5-28-2018

Exploration of New Complexity Metrics for Curriculum-Based Measures of Writing

Kyle Wagner

Alex Smith

Abigail A. Allen

Kristen L. McMaster

Apryl L. Poch

See next page for additional authors

Authors

Kyle Wagner, Alex Smith, Abigail A. Allen, Kristen L. McMaster, Apryl L. Poch, and Erica S. Lembke

Exploration of New Complexity Metrics for Curriculum-Based Measures of Writing

Kyle Wagner, SSP[1], