that text is accurate in form and meaning (CWS, C-IWS).

In general, although all scoring procedures (WW, WSC, CWS, and C-IWS) have shown evidence of reliability (*r* 2≥ 0.70) and criterion validity (*r* 2≥ 0.60) when used to score story prompt and picture–word tasks, CWS consistently demonstrated the largest criterion validity coefficients in grades K-3 (*r* = 0.53–0.57, Coker & Ritchey, 2010; *r* = 0.45–0.63, McMaster et al., 2009; McMaster, Du, et al. 2011; *r* = 0.42–0.92, Lembke et al., 2003; Ritchey & Coker, 2013; *r* = 0.42–0.56, Parker, Tindal, & Hasbrouck, 1991; Ritchey & Coker, 2013; *r* = 0.41–0.71 McMaster & Campbell, 2008; Tindal & Parker, 1991). Words written, WSC, and CWS also demonstrated sensitivity to growth for use as progress monitoring tools in various early elementary studies (McMaster, Du et al., 2011; Parker, McMaster, Medhanie, & Silberglitt, 2011; Ritchey & Coker, 2013).

Summary and next steps

The findings reported here indicate promising evidence that story prompt and picture–word CBM-W tasks as well as a variety of scoring methods demonstrate technical adequacy for identification and monitoring of student writing performance. Both types of tasks allow for flexibility in assessing writers at different levels of performance. Scoring methods like WW and WSC can capture basic transcription and text-generation skills for very young or developing writers, whereas CWS and C-IWS capture writing complexity, word form and use, and errors. Because any scoring method can be used with either type of task, researchers and educators can customize a combination of tasks and scoring procedures to best capture the writing performance of young students who may be struggling with different skills. However, more research is needed into how to best use these assessments to look at performance and growth across multiple grades.

Despite existing CBM-W research that provides evidence of technical adequacy for a range of tasks and scoring procedures across elementary grade levels, further validation of these tasks for use across grade levels is needed (McMaster, Parker, & Jung, 2012). For example, no research has evaluated the technical adequacy of picture–word and story prompt measures in first through third grade in one sample. The most recent CBM-W research has been conducted with students in one or two grade levels at a time. Educators need to know what performance to expect within an academic year at a single grade level as well as across grade levels, requiring assessments that are flexible and usable at multiple grade levels (McMaster, Ritchey et al., 2011). There is also a need to identify which tasks and scoring methods are most appropriate across grade levels.

Current study

The picture–word and story prompt data in this manuscript were drawn from a large screening study that also included a word dictation CBM-W task measuring word-level spelling. The word dictation data were not included in this manuscript because the word dictation task underwent additional development over the course of the study, and that analysis warrants a separate paper. Ultimately, the picture–word and story prompt tasks were chosen for this article for three rea- sons. First, the picture–word and story prompt tasks have yielded promising technical adequacy data in the previous studies outlined above for young writers. Second, students in early grades are developing sentence- and text-level writing skills, indicating a need for measures that capture both sentence- and text-level writing. Third, the two tasks are similar to common writing tasks students are asked to complete in the classroom, such as open-ended journal responses and describing pictures or stories (Cutler & Graham, 2008). The alignment between CBM-W tasks and common classroom practice yields important social and face validity for teachers and instructional utility of the CBM-W scores (Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006). This study includes a larger and broader sample than previous studies to yield more generalizable information about the technical adequacy of picture–word and story prompt measures across grade levels by exploring two primary research questions:

1. Do picture–word and story prompt scores demonstrate evidence of reliability and criterion validity in first, second, and third grades?
2. Do picture–word and story prompt scores discriminate between students at different grade levels?

The data in this article are from two data collection “waves” under our stated research questions, which is the reason why they are presented in one manuscript. The first wave of data col- lection was conducted during one academic year in one school district (District 1). A replication was completed during a second academic year in a different school district in another state (District 2) with modifications to the criterion measure. For ease of analysis and interpretation, first we present the methodology used in both districts, then we present results from each district separately, and finally, we discuss conclusions and implications drawn from the entire study.

Method

Participants and setting

Total participants across both districts included 612 students in grades 1–3 in two Midwestern states. Data from District 1 (*n* = 338) were collected during the 2013–2014 academic year and data from District 2 (*n* = 274) were collected during the 2014–2015 academic year. Demographic data for the participants are summarized in Table 1. All teachers volunteered to participate.

District 1

District 1 was in a small city serving 16,990 K–12 students during the 2013–2014 academic year. Participants (*n* = 338) in District 1 included 96 first graders, 118 second graders, and 124 third graders across 27 classrooms in two schools within one school district in a small Midwestern city.

District 2

District 2 was a large, urban school district serving 35,400 K–12 students during the 2014–2015 academic year. Participants (*n* = 274) in District 2 included 94 first graders, 100 second graders, and 80 third graders across 18 classrooms in two schools within one school district in a large Midwestern city.

Curriculum-based measures of writing tasks

Two CBM-W tasks, picture–word and story prompt, were used to capture different writing skills and reflect the Simple View of Writing (Berninger et al., 2002).

Picture–word

Picture–word assesses transcription and text-generation at the sentence level and is group- administered for 3 min. Four alternate forms (A, B, C, D) of the picture–word task were developed for use in this study. A list of 48 distinct words was generated from previous research (McMaster et al., 2009) and from common objects and activities students would likely encounter through their school experiences (e.g. paper, walk, eat). The master list of 48 words was randomly grouped into 16 sets of three words. The three-word sets were then randomly assigned to either Form A, B, C, or D so that each form contained four single-sided pages with three words per page (12 words per form). Each word was paired with a corresponding picture of the word. Students were provided a packet with two *forms* of the picture–word task (e.g., Form A and Form B). Alternate forms were counterbalanced across classrooms.

Story prompt

Story prompt assesses text generation skills at the passage level. The task is group-administered for 3 min. Four alternate forms (A, B, C, D) of the story prompt task were developed for use in this study. The research team generated 16 narrative prompts cited in the previous research (e.g. McMaster et al., 2009) and from common school experiences (e.g. "One day, we were playing outside the school and … "). The prompts were reviewed by the research team to be as free of cultural bias as possible. Four total prompts were randomly selected from the master list and were randomly assigned to either form A, B, C, or D. Students were provided with a packet containing two story prompt *forms*. Each form contained one prompt with lines for composing a passage. Alternate forms were counterbalanced across classrooms.

Although the alternate test forms were counterbalanced across classrooms, we chose not to counterbalance the actual picture–words and writing prompts contained within each form. We wanted to ensure that if there was indeed an order effect that it was uniform so that results would still be interpretable for the types of analyses planned (see Data analysis section).

Curriculum-based measures of writing scoring procedures

The following scoring procedures were used with both picture–word and story prompt samples. *Words written* (WW) is the total number of words written where a "word" is a sequence of letters separated by a space from another sequence of letters (Parker et al., 1991). *Words spelled correctly* (WSC) is the number of correctly spelled English words in the sample regardless of context. *Correct word* *sequences* (CWS) is any two adjacent words that are correct in terms of spelling, grammar, capitalization, and punctuation in the context of a sentence (Parker et al., 1991; Videen et al., 1982). *Correct minus incorrect word sequences* (C-IWS) is the number of CWS minus the number of incorrect word sequences in a sample (Espin et al., 1999).

Criterion writing measures

To assess criterion-related validity, a standardized criterion assessment of writing was administered to a subsample of participants at each site during the spring of each year.

District 1

A subsample of participants (*n* = 150) were administered the Spelling and Sentence Composition subtests from the Weschler Individual Achievement Test-III (WIAT-III; Pearson, 2009). The WIAT-III is an individually administered comprehensive assessment of student academic achievement designed for children in grades Pre-K through 12 (Pearson, 2009). The reported validity coefficients for the subtests were *r* = .65–.66 (Breaux, 2009). The Spelling subtest requires stu- dents to correctly spell a series of dictated sounds or words (reliability: *r* = 0.94–0.95; Breaux, 2009). The Sentence Composition subtest includes both Sentence Building and Sentence Combining tasks. In Sentence Building, students write a sentence for a target word. For Sentence Combining, students must combine two or three target sentences into one sentence (reliability: *r* = .84–0.90; Breaux, 2009).

Prior to the spring CBM-W data collection, teachers from District 1 rank-ordered all students in their classrooms according to their judgment of each student's overall writing performance level. Based on these rankings, researchers administered the WIAT-III using standardized procedures to the middle 50 students in each grade to obtain criterion assessment data on an average performing sample.

District 2

Criterion validity results from District 1 using the WIAT-III were less than ideal (see District 1 Results below). The sample (middle 50 from each grade) produced a restricted range of scores and created limitations in the interpretability and generalizability of the data. Additionally, the criterion assessment provided limited information because the participants were assessed only on sentence construction and spelling and were not assessed on connected writing. Therefore, the research team decided to use a different criterion measure and a different subsampling procedure for the second wave of data collection in District 2 in the hopes of obtaining stronger criterion validity coefficients and to address the potential limitation of a restricted range of scores from using a narrow band of participants.

In District 2, a subsample of participants (*n* =94) were administered the Spelling, Writing Samples, and Sentence Writing Fluency subtests of the Woodcock–Johnson Tests of Achievement IV (WJ-IV; Schrank, Mather, & McGrew, 2014). The WJ-IV is an individually administered battery of achievement tests designed for ages 2–90 years (Schrank et al., 2014). In the Spelling subtest, students spell a series of dictated words (reliability: *r* = 0.91; Schrank et al., 2014). In the Writing Samples subtest, students are asked to write a variety of sentences according to given prompts. Items are scored for length, vocabulary, and sophistication (reliability: *r* = 0.90; Schrank et al., 2014). In the Sentence Writing Fluency subtest, students are asked to write simple sentences that contain three target words (reliability: *r* = 0.83; Schrank et al., 2014). The Broad Written Language cluster is an aggregate measure of the Spelling, Writing Samples, and Sentence Writing Fluency subtests (reliability: *r* = 0.95; Schrank et al., 2014). The Written Expression cluster is an aggregate measure of the Writing Samples and Sentence Writing Fluency subtests (reliability: *r* = 0.91; Schrank et al., 2014).

Prior to the spring CBM-W data collection, teachers in District 2 rank-ordered all students in their classrooms according to their judgment of each student's overall writing performance level. Based on these rankings, researchers identified a stratified sample of participants with high-, middle-, and lower-level writing performance within each grade level (*n* = 94; 29 first grade, 33 second grade, and 32 third grade) to obtain criterion assessment data from a more representative sample of students. Participants in this stratified sample were given the subtests from the WJ-IV using standardized procedures after completing all other CBM-W tasks.

General procedures

Curriculum-based measures of writing data collection occurred during the fall (Oct–Nov), winter (Jan), and spring (Apr) in both districts (District 1: 2013–14; District 2: 2014–15). During each data collection wave, participants completed two forms of picture–word and two forms of story prompt, both group-administered to whole classes of approximately 20–25 students by trained graduate research assistants and site project coordinators. For the picture–word task, students were instructed to, "Write one sentence for each picture in your packet." Students completed a practice item ("car") on the front of their test packets with the assessment administrators, and then followed along in the packet while the administrators read each word aloud. Students were then given 3 min to "write your best sentences." For the story prompt task, students were told, "I'm going to ask you to write a story. I will give you a story starter to give you ideas for your story. Before you write, I want you to think about your story. You will have 30 s to think and 3 min to write." Administrators then read the prompt aloud. The administration length for both tasks was drawn from the previous research, demonstrating evidence of technical adequacy of CBM-W tasks given for 3 min to early elementary grades (Coker & Ritchey, 2010; Deno, Mairkin et al., 1982; Marston & Deno, 1981; McMaster et al., 2009; Ritchey & Coker, 2014).

Each set of alternate forms for both tasks was administered successively with a short break in between for a repetition of directions. Students took picture–word first followed by story prompt during the same administration period. Each class took a different combination of two forms of each task at each time point (e.g. picture–word A, picture–word B; story prompt A, story prompt B). Form combinations were counterbalanced across classrooms within grades. All students were intended to complete Form A of each CBM-W task with a second form (B, C, or D) at each time point; however, due to a collating error, all participants in District 1 did not complete picture– word combination AB and instead received combination BC.

In spring, a subsample of participants also completed the criterion measure. Participants in the subsamples completed the subtests from the criterion measures after completing all other CBM- W tasks. All criterion measures were administered by trained graduate students and site project coordinators according to the standardized administration procedures.

Reliability procedures

Curriculum-based measures of writing administrators were research assistants with experience with CBM-W procedures. Test administrators completed a 1-hr small group training with the co- Principal Investigator (PI) or site project coordinator to establish fidelity of administration prior to data collection. All administrators scored at or above 90% fidelity of administration using a 10-item fidelity checklist after training. Team members administered all measures in pairs to informally ensure accurate administration. The same administrators were also trained by the co- PIs to score CBM-W assessments. Inter-rater agreement for scorers was completed on 14% of all writing samples across both sites and was above 90% at each site (District 1: picture–word = 97–99%, story prompt = 95–100%; District 2: picture–word = 94–100%, story prompt = 91–99%). Full inter-rater agreement results by individual scoring procedure are available upon request.

For the criterion measures, test administrators were graduate research assistants and a site project coordinator. Test administrators completed administration and scoring training by experts (a retired school psychologist with an educational specialist degree who was a project coordinator at District 1 and a nationally certified school psychologist at District 2). Administration fidelity was above 90% for all test administrators as measured by an observation checklist. Scoring inter-rater agreement was above 90% for all subtests except Writing Samples on the WJ-IV in District 2. A second round of scoring training on the

Writing Samples subtest was provided and all scorers' inter-rater agreement improved to >90%.

Classroom writing instruction

Although not a focus of this study, researchers collected general information about classroom writing instructional practices (e.g. writing tasks, student groupings, etc.) by teacher self-report. Teachers in both districts reported using a combination of skill-specific lessons teaching spelling and grammar and workshop-style activities where students planned and wrote stories with feed- back and edited their work. Observational data were not collected, which is a limitation of the study; however, teacher reports suggest little variation in writing instruction across classrooms or sites.

Data analysis

All analyses were completed on all data from each district. All data were entered into SPSS v. 24.0 for analysis and descriptive statistics (mean, *SD*, range, skewness, and kurtosis) were calculated for all measures. For alternate form reliability, a Pearson's *r* correlation coefficient was calculated between forms of each task (picture–word and story prompt) at each time point (fall/ winter/spring) by grade level. To determine criterion-related predictive and concurrent validity, a series of Pearson's *r* correlation coefficients were calculated. For predictive validity, the age- normed standardized scores of the spring criterion measures were correlated with the mean of the two fall forms of picture–word and story prompt. For concurrent validity, the age-normed spring criterion measures were correlated with the mean of the two spring forms of picture–word and story prompt. While we acknowledge that using the mean of two forms may have influenced validity results, we used the mean of two forms to obtain a more stable estimate of performance at each time point and to account for the one form combination difference between sites on the picture–word task (District 1 took forms BC and District 2 took forms AB in fall). To keep the analysis consistent and provide a more stable estimate of validity across tasks, the mean of both forms at each time point was also used for story prompt. To determine grade-level differences, a multivariate analysis of variance (MANOVA) was conducted. A Bonferroni correction was used (0.05/12 = 0.004) to control for Type I error and a Cohen's *d* statistic was calculated to measure the magnitude of the significant grade level differences. Cohen's *d* statistics were interpreted using Cohen's (1988) standards (small 0.2, medium 0.5, large 0.80, very large 2≥1.0). Due to space constraints, condensed data tables are presented; full results are available upon request.

Results: District 1

Descriptive data

In grades 1 and 2, participants on average scored higher at the spring time point compared to the fall time point on both picture–word (Table 2) and story prompt (Table 3) across all scoring methods. From fall to winter, mean WW and WSC scores on picture–word and story prompt dipped slightly or remained flat, and while CWS on story prompt also fell from fall to winter, the CWS scores on picture–word increased slightly. Mean C-IWS scores steadily increased on both tasks from fall to winter and winter to spring. The patterns for grade 3 are slightly different, however. While WW, WSC, and CWS on story prompt decreased from fall to winter and then increased from winter to spring, all scores on the picture–word task increased steadily from fall to winter and from winter to spring. Correct minus incorrect word sequences on both tasks steadily increased across all time points, similar to the pattern observed in grades 1 and 2. Overall, all participants increased their scores from fall to spring on both tasks using all scoring methods. Descriptive data for the criterion measure are summarized in Table 4.

Alternate form reliability

The criterion level for acceptable reliability in previous CBM-W studies has been *r* 2≥.70 (Lembke et al., 2003; McMaster et al., 2009). Previous research on standardized writing measures suggested that sufficient reliability estimates for screening in the area of writing should range from .70 to .90 (Shinn, 1989; Taylor, 2003). Coefficients that met the criterion level are bolded in Tables 5 and 6.

Picture-word alternate form reliability coefficients (Table 5) were above *r* = 0.70 and significant ($p \leq .01$) for nearly all forms and scoring procedures in every grade. The only coefficients that did not meet the criterion across time points were CWS (*r* = 0.60–0.61) and C-IWS (*r* = 0.27–0.69) in first grade. Story prompt had less evidence of reliability in first grade across all scores and forms at the fall and winter time points (Table 6); however, in spring, nearly all of the form combinations and scores reached the reliability criterion. In second grade, most coefficients met the criterion at each time point except for C-IWS (*r* = 0.53–0.66). In third grade, nearly all coefficients met criterion across time points; however, C-IWS showed weaker evidence of reliability (*r* = 0.46–0.65).

Criterion validity

In previous studies, researchers have established *r* 2≥ 0.50 (McMaster, Du et al., 2011; McMaster et al., 2009) as acceptable evidence of concurrent criterion validity to identify promising measures, accounting for the fact that writing measures historically have shown evidence of modest criterion validity (Shinn, 1989; Taylor, 2003). We adopted the same threshold for this study. Coefficients that met validity criterion are bolded in Tables 7 and 8.

Predictive validity

For picture–word (Table 7), in second grade, predictive validity coefficients for CWS and C-IWS met the validity criterion for the Sentence Composition subtest of the WIAT-III, but not the Spelling subtest. In third grade, CWS and C-IWS met criterion for the Spelling subtest but not the Sentence Composition subtest. For story prompt (Table 8), predictive validity coefficients met criterion with the Spelling subtest in first grade for C-IWS, in second grade for WSC, CWS, and C-IWS, and in third grade for CWS and C-IWS. For the Sentence Composition subtest, predictive validity coefficients met criterion only in second grade for C-IWS.

Concurrent validity

For picture–word (Table 7), coefficients for CWS and C-IWS met criterion with Sentence Composition in second grade and for CWS and C-IWS with the Spelling subtest in third grade. For story prompt (Table 8) C-IWS met criterion for the Spelling subtest in first grade. In second grade, WSC and CWS met criterion for the Spelling and Sentence Composition subtests. In third grade, all coefficients but WW met criterion for Spelling.

Grade-level differences

Using Wilks' lambda, significant differences were detected for picture–word between grade levels for all scoring procedures at all time points, *F*(8, 674) = 22.68, *p* ≤ .001 (Table 9). Follow-up univariate ANOVAs (available upon request) revealed significant differences between all grades for all scoring procedures at each time point (*p*≤ .001). Least squares difference (LSD) tests revealed statistically significant differences between all grade levels on all scoring procedures (*p* ≤.001) except between second and third grade on WW in fall. Effect sizes were large (*d* 2≥ 0.80) or very large (*d* 2≥ 1.0; Cohen, 1988) between grades 1 and 2 (*d* = 0.83–1.42) and between grades 1 and 3 (*d* = 1.26–2.1) at all time points and on nearly all scoring procedures. Effect sizes for the difference between grades 2 and 3 ranged from small (*d* = 0.33) to medium (*d*= 0.67) at all time points and scoring methods. For story prompt (Table 9), using Wilks' lambda, significant differences were detected between grade levels for all scoring procedures at each time point *F*(8, 674) = 19.13, *p* ≤ .001. Follow-up univariate ANOVAs revealed significant grade-level differences for all scoring procedures at all time points. Least square difference tests revealed significant differences between all grade levels on nearly all scoring procedures at all time points (*p* ≤ .001). Large effect sizes were found between grades 1 and 2 in fall and winter (*d* 2≥ 0.80) and were small to medium (*d*=0.36–0.072) in spring. Effect sizes were large to very large between grades 1 and 3 (*d* 2≥ 1.0) and small to medium (*d* =0.33–0.76) between grades 2 and 3 at all time points and on all scoring procedures.

Results: District 2

Descriptive data

In grades 1 and 2, scores on all scoring methods for both picture–word (Table 2) and story prompt tasks (Table 3) increased from fall to winter and winter to spring. However, in grade 3, while all scores on both tasks increased from fall to winter, scores stagnated or decreased slightly from winter to spring on all scores on both types of CBM-W tasks. Criterion measure descriptive data are summarized in Table 4.

Alternate form reliability

Coefficients that met the 0.70 criterion level (cited under District 1 Results) are bolded in Tables 5 and 6. Most alternate form reliability coefficients of picture–word (Table 5) were above *r* = .70 across scoring procedures at each time point (*p* ≤ .01) in all grades. Coefficients that did not meet the .70 criterion were in first and second grade in fall (WW, WSC, CWS, C-IWS across various forms), but in winter and spring at these same grade levels nearly all coefficients met criterion. All forms and scoring procedures met criterion in third grade at all time points. For story prompt (Table 6), reliability coefficients met criterion more consistently in winter and spring across scoring procedures and grades, but were more variable in fall across grades. In fall, the scoring procedures that met criterion across form combinations were WW in first grade, and CWS and C-IWS in third grade. In winter and spring, nearly all coefficients met criterion across scoring procedures and grades.

Criterion validity

Coefficients that met validity criterion (*r* 2 ≥ 0.50; see citations from District 1 Results) are bolded in Tables 7 and 8.

Predictive validity

For picture–word (Table 7), predictive validity coefficients for C-IWS met the validity criterion for the Spelling subtest and Broad Written Language cluster of the WJ-IV in first grade. In second grade, WSC, CWS, and C-IWS met criterion for the Spelling subtest and Broad Written Language cluster, and CWS and C-

IWS met criterion for the Sentence Writing Fluency and Writing Samples subtests and the Written Expression cluster. In third grade, predictive validity coefficients met the criterion across scoring procedures for all subtests and clusters. The overall pattern of predictive validity coefficients for story prompt (Table 8) was similar to picture–word. Predictive validity coefficients met criterion for CWS and C-IWS in first grade for all subtests and clusters, except Writing Samples. In second grade, WSC, CWS, and C-IWS met criterion on the Spelling and Written Expression subtests and the Broad Written Language cluster, and C- IWS on Sentence Writing Fluency subtest. At third grade, predictive validity coefficients met criterion across most scoring procedures for all subtests and clusters.

Concurrent validity

For picture–word (Table 7), in second grade, WSC and C-IWS demonstrated acceptable validity coefficients with the Sentence Writing Fluency subtest, and C-IWS also demonstrated acceptable validity coefficients with the Spelling subtest and Broad Written Language cluster. Most coefficients were acceptable at third grade across scoring procedures for all subtests and cluster scores. Story prompt showed a similar pattern as picture–word (Table 8). In second grade, C-IWS met the validity criterion for the Spelling and Sentence Writing Fluency subtests and the Broad Written Language cluster. Most coefficients met the criterion in third grade across scoring procedures for all subtests and cluster scores.

Grade-level differences

For picture–word (Table 9), using Wilks' lambda, statistically significant grade-level differences were found on all scoring procedures across all time points *F*(8, 474) =12.98, *p* ≤ .001. Follow- up univariate ANOVAs showed that each scoring procedure was statistically significantly different between grades (*p* ≤ .001). Least square difference post-hoc tests showed that third graders achieved statistically higher mean scores than first and second graders, and second-grade students showed statistically higher mean scores than first-grade students across scoring procedures and across time points, except for mean score differences between second- and third-grade students in spring. Large to very large effect sizes were found between grades 1 and 2 (*d* = 0.87–1.21) and between grades 1 and 3 (*d* = 0.95–1.61) for all scoring procedures across all time points. Between grades 2 and 3, medium effect sizes (*d* = 0.43–0.70) were found for fall and winter across scoring procedures. Spring effect sizes were negligible (*d* =–0.03).

For story prompt, using Wilks' lambda, a statistically significant grade level difference was found for all scoring procedures across all time points *F*(8, 470) = 17.27, *p* ≤ .001. Follow-up univariate ANOVAs showed statistically significant differences between grades for all scoring procedures (*p* ≤ .001). Least square difference post-hoc tests showed that third graders achieved statistically higher mean scores than first- and second-grade students, and second-grade students showed statistically higher mean scores than first-grade students across scoring procedures and across seasons, except for mean score differences between second- and third-grade students in spring (*p* = .31 for WW, *p* = .10 for WSC). Large to very large effect sizes were found between grades 1 and 2 (*d* = 0.94–1.36) and between grades 1 and 3 (*d* = 1.17–1.79) for all scoring procedures across all time points. Effect sizes between grades 2 and 3 were medium to large (*d* = 0.53–0.85) in fall and winter and small in spring (*d*=0.19–0.47) for all scoring procedures.

Discussion

Alternate form reliability

Results indicate that generally, while both picture–word and story prompt demonstrate sufficient alternate form reliability in grades 1–3 with all scoring methods, the largest coefficients were found in grades 2–3, which is consistent with the previous research (Deno et al., 1980; McMaster & Campbell, 2008). Many of the reliability coefficients that did not meet the .70 criterion were from data collected in fall and winter of first grade in both districts, suggesting that the tasks might produce less reliable data in first grade or with very young or inexperienced writers. While the weaker reliability evidence in fall and winter of first grade is seemingly at odds with other first-grade studies (McMaster et al., 2009; McMaster & Du et al., 2011), upon closer inspection, the current results provide new information. The previous studies obtained first-grade reliability data in the spring only, whereas this study measured CBM-W performance at three points during the year. When comparing spring alternate form reliability coefficients across studies, our results do reflect past findings that picture–word and story prompt tasks demonstrate adequate evidence of reliability in first grade (McMaster et al., 2009; McMaster & Du et al., 2011). The weaker evidence for reliability in fall and winter with CWS and C-IWS in particular in this study could be an indication of floor effects. Early in the academic year, first graders are still developing basic writing skills and the CWS and C-IWS scoring methods in particular may not be appropriate to capture very early writing development at that point in time. It may be more appropriate to use WW and WSC early in first grade and transition to using CWS and C-IWS later in the year. It is also possible that the less consistent reliability coefficients in fall and winter of first grade were due to the construction of a specific form rather than the picture–word task as a whole being unreliable. For example, forms AC in fall of first grade did not meet the .70 criterion with CWS and C-IWS at either site. It is also possible that CBM-W tasks that require sentence or passage writing are too difficult for young writers early in the academic year and therefore are not a good representation of their writing abilities.

Additionally, fewer studies have examined the C-IWS scoring procedure with young writers compared to other scoring procedures. Results from this study suggest that C-IWS may yield less reliable scores in first grade, particularly in the fall, compared to later time points and later grades. In other words, the C-IWS procedure has stronger evidence of reliability in second and third grades compared to first, but less evidence of reliability overall compared to the other scoring procedures used. This outcome is consistent with limited research that has included C-IWS with young students (McMaster et al., 2009; McMaster, Du et al., 2011). Several scoring procedures, particularly CWS, may be used to reliably capture early writing performance of students who can write at the sentence and passage levels of language. Combined, there is converging evidence across studies to suggest that picture–word and story prompt measures may demonstrate sufficient evidence of alternate form reliability with young writers.

Criterion validity

The validity results imply that broadly, across two different criterion measures, the picture–word and story prompt tasks show the strongest evidence of validity in third grade. This finding indicates that picture–word and story prompt may be a more accurate representation of writing proficiency for older students compared to younger students, which reflects some previous research (Deno et al., 1980; McMaster & Campbell, 2008) but seems to conflict with findings from other first-grade studies (McMaster et al., 2009; McMaster, Du et al., 2011). However, there are several points worth noting. In McMaster et al. (2009), the criterion measures were not, in fact, standardized writing tests, but were teacher ratings of student performance and a district writing rubric. Additionally, tasks that met validity criterion were administered for 5 min; the 3-min administration did not meet validity criterion for any of the criterion measures. Therefore, it could be that the validity results in McMaster et al. (2009) were influenced by the relatively subjective nature of rating scales and trait-based writing rubrics and a longer administration time.

Also, the current validity results may have been influenced by the way the extended writing subtests on the criterion measures were scored. On the Sentence Composition subtest of the WIAT-III and the Writing Samples subtest of the WJ-IV, part of a student's score depends on whether they used given words correctly or responded to a prompt appropriately, while the scoring procedures used with the CBM-W tasks do not score whether the target word in picture– word or story starter in story prompt were actually used. It could also be that the scoring procedures used on the CBM-W tasks are more heavily influenced by spelling than the scores obtained on the criterion measures, which would also explain the stronger relation between the Spelling subtests on both criterion measures and the CBM-W tasks across grades. Finally, using the mean of two forms for the validity analyses to account for the form combination differences across sites (see Measures) could have resulted in higher coefficients in this study, and validity results may not be generalizable to using single forms. Further research is needed with picture–

word and story prompt in grade 1 to investigate criterion validity evidence.

Grade-level differences

Across the two districts, scores on picture–word and story prompt demonstrated significant differences between grade levels for nearly all scoring procedures at each time point. Given that significant grade level differences in CBM-W scores provides one way to infer each measure's sensitivity to differences across grades and skill levels (Ritchey & Coker, 2013), findings of this study indicate that picture–word and story prompt are sensitive to writing skill development across grades 1–3. On both tasks, older students generally outperformed younger students, which is logical considering older students are generally more developed and have more experience writing. Overall, the picture–word task appeared to distinguish between first- and second-graders' performance in writing throughout the academic year and the story prompt task distinguished between second- and third-graders' performance in the beginning and middle of the academic year. Taken together, these results indicate that picture–word captured some essential component or difference in writing performance that develops throughout first and second grade that becomes diminished between second and third grades, while story prompt seems to be better at capturing grade-level differences after students have had more experience writing. The previous studies also concluded that the picture–word task was more appropriate for detecting growth in first grade (McMaster, Du et al., 2011; Parker et al., 2011) as compared to the story prompt task detecting differences from second to third grade (Ritchey & Coker, 2013).

In terms of scoring, the largest differences between grades 1 and 2 were found with the WW and WSC scoring procedures *on both tasks*. The reverse is true for grade-level differences between grades 2 and 3; CWS and C-IWS had the largest effect sizes *on both tasks*. The results here suggest that in first grade and into second grade, the amount of text and correctly spelled words students produce is enough to detect performance differences between grades, but as students mature and grow as writers, tasks and scoring procedures need to be more sensitive to the complexities and sophistication of writing to capture growth and grade-level differences, such as the CWS and C-IWS scoring procedures. The differences in scoring procedures are reflective of past research finding that scoring procedures such as WSC had stronger evidence of technical adequacy in early compared to later elementary grades (McMaster, Du et al., 2011). These grade level difference results also reflect our reliability results and previous research that found C-IWS was more reliable with second and third grades compared to first (McMaster et al., 2009; McMaster, Du et al., 2011). Overall, the grade-level differences suggest that picture–word and story prompt tasks can capture grade-level differences in writing for grades 1 through 3.

When taken as a whole, the results from this study suggest that educators can use picture– word and story prompt tasks, particularly with the CWS and C-IWS scoring procedures, to reliably identify students at risk for writing difficulty in second and third grade. First-grade teachers should use caution when using these tasks, particularly early in the academic year, as their students may not have developed the skills necessary for connected writing and they may be misidentified as struggling. In terms of scoring procedures, findings suggest CWS may be more appropriate in grades 2 and 3, and first-grade teachers should consider using WW and WSC, given young students' relative lack of experience and proficiency in writing.

Contextual factors

It is possible that contextual factors, rather than the actual measures, influenced results or cross- site differences. The participating school districts *as a whole* were somewhat different in terms of demographics. Overall, District 2 was larger, had a more even distribution of racial and ethnic groups, had a higher rate of free/reduced lunch eligibility (a common proxy for low income back- ground), and had more students receiving special education services than District 1. However, the *actual* participant samples from each district were relatively similar. Both samples were similar in terms of racial and ethnic makeup. One difference between the samples was the sample from District 1 had a larger percentage of participants eligible for free/reduced lunch (50 vs. 40% in District 2), and District 2 had a larger percentage of students receiving special education services (11 vs. 9% in District 1); however, it is not clear whether these differences influenced results, and analysis of these potential differences is outside the scope of this study. At face value, the demo- graphic differences do not seem to be large enough to have had a significant impact on the study results, although this is a limitation that could be addressed in future iterations of this work.

Although it is possible that differences in classroom writing instruction could have impacted the screening results, it is not clear whether this is true for this study. At this time, a deeper analysis of the writing instruction in participating classrooms is outside the scope of this study, but future work could incorporate an observational measure of writing instruction and account for instructional differences on students' writing performance on CBM-W tasks.

Implications

When interpreting all results in the context of the Simple View of Writing (Berninger et al., 2002), it appears that two of the key facets in the Simple View writing model, transcription and text-generation, are important aspects of young students' writing performance and can be captured by picture–word and story prompt CBM tasks in first through third grade, and each task shows differences in performance across grades. Although reliability and validity estimates improve as grade level increases on both tasks, it may be more advantageous in practice to use a more progressive model of writing assessment. In other words, to more accurately and consistently capture early writing skills, it may be advisable to use varying combinations of CBM-W tasks and scoring methods depending on the grade and time of year. Measures of transcription and production, such as WW and WSC with the picture–word task, may be more appropriate early in the academic year in first and second grades, whereas the more complex CWS and C- IWS scoring procedures with either CBM-W task in late second and third grade may be more accurate at capturing more advanced skills. Although self-regulation was not explicitly measured by the CBM tasks in this study, it would be worth considering in future research and practice how educators may want to tap into or reflect young students' functioning in self-regulation and working memory and how these cognitive processes impact writing performance.

Limitations and future research

In general, results of this study are limited by moderate-sized samples of students from two mid- to large-sized school districts. Whether results generalize to students in other types of settings is unknown and warrants additional examination. While demographic data were collected and some differences across sites were noted, it is not clear whether these differences influenced results. Additionally, while information about classroom writing instruction were obtained, this information was gained by teacher self-report and writing instruction was not directly observed. Future studies should account for and model the impact of contextual factors on CBM-W performance. Furthermore, all CBM-W scoring was completed by researchers and trained graduate students. The extent to which outcomes would be similar if assessments were scored by teachers should be examined to infer the feasibility of CBM-W as a practical, accurate assessment for young writers in schools. In addition, CBM-W administration periods across fall, winter, and spring at each site were not identical and may have caused variation in outcomes. The use of two different writing criterion measures and a potentially restricted range of criterion scores in District 1 was also a limitation that restricts cross-site comparisons; future research should account for such discrepancies across settings to facilitate more comprehensive inferences regarding the technical adequacy of picture–word and story prompt with young writers. Additionally, the criterion measures used did not measure text-level writing. Future studies should use a criterion measure with a text-level writing component normed on early elementary grade students to more precisely pinpoint the validity evidence of the CBM-W tasks and to obtain a fuller picture of students' writing ability.

In terms of data analysis, future studies should include test–retest and internal consistency reliability to further examine the utility of CBM-W as a progress monitoring tool for young writers. It is also important to continue to examine which CBM-W measures have the strongest validity evidence for students at each grade level (McMaster & Espin, 2007; McMaster, Ritchey et al., 2011). Additionally, the technical adequacy of the C-IWS scoring procedure should further be investigated. Across the two sites and grade levels of this study, C-IWS consistently showed acceptable levels of predictive and concurrent validity but appeared to yield insufficiently reliable scores for young writers, therefore further research on C-IWS is warranted.

Finally, for CBM-W to be used to screen students who may need intensive writing intervention, the classification accuracy (e.g., sensitivity and specificity) of the measures should be deter- mined for each grade level (McMaster, Du et al., 2011). For picture–word and story prompt tasks to be used for regular progress monitoring, additional “stage 2” research is needed (Fuchs, 2004), which examines the validity of slopes derived from scores from repeated CBM-W administrations (McMaster, Du et al., 2011). Although these analyses were considered beyond the scope of this study, in future iterations of this work, sensitivity to growth of CBM-W tasks for multiple scoring procedures should be examined.

Conclusion

This study examined the technical adequacy of picture–word and story prompt CBM-W tasks with writers in grades 1–3. This study replicates and extends prior CBM-W research by including larger numbers of young writers and examining the reliability, validity, and grade-level differences of CBM-W across multiple scoring procedures and time points. Study results indicated that both CBM-W tasks can be used in grades 2 and 3, whereas more research is necessary in first grade. These findings are important to the development of systems for screening and progress monitoring that are psychometrically sound, feasible, and ultimately result in meaningful decision making that improves educational outcomes for young writers.

Table 1. Demographic data.

	District 1 (*n* = 338) *n* (%)	District 2 (*n* = 274) *n* (%)	All (*n* = 612) *n* (%)
Grade			
First	96 (28.4)	94 (34.3)	190 (31.0)
Second	118 (34.9)	100 (36.5)	218 (35.6)
Third	124 (36.7)	80 (29.2)	204 (33.3)
Gendera			
Male	166 (49.1)	140 (51.1)	306 (50.0)
Female	172 (50.9)	132 (48.2)	304 (49.7)
Ethnicitya			
White	215 (63.6)	169 (61.7)	384 (62.7)
African American	88 (26.0)	63 (23.0)	151 (24.7)
Hispanic	13 (3.9)	19 (6.9)	32 (5.2)
Asian/Pacific Islander	7 (2.1)	18 (6.6)	25 (4.1)
American Indian	—	3 (1.1)	3 (.5)
Multi-racial	15 (4.4)	—	15 (2.5)
SESa			
Free/Reduced	181 (53.6)	112 (40.9)	293 (47.9)
Paid	157 (46.4)	160 (58.4)	317 (51.8)
Special Education[a]			
Yes	30 (8.9)	31(11.3)	61 (10.0)
No	303 (89.6)	241 (88.0)	544 (88.9)
504	5 (1.5)	NC	5 (.8)
Gifted Education			
Yes	28 (8.3)	NC	28 (4.6)
No	310 (91.7)	NC	310 (50.7)
ELL[a]			
Yes	7 (2.1)	23 (8.4)	30 (4.9)
No	331 (97.9)	249 (90.9)	580 (94.8)

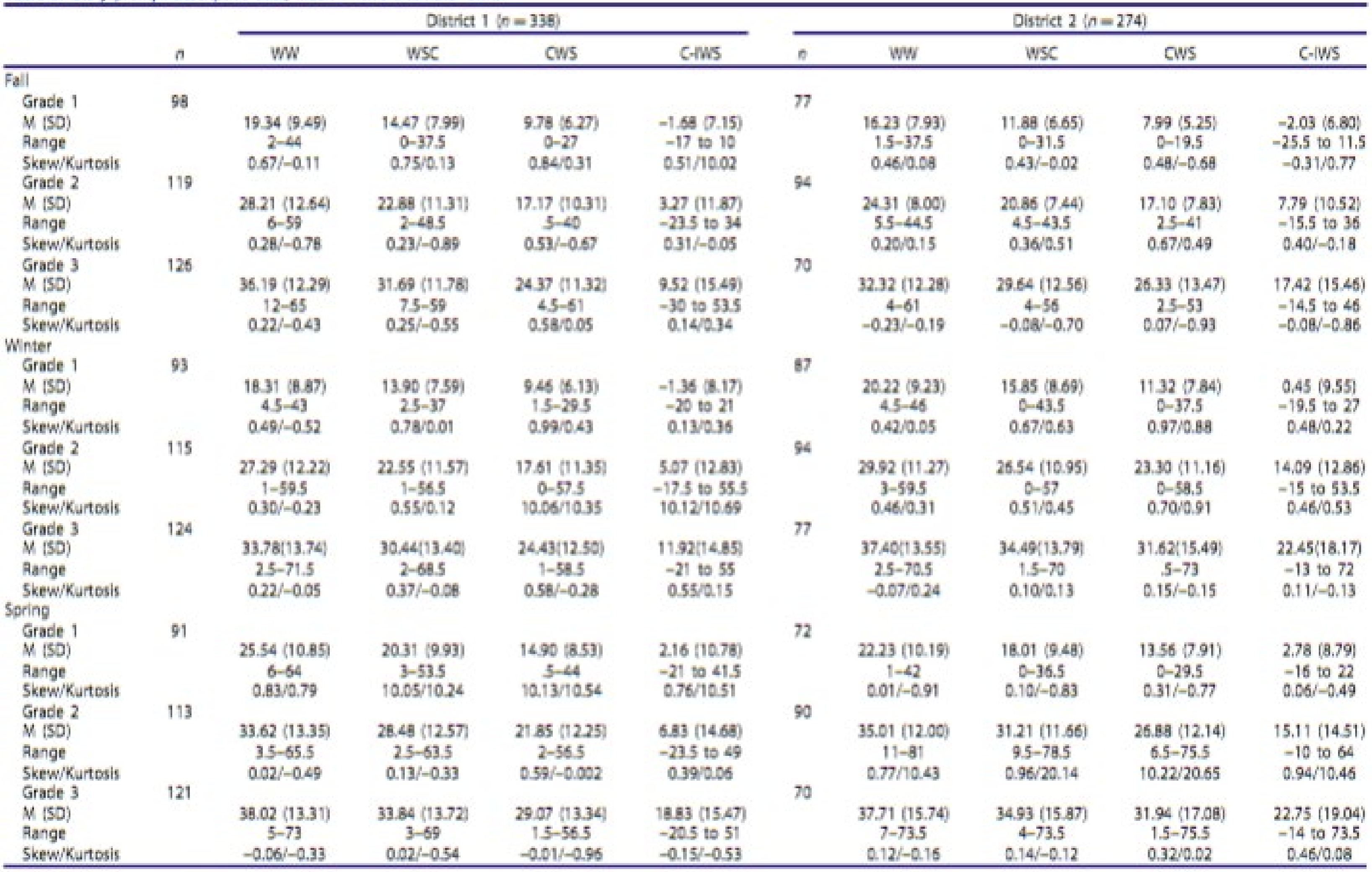
Note: [a]Percentages may not add to 100 due to missing demographic data for two students from District 1; NC: not collected.

Table 2. Picture–word descriptive data, mean of two alternate forms.

		District 1 (*n* = 338)					District 2 (*n* = 274)			
	n	WW	WSC	CWS	C-IWS	*n*	WW	WSC	CWS	C-IWS
Fall										
Grade 1	98					80				
M(SD)		21.05 (8.99)	17.05 (8.08)	13.50 (8.40)	1.55 (9.27)		18.22 (8.26)	15.35 (8.06)	13.66 (8.42)	5.25 (9.45)
Range		3–47	2.5–39	0–40.5	−18.5 to 35		2–44	0–41	0–37.5	−20.5 to 26.5
Skew/Kurtosis		0.46/0.11	0.71/0.22	1.32/1.65	1.01/2.44		0.73/10.22	0.89/10.59	0.74/0.36	0.25/−0.41
Grade 2	119					93				
M(SD)		35.61 (12.66)	31.10 (12.35)	27.54 (13.23)	12.65 (16.39)		25.98 (7.88)	23.54 (7.93)	23.21 (9.25)	15.65 (11.77)
Range		9–71	7.5–64	2.5–63	−21 to 54.5		8–48	3.5–46.5	3–52	−15.5 to 48
Skew/Kurtosis		0.35/−0.02	0.34/−0.17	0.25/−0.44	0.12/−0.70		0.20/0.36	0.23/0.73	0.48/0.67	0.27/0.40
Grade 3	126					70				
M (SD)		39.71 (12.30)	36.69 (12.13)	35.29 (13.85)	23.51 (17.02)		30.90 (10.52)	29.45 (10.91)	30.84 (13.09)	25.12 (14.93)
Range		12–73.5	9.5–72	6.5–75.5	−13.5 to 73		4–58	3.5–57.5	3.5–68.5	−5 to 67
Skew/Kurtosis		0.19/−0.01	0.37/0.21	0.58/0.28	0.43/0.07		0.14/−0.04	0.21/−0.13	0.33/0.02	0.19/0.12
Winter										
Grade 1	94					87				
M (SD)		19.77 (10.93)	16.02 (9.91)	13.61 (10.23)	3.07 (10.33)		24.45 (10.45)	21.53 (10.46)	19.87 (12.52)	10.11 (14.74)
Range		0–52.5	0–49	0–44.5	−22.5 to 37.5		3–51.5	.5–50.5	0–52	−20 to 46.5
Skew/Kurtosis		0.64/0.20	0.86/0.61	10.07/0.80	0.64/0.73		0.20/−0.31	0.37/−0.07	0.53/−0.29	0.38/−0.57
Grade 2	116					93				
M (SD)		35.66 (12.47)	31.70 (12.12)	31.07 (14.10)	18.72 (17.32)		34.06 (11.47)	31.90 (11.66)	33.48 (13.97)	26.17 (16.07)
Range		0–62.5	0–61.5	0–68	−11 to 62.5		6.5–66.5	1–65.5	.5–75	−14.5 to 72
Skew/Kurtosis		−0.25/−0.13	−0.06/−0.12	0.27/−0.22	0.49/−0.52		0.08/−0.13	0.14/−0.01	0.34/−0.09	0.14/0.05
Grade 3	124					78				
M (SD)		40.47 (13.35)	39.32 (12.73)	39.68 (14.27)	29.65 (16.70)		40.85 (12.32)	39.33 (12.51)	41.89 (14.73)	35.19 (17.05)
Range		3.5–69	2.5–67	1.5–71.5	−13 to 68		10.5–77.5	9–75	4.5–82	−3.5 to 80.5
Skew/Kurtosis		−0.29/−0.20	−0.09/−0.08	−0.07/−0.47	0.03/−0.52		0.02/0.90	0.03/0.70	−0.12/0.62	−0.18/0.02
Spring										
Grade 1	91					73				
M(SD)		31.13 (11.08)	26.91 (11.21)	25.59 (14.14)	13.17 (17.16)		28.35 (11.13)	25.40 (10.65)	24.93 (12.69)	15.26 (14.41)
Range		12.5–64.5	9.5–60.5	5.5–60	−25.5 to 56.5		1–48	0–47.5	0–55.5	−18.5 to 51.5
Skew/Kurtosis		0.60/0.01	0.68/−0.10	0.79/−0.23	0.45/0.05		−0.17/−0.79	−0.03/−0.43	0.31/0.01	0.39/0.01
Grade 2	113					91				
M(SD)		41.43 (10.62)	37.82 (10.57)	37.66 (13.68)	24.53 (18.31)		40.82 (11.19)	38.63 (11.28)	40.96 (13.87)	32.97 (16.14)
Range		13–67	9–62	6–67	−24.5 to 60.5		0–76.5	0–75.5	0–81.5	−15.5 to 75
Skew/Kurtosis		−0.60/0.15	−0.47/0.07	−0.06/−0.54	0.05/−0.54		−0.19/10.60	−0.13/10.34	−0.02/0.41	−0.10/0.23
Grade 3	120					72				
M(SD)		45.48 (11.67)	43.09 (11.95)	44.74 (14.33)	35.42 (17.75)		40.43 (14.02)	38.61 (13.87)	40.74 (16.40)	33.46 (18.60)
Range		9–73	5.5–70.5	3.5–79	−7.5 to 75.5		13.5–75.5	12.5–73	4–80	−16 to 74
Skew/Kurtosis		−0.25/0.49	−0.25/0.37	−0.16/0.09	−0.20/−0.17		−0.04/−0.52	−0.02/−0.57	0.02/−0.50	−0.22/0.18

Note: WW: words written; WSC: words spelled correctly; CWS: correct word sequences; C-IWS: correct minus incorrect word sequences.

Table 3. Story prompt descriptive data, mean of two alternate forms.



		District 1 (*n* = 338)					District 2 (*n* = 274)			
	n	WW	WSC	CWS	C-IWS	*n*	WW	WSC	CWS	C-IWS
Fall										
Grade 1	98					77				
M (SD)		19.34 (9.49)	14.47 (7.99)	9.78 (6.27)	−1.68 (7.15)		16.23 (7.93)	11.88 (6.65)	7.99 (5.25)	−2.03 (6.80)
Range		2–44	0–37.5	0–27	−17 to 10		1.5–37.5	0–31.5	0–19.5	−25.5 to 11.5
Skew/Kurtosis		0.67/−0.11	0.75/0.13	0.84/0.31	0.51/10.02		0.46/0.08	0.43/−0.02	0.48/−0.68	−0.31/0.77
Grade 2	119					94				
M (SD)		28.21 (12.64)	22.88 (11.31)	17.17 (10.31)	3.27 (11.87)		24.31 (8.00)	20.86 (7.44)	17.10 (7.83)	7.79 (10.52)
Range		6–59	2–48.5	.5–40	−23.5 to 34		5.5–44.5	4.5–43.5	2.5–41	−15.5 to 36
Skew/Kurtosis		0.28/−0.78	0.23/−0.89	0.53/−0.67	0.31/−0.05		0.20/0.15	0.36/0.51	0.67/0.49	0.40/−0.18
Grade 3	126					70				
M (SD)		36.19 (12.29)	31.69 (11.78)	24.37 (11.32)	9.52 (15.49)		32.32 (12.28)	29.64 (12.56)	26.33 (13.47)	17.42 (15.46)
Range		12–65	7.5–59	4.5–61	−30 to 53.5		4–61	4–56	2.5–53	−14.5 to 46
Skew/Kurtosis		0.22/−0.43	0.25/−0.55	0.58/0.05	0.14/0.34		−0.23/−0.19	−0.08/−0.70	0.07/−0.93	−0.08/−0.86
Winter										
Grade 1	93					87				
M (SD)		18.31 (8.87)	13.90 (7.59)	9.46 (6.13)	−1.36 (8.17)		20.22 (9.23)	15.85 (8.69)	11.32 (7.84)	0.45 (9.55)
Range		4.5–43	2.5–37	1.5–29.5	−20 to 21		4.5–46	0–43.5	0–37.5	−19.5 to 27
Skew/Kurtosis		0.49/−0.52	0.78/0.01	0.99/0.43	0.13/0.36		0.42/0.05	0.67/0.63	0.97/0.88	0.48/0.22
Grade 2	115					94				
M (SD)		27.29 (12.22)	22.55 (11.57)	17.61 (11.35)	5.07 (12.83)		29.92 (11.27)	26.54 (10.95)	23.30 (11.16)	14.09 (12.86)
Range		1–59.5	1–56.5	0–57.5	−17.5 to 55.5		3–59.5	0–57	0–58.5	−15 to 53.5
Skew/Kurtosis		0.30/−0.23	0.55/0.12	10.06/10.35	10.12/10.69		0.46/0.31	0.51/0.45	0.70/0.91	0.46/0.53
Grade 3	124					77				
M (SD)		33.78(13.74)	30.44(13.40)	24.43(12.50)	11.92(14.85)		37.40(13.55)	34.49(13.79)	31.62(15.49)	22.45(18.17)
Range		2.5–71.5	2–68.5	1–58.5	−21 to 55		2.5–70.5	1.5–70	.5–73	−13 to 72
Skew/Kurtosis		0.22/−0.05	0.37/−0.08	0.58/−0.28	0.55/0.15		−0.07/0.24	0.10/0.13	0.15/−0.15	0.11/−0.13
Spring										
Grade 1	91					72				
M (SD)		25.54 (10.85)	20.31 (9.93)	14.90 (8.53)	2.16 (10.78)		22.23 (10.19)	18.01 (9.48)	13.56 (7.91)	2.78 (8.79)
Range		6–64	3–53.5	.5–44	−21 to 41.5		1–42	0–36.5	0–29.5	−16 to 22
Skew/Kurtosis		0.83/0.79	10.05/10.24	10.13/10.54	0.76/10.51		0.01/−0.91	0.10/−0.83	0.31/−0.77	0.06/−0.49
Grade 2	113					90				
M (SD)		33.62 (13.35)	28.48 (12.57)	21.85 (12.25)	6.83 (14.68)		35.01 (12.00)	31.21 (11.66)	26.88 (12.14)	15.11 (14.51)
Range		3.5–65.5	2.5–63.5	2–56.5	−23.5 to 49		11–81	9.5–78.5	6.5–75.5	−10 to 64
Skew/Kurtosis		0.02/−0.49	0.13/−0.33	0.59/−0.002	0.39/0.06		0.77/10.43	0.96/20.14	10.22/20.65	0.94/10.46
Grade 3	121					70				
M (SD)		38.02 (13.31)	33.84 (13.72)	29.07 (13.34)	18.83 (15.47)		37.71 (15.74)	34.93 (15.87)	31.94 (17.08)	22.75 (19.04)
Range		5–73	3–69	1.5–56.5	−20.5 to 51		7–73.5	4–73.5	1.5–75.5	−14 to 73.5
Skew/Kurtosis		−0.06/−0.33	0.02/−0.54	−0.01/−0.96	−0.15/−0.53		0.12/−0.16	0.14/−0.12	0.32/0.02	0.46/0.08

Note: WW: words written; WSC: words spelled correctly; CWS: correct word sequences; C-IWS: correct minus incorrect word sequences.

Table 4. Criterion measures descriptive data.

	District 1 (WIAT-III; *n* = 150)			District 2 (WJ-IV; *n* = 94)					
	n	Spelling	Sentence composition	*n*	Spelling	Sentence writing fluency	Writing samples	Broad written language	Written expression
Grade 1	50			29					
M (SD)		103.20 (10.65)	100.00 (15.78)		104.24 (14.01)	109.00 (12.37)	101.83 (15.37)	106.14 (13.89)	106.76 (13.47)
Range		87–126	69–130		74–139	81–132	66–130	77–137	80–132
Skewness		0.58	0.07		0.20	−0.18	−0.62	−0.001	−0.18
Kurtosis		−0.49	−0.68		0.81	−0.34	0.15	−0.18	−0.59
Grade 2	50			33					
M (SD)		94.24 (12.86)	94.58 (13.39)		108.27 (16.84)	107.79 (10.39)	108.46 (13.49)	109.79 (13.42)	109.39 (11.67)
Range		64–119	69–128		74–141	89–136	68–125	80–132	85–133
Skewness		−0.15	0.12		−0.29	0.48	−0.98	−0.47	−0.35
Kurtosis		−0.29	−0.30		−0.32	0.35	10.00	−0.49	−0.54
Grade 3	50			32					
M (SD)		94.48 (10.58)	95.56 (12.95)		103.56 (19.56)	107.00 (9.85)	103.44 (15.22)	106.13 (15.55)	106.78 (13.03)
Range		63–115	60–121		64–133	87–129	63–127	69–127	75–127
Skewness		0.03	−0.73		−0.36	−0.49	−0.85	−0.65	−0.72
Kurtosis		0.48	0.61		−0.60	0.21	0.25	−0.37	0.14

Note: WIAT-III: Wechsler individual achievement test; WJ-IV: Woodcock–Johnson tests of achievement IV.

Table 5. Picture-word alternate form reliability coefficients.

	District 1 ($n = 338$)				District 2 ($n = 274$)			
	WW	WSC	CWS	C-IWS	WW	WSC	CWS	C-IWS
Fall								
Grade 1								
AB[a]	–	–	–	–	**.88****	**.88****	**.84****	**.75****
AC	**.70****	**.73****	.61**	.27	**.71****	**.79****	.68**	.61**
AD	**.77****	**.74****	**.76****	**.71****	.64**	.61**	**.70****	**.81****
BC[b]	**.82****	**.80****	**.80****	.69**	–	–	–	–
Grade 2								
AB[a]	–	–	–	–	**.72****	.63**	.55**	.58**
AC	**.81****	**.82****	**.82****	**.78****	.52**	.49**	.43*	.51**
AD	**.78****	**.77****	**.80****	**.76****	**.81****	**.82****	**.80****	**.73****
BC[b]	**.72****	**.74****	**.71****	**.73****	–	–	–	–
Grade 3								
AB[a]	–	–	–	–	**.83****	**.85****	**.87****	**.80****
AC	**.71***	**.70***	**.73***	**.78***	**.86****	**.87****	**.85****	**.75****
AD	**.75***	**.74***	**.72***	**.74***	**.79****	**.81****	**.84****	**.81****
BC[b]	**.70***	**.70***	.69*	**.72***	–	–	–	–
Winter								
Grade 1								
AB[a]	–	–	–	–	**.88****	**.89****	**.91****	**.86****
AC	**.85****	**.82****	**.75****	.63**	**.86****	**.87****	**.89****	**.84****
AD	**.84****	**.82****	**.79****	.60**	**.71****	.64**	.63**	.66**
BC[b]	**.86****	**.82****	**.75****	.67**	–	–	–	–
Grade 2								
AB[a]	–	–	–	–	**.92****	**.93****	**.87****	**.81****
AC	**.76****	**.87****	**.90****	**.86****	**.83****	**.87****	**.85****	**.82****
AD	**.88****	**.89****	**.90****	**.84****	**.76****	**.71****	.69**	.53**
BC[b]	**.78****	**.76****	**.78****	**.74****	–	–	–	–
Grade 3								
AB[a]	-	-	-	-	**.86****	**.86****	**.86****	**.84****
AC	**.90****	**.90****	**.87****	**.75****	**.80****	**.84****	**.85****	**.80****
AD	**.83****	**.88****	**.87****	**.83****	**.78****	**.79****	**.79****	**.79****
BC[b]	**.80****	**.81****	**.81****	**.74****				
Spring								
Grade 1								
AB[a]	–	–	–	–	**.90****	**.88****	**.88****	**.78****
AC	**.85****	**.87****	**.88****	**.82****	**.88****	**.87****	**.81****	**.78****
AD	**.84****	**.86****	**.93****	**.92****	**.76****	**.78****	**.82****	**.84****
BC[b]	**.75****	**.72****	**.60****	.38	–	–	–	–
Grade 2								
AB[a]	–	–	–	–	**.71****	.68**	.67**	.67**
AC	**.82****	**.80****	**.85****	**.83****	**.78****	**.80****	**.83****	**.82****
AD	**.82****	**.80****	**.79****	**.74****	**.73****	**.71****	**.78****	**.78****
BC[b]	**.84****	**.85****	**.90****	**.92****	–	–	–	–
Grade 3								
AB[a]	–	–	–	–	**.85****	**.83****	**.77****	.67**
AC	**.72****	**.80****	**.80****	**.77****	**.94****	**.93****	**.93****	**.84****
AD	**.90****	**.89****	**.86****	**.80****	**.71****	**.75****	**.81****	**.83****
BC[b]	**.86****	**.84****	**.83****	**.80****	–	–	–	–

Note. Bold statistics indicate $r \geq .70$. WW = Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; C-IWS = Correct Minus Incorrect Word Sequences; [a]AB was only given to District 2; [b]BC was only given to District 1; $^{*}p \leq .05$, $^{**}p \leq .01$.

Table 6. Story prompt alternate form reliability coefficients.

	District 1 (*n* = 338)				District 2 (*n* = 274)			
	WW	WSC	CWS	C-IWS	WW	WSC	CWS	C-IWS
Fall								
Grade 1								
AB	**.75****	**.77****	**.71****	.46*	**.74****	**.76****	**.68****	.48**
AC	**.76****	.64**	.46*	.32	**.80****	.68**	.54**	.61**
AD	.61**	.56**	.66**	.57**	**.79****	**.79****	**.77****	.31
Grade 2								
AB	**.76****	**.76****	.69**	.62**	.50**	.59**	**.70****	.67**
AC	**.75****	**.78****	**.84****	**.81****	.65**	.52**	.47**	.52**
AD	**.80****	**.78****	**.78****	.62**	**.79****	**.78****	**.80****	**.77****
Grade 3								
AB	**.76****	**.79****	**.83****	**.83****	.62**	.65**	**.74****	**.80****
AC	**.80****	**.84****	**.82****	**.79****	**.75****	**.76****	**.84****	**.82****
AD	**.70****	.67**	.68**	.62**	**.83****	**.83****	**.81****	**.72****
Winter								
Grade 1								
AB	**.72****	.68**	.67**	.68**	**.81****	**.83****	**.80****	**.75****
AC	**.77****	**.74****	.66**	.46**	**.80****	**.81****	**.79****	**.70****
AD	.52**	.59**	.55**	.54**	.64**	.62**	.63**	**.71****
Grade 2								
AB	**.85****	**.82****	**.81****	.65**	**.90****	**.89****	**.86****	**.76****
AC	**.88****	**.80****	.69**	.53**	**.77****	**.79****	**.79****	**.68****
AD	**.81****	**.84****	**.86****	**.81****	.59**	.51**	.48*	.54**
Grade 3								
AB	**.88****	**.89****	**.82****	**.77****	**.82****	**.85****	**.86****	**.82****
AC	**.84****	**.80****	**.72****	.65**	**.72****	**.76****	**.82****	**.78****
AD	**.86****	**.87****	**.86****	**.78****	**.79****	**.71****	**.70****	.61**
Spring								
Grade 1								
AB	**.87****	**.76****	.69**	.46*	**.84****	**.84****	**.82****	**.71****
AC	**.70****	.62**	**.75****	**.78****	**.91****	**.94****	**.88****	**.72****
AD	**.87****	**.90****	**.75****	.51**	**.76****	**.80****	.43*	.46*
Grade 2								
AB	**.89****	**.88****	**.90****	**.86****	**.81****	**.75****	.61**	.54**
AC	**.86****	**.85****	**.80****	.66**	**.90****	**.89****	**.89****	**.87****
AD	**.79****	**.74****	.65**	.57**	**.71****	**.68****	.65**	**.71****
Grade 3								
AB	**.90****	**.90****	**.86****	**.77****	**.87****	**.85****	**.83****	**.81****
AC	**.75****	**.76****	**.77****	**.72****	**.88****	**.91****	**.92****	**.92****
AD	**.75****	**.72****	**.58****	.46*	**.78****	**.80****	**.78****	**.70****

Note. Bold statistics indicate $r \geq .70$. WW = Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; C-IWS = Correct Minus Incorrect Word Sequences;
$^{*}p \leq .05$, $^{**}p \leq .01$

Table 7. Picture-word validity coefficients, mean of two alternate forms.

	District 1 (WIAT-III; *n* = 150)		District 2 (WJ-IV; *n* = 92)				
	Spelling	*Sentence Composition*	*Spelling*	*Sentence Writing*	*Writing Samples*	*Broad Written Language*	*Written Expression*
Fall							
Grade 1							
WW	.14	.12	.28	.34	.22	.31	.28
WSC	.22	.19	.37	.35	.27	.37	.31
CWS	.24	.27	.44*	.38	.31	.43*	.36
C-IWS	.29*	.36*	**.58****	.39*	.39	**.52****	.41*
Grade 2							
WW	.29*	.26	.44*	.24	.27	.41*	.29
WSC	.39**	.41**	**.56****	.34	.34	**.52****	.38*
CWS	.42**	**.50****	**.68****	.46**	.45**	**.66****	**.52****
C-IWS	.42**	**.55****	**.73****	**.54****	**.50****	**.73****	**.59****
Grade 3							
WW	.43**	.13	**.50****	**.50****	.38	**.53****	.48*
WSC	.49**	.18	**.51****	**.53****	.38	**.54****	**.50****
CWS	**.56****	.20	**.55****	**.61****	.42*	**.60****	**.57****
C-IWS	**.62****	.26	**.58****	**.65****	.44*	**.63****	**.60****
Spring							
Grade 1							
WW	.11	-0.01	.30	.16	.26	.26	.24
WSC	.22	.10	.38	.16	.26	.31	.25
CWS	.33*	.16	.39	.14	.23	.30	.21
C-IWS	.46**	.30*	.48*	.17	.22	.35	.23
Grade 2							
WW	.41**	.45**	.21	.47**	-0.08	.26	.25
WSC	**.52****	**.54****	.33	**.53****	-0.04	.35	.31
CWS	**.62****	**.60****	.38*	**.54****	.04	.41*	.36
C-IWS	**.66****	**.60****	**.50****	**.56****	.17	**.51****	.44*
Grade 3							
WW	**.52****	.19	.49**	.40*	**.50****	**.53****	**.50****
WSC	**.55****	.22	**.54****	.44*	**.52****	**.58****	**.54****
CWS	**.63****	.29*	**.63****	**.55****	**.57****	**.67****	**.62****
C-IWS	**.66****	.36*	**.73****	**.63****	**.62****	**.76****	**.68****

Note. Bold statistics indicate $r \geq .50$. WW = Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; C-IWS = Correct Minus Incorrect Word Sequences.

$^{*}p \leq .05$, $^{**}p \leq .01$

Table 8. Story prompt validity coefficients, mean of two alternate forms.

	District 1 (WIAT-III; $n = 150$)		District 2 (WJ-IV; $n = 92$)				
	Spelling	*Sentence Composition*	*Spelling*	*Sentence Writing Fluency*	*Writing Samples*	*Broad Written Language*	*Written Expression*
Fall							
Grade 1							
WW	−.03	-0.05	.30	.28	.24	.30	.28
WSC	.13	.07	.46*	.44*	.37	.47*	.43*
CWS	.33*	.24	**.60****	**.55****	.48*	**.62****	**.56****
C-IWS	**.56****	.45**	**.58****	**.54****	.47*	**.62****	**.56****
Grade 2							
WW	.37**	.25	.40*	.27	.29	.40*	.32
WSC	**.53****	.38*	**.56****	.41*	.37*	.56**	.45**
CWS	**.64****	.47**	**.69****	.49**	.41*	**.67****	**.53****
C-IWS	**.65****	**.51****	**.69****	**.51****	.39*	**.68****	**.53****
Grade 3							
WW	.35*	.16	**.56****	**.52****	.49**	**.59****	**.54****
WSC	.44**	.23	**.61****	**.56****	**.52****	**.64****	**.58****
CWS	**.56****	.33*	**.67****	**.62****	**.59****	**.71****	**.65****
C-IWS	**.50****	.39**	**.70****	**.63****	**.59****	**.74****	**.66****
Spring							
Grade 1							
WW	.18	.00	.22	.39	.23	.28	.31
WSC	.33*	.13	.35	.43*	.24	.36	.32
CWS	.49**	.25	.31	.30	.21	.30	.24
C-IWS	**.59****	.41**	.44*	.20	.19	.33	.18
Grade 2							
WW	**.50****	.39**	.27	.36	.01	.28	.25
WSC	**.62****	**.50****	.35	.45*	.06	.37*	.33
CWS	**.70****	**.59****	.47**	.47**	.13	.47**	.38*
C-IWS	**.74****	**.57****	**.57****	**.52****	.22	**.57****	.46*
Grade 3							
WW	**.56****	.32*	**.65****	**.52****	**.50****	**.67****	**.57****
WSC	**.54****	.30*	**.72****	**.53****	**.52****	**.71****	**.58****
CWS	**.59****	.38**	**.77****	**.52****	**.52****	**.75****	**.58****
C-IWS	**.60****	.41**	**.81****	**.50****	**.50****	**.76****	**.55****

Note. Bold statistics indicate $r \geq .50$. WW = Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; C-IWS = Correct Minus Incorrect Word Sequences.

$^{*}p \leq .05$, $^{**}p \leq .01$.

Table 9. MANOVA results.

Multivariate results

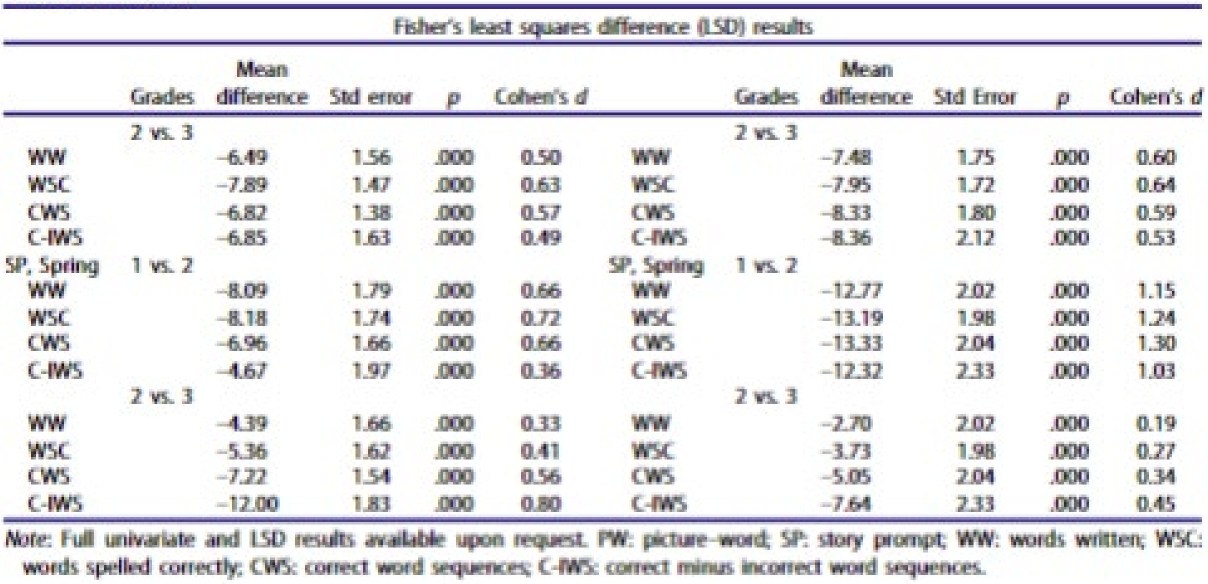
	District 1 ($n = 338$)							District 2 ($n = 274$)					
	df	Error *df*	Wilks' Λ	*F*	*p* Value	Partial η^2		*df*	Error *df*	Wilks' Λ	*F*	*p* Value	Partial η^2
PW							PW						
Fall	8	470	0.63	15.44	<.001	0.21	Fall	8	474	0.74	9.80	<.001	0.14
Winter	8	374	0.64	11.86	<.001	0.20	Winter	8	508	0.70	12.36	<.001	0.16
Spring	8	432	0.65	12.89	<.001	0.19	Spring	8	458	0.75	8.74	<.001	0.13
SP							SP						
Fall	8	672	0.70	16.09	<.001	0.16	Fall	8	470	0.60	17.24	<.001	0.23
Winter	8	628	0.69	15.73	<.001	0.17	Winter	8	504	0.67	14.02	<.001	0.18
Spring	8	638	0.73	13.72	<.001	0.15	Spring	8	452	0.72	10.00	<.001	0.15

Fisher's least squares difference (LSD) results

	Grades	Mean difference	Std error	*p*	Cohen's *d*		Grades	Mean difference	Std Error	*p*	Cohen's *d*
PW, Fall	1 vs. 2					PW, Fall	1 vs. 2				
WW		−14.57	1.58	.000	1.33	WW		−7.77	1.35	.000	0.96
WSC		−14.05	1.53	.000	1.35	WSC		−8.19	1.36	.000	1.02
CWS		−14.04	1.68	.000	1.27	CWS		−9.55	1.57	.000	1.08
C-IWS		−11.11	2.04	.000	0.83	C-IWS		−10.40	1.85	.000	0.97
	2 vs. 3						2 vs. 3				
WW		−4.10	1.48	.01	0.33	WW		−4.92	1.40	.000	0.53
WSC		−5.60	1.43	.000	0.46	WSC		−5.91	1.41	.000	0.62
CWS		−7.75	1.57	.000	0.57	CWS		−7.63	1.62	.000	0.67
C-IWS		−10.86	1.91	.000	0.65	C-IWS		−9.47	1.92	.000	0.70
PW, Winter	1 vs. 2					PW, Winter	1 vs. 2				
WW		−15.89	1.72	.000	1.36	WW		−9.59	1.70	.000	0.87
WSC		−15.69	1.63	.000	1.42	WSC		−10.37	1.72	.000	0.94
CWS		−17.45	1.83	.000	1.42	CWS		−13.61	2.05	.000	1.03
C-IWS		−15.66	2.14	.000	1.10	C-IWS		−16.06	2.38	.000	1.04
	2 vs. 3						2 vs. 3				
WW		−4.81	1.60	.000	0.37	WW		−6.79	1.75	.000	0.57
WSC		−7.61	1.52	.000	0.61	WSC		−7.44	1.77	.000	0.43
CWS		−8.61	1.70	.000	0.61	CWS		−8.41	2.11	.000	0.59
C-IWS		−10.93	1.99	.000	0.64	C-IWS		−9.02	2.45	.000	0.54
PW, Spring	1 vs. 2					PW, Spring	1 vs. 2				
WW		−10.30	1.57	.000	0.95	WW		−12.47	1.90	.000	1.12
WSC		−10.91	1.59	.000	1.00	WSC		−13.23	1.88	.000	1.21
CWS		−12.08	1.98	.000	0.87	CWS		−16.03	2.26	.000	1.21
C-IWS		−11.36	2.51	.000	0.64	C-IWS		−17.71	2.58	.000	1.16
	2 vs. 3						2 vs. 3				
WW		−4.04	1.46	.01	0.36	WW		0.39	1.91	0.84	0.03
WSC		−5.27	1.48	.000	0.47	WSC		0.02	1.89	0.99	0.0
CWS		−7.08	1.84	.000	0.51	CWS		0.23	2.26	0.92	0.01
C-IWS		−10.89	2.33	.000	0.60	C-IWS		−0.49	2.59	0.85	0.03
SP, Fall	1 vs. 2					SP, Fall	1 vs. 2				
WW		−9.00	1.58	.000	0.79	WW		−8.07	1.45	.000	1.01
WSC		−8.58	1.44	.000	0.86	WSC		−8.97	1.39	.000	1.27
CWS		−7.50	1.33	.000	0.87	CWS		−9.10	1.42	.000	1.36
C-IWS		−5.03	1.68	.000	0.51	C-IWS		−9.82	1.73	.000	1.11
	2 vs. 3						2 vs. 3				
WW		−8.15	1.48	.000	0.64	WW		−8.01	1.49	.000	0.77
WSC		−8.94	1.35	.000	0.76	WSC		−8.78	1.43	.000	0.85
CWS		−7.28	1.24	.000	0.66	CWS		−9.23	1.46	.000	0.84
C-IWS		−6.21	1.57	.000	0.45	C-IWS		−9.62	1.78	.000	0.73
SP, Winter	1 vs. 2					SP, Winter	1 vs. 2				
WW		−8.98	1.68	.000	0.84	WW		−9.70	1.69	.000	0.94
WSC		−8.65	1.59	.000	0.88	WSC		−10.69	1.67	.000	1.08
CWS		−8.16	1.49	.000	0.89	CWS		−11.98	1.75	.000	1.24
C-IWS		−6.43	1.75	.000	0.60	C-IWS		−13.64	2.05	.000	1.20

(continued)

Table 9. Continued.



Fisher's least squares difference (LSD) results											
	Grades	Mean difference	Std error	*p*	Cohen's *d*		Grades	Mean difference	Std Error	*p*	Cohen's *d*
	2 vs. 3						2 vs. 3				
WW		−6.49	1.56	.000	0.50	WW		−7.48	1.75	.000	0.60
WSC		−7.89	1.47	.000	0.63	WSC		−7.95	1.72	.000	0.64
CWS		−6.82	1.38	.000	0.57	CWS		−8.33	1.80	.000	0.59
C-IWS		−6.85	1.63	.000	0.49	C-IWS		−8.36	2.12	.000	0.53
SP, Spring	1 vs. 2					SP, Spring	1 vs. 2				
WW		−8.09	1.79	.000	0.66	WW		−12.77	2.02	.000	1.15
WSC		−8.18	1.74	.000	0.72	WSC		−13.19	1.98	.000	1.24
CWS		−6.96	1.66	.000	0.66	CWS		−13.33	2.04	.000	1.30
C-IWS		−4.67	1.97	.000	0.36	C-IWS		−12.32	2.33	.000	1.03
	2 vs. 3						2 vs. 3				
WW		−4.39	1.66	.000	0.33	WW		−2.70	2.02	.000	0.19
WSC		−5.36	1.62	.000	0.41	WSC		−3.73	1.98	.000	0.27
CWS		−7.22	1.54	.000	0.56	CWS		−5.05	2.04	.000	0.34
C-IWS		−12.00	1.83	.000	0.80	C-IWS		−7.64	2.33	.000	0.45

Note: Full univariate and LSD results available upon request. PW: picture–word; SP: story prompt; WW: words written; WSC: words spelled correctly; CWS: correct word sequences; C-IWS: correct minus incorrect word sequences.

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. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 345–363). New York, NY: Guilford Press.

Berninger, V. W., Rutberg, J. E., Abbott, R. D., Garcia, N., Anderson-Youngstrom, M., Brooks, A., & Fulton, C. (2006). Tier 1 and tier 2 early intervention for handwriting and composing. *Journal of School Psychology*, *44*(1), 3–30. doi:10.1016/j.jsp.2005.12.003

Berninger, V. W., Vaughan, K., Abbott, R. D., Begay, K., Coleman, K. B., Curtin, G., … Graham, S. (2002). Teaching spelling and composition alone and together: Implications for the simple view of writing. *Journal of Educational Psychology*, *94*(2), 291–304. doi:10.1037/0022-0663.94.2.291

Breaux, K. C. (2009). *WIAT-III technical manual*. San Antonio, TX: Pearson.

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

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

Cutler, L., & Graham, S. (2008). Primary grade writing instruction: A national survey. *Journal of Educational Psychology*, *100*(4), 907–919. doi:10.1037/a0012656

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*(3), 219–232.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, *37*(3), 184–192. doi:10.1177/00224669030370030801

Deno, S. L., Marston, D., & Mirkin, P. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children*, *48*, 368–371. doi:10.1177/001440298204800417

Deno, S. L., Mirkin, P. K., & Marston, D. (1980). *Relationships among simple measures of written expression and performance on standardized achievement tests* (No. IRLDRR-22). Minneapolis: MN: University of Minnesota, Institute for Research on Learning Disabilities.

Deno, S. L., Mirkin, P. K., Marston, D., Lowry, L., Sindelar, P., & Jenkins, J. (1982). *The use of standard tasks to measure achievement in reading, spelling, and written expression: A normative and developmental study* (No. IRLDRR-87). Minneapolis: MN: University of Minnesota, Institute for Research on Learning Disabilities.

Espin, C. A., Scierka, B. J., Skare, S., & Halverson, N. (1999). Criterion-related validity of curriculum-based measures in writing for secondary school students. *Reading and Writing Quarterly*, *15*(1), 5–27. doi:10.1080/ 105735699278279

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review*, *33*, 188–193.

Fuchs, L. S., & Deno, S. L. (1991). Effects of curriculum within curriculum-based measurement. *Exceptional Children*, *58*, 232–243. doi:10.1177/001440299105800306

Fuchs, L., Fuchs, D., & Hamlett, C. (1989). Computers and curriculum-based measurement: Effects of teacher feed- back systems. *School Psychology Review*, *18*, 112–125. doi:10.1177/002221948902200110

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., Harris, K. R., & McKeown, D. (2013). The writing of students with learning disabilities, meta-analysis of self-regulated strategy development writing intervention studies, and future directions: Redux. In L. Swanson,

K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 405–438). New York: Guilford Press.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*(3), 445–476. doi:10.1037/0022-0663.99.3.445

Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology*, *78*(4), 243–255. doi:10.1037/0022-0663.78.4.243

Jung, P., McMaster, K. L., & delMas, R. (2017). Effects of early writing intervention delivered within a data-based instruction framework. *Exceptional Children*, *83*, 281–297. doi:10.1177/0014402916667586

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*(3–4), 23–35. doi:10.1177/073724770302800304 Marston, D., & Deno, S. (1981). *The reliability of simple, direct measures of written expression* (No. IRLDRR-50).

Minneapolis: MN: University of Minnesota, Institute for Research on Learning Disabilities.

McMaster, K. L., & Campbell, H. (2008). New and existing curriculum-based writing measures: Technical features within and across grades. *School Psychology Review*, *37*, 550–566.

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *The Journal of Special Education*, *41*(2), 68–84.

McMaster, K. L., Du, X., & Petursdottir, A. L. (2009). Technical features of curriculum-based measures for begin- ning 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, L. (2018). Early intervention: A best-evidence synthesis. *Journal of Learning Disabilities*, *51*(4), 363–380. doi:10.1177/0022219417708169

McMaster, K. L., Parker, D., & Jung, P. (2012). Using curriculum-based measurement for beginning writers within a response to intervention framework. *Reading Psychology*, *33*(1–2), 190–216. doi:10.1080/02702711.2012.631867

McMaster, K. L., Ritchey, K. D., & Lembke, E. (2011). Curriculum-based measurement for beginning writers: Recent developments and future directions. In *Assessment and Intervention* (pp. 111–148). Bingley, UK: Emerald Group Publishing Limited.

National Center for Education Statistics. (2011). *The nation's report card: Writing 2011* (NCES 2012–470). Washington, D.C.: Institute of Education Sciences, U.S. Department of Education.

Parker, D. C., McMaster, K. L., Medhanie, A., & Silberglitt, B. (2011). Modeling early writing growth with curriculum-based measures. *School Psychology Quarterly*, *26*(4), 290–304. doi:10.1037/a0026833

Parker, R., Tindal, G., & Hasbrouck, J. (1991). Progress monitoring with objective measures of writing performance for students with mild disabilities. *Exceptional Children*, *58*, 61–73. doi:10.1177/001440299105800106

Pearson (2009). *Weschler individual achievement test (3rd edition)*. San Antonio, TX: The Psychological Corporation.

Ritchey, K. D., & Coker, D. L. (2013). An investigation of the validity and utility of two curriculum-based measurement writing tasks. *Reading and Writing Quarterly*, *29*(1), 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 and Practice*, *29*, 54–65. doi:10.1111/ldrp.12030

Schrank, F. A., Mather, N., & McGrew, K. S. (2014). *Woodcock-Johnson IV tests of achievement*. Rolling Meadows, IL: Riverside.

Shinn, M. R. (Ed.). (1989). *Curriculum-based measurement: Assessing special children*. New York, NY: Guilford Press.

Swanson, H. L., & Zheng, X. (2013). Memory difficulties in children and adults with learning disabilities. In L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 214–238). New York: The Guilford Press.

Taylor, R. L. (2003). *Assessment of exceptional students: Educational and psychological procedures* (6th ed.). Boston: Allyn & Bacon.

Tindal, G., & Parker, R. (1991). Identifying measures for evaluating written expression. *Learning Disabilities Research & Practice*, *58*(1), 61–73.

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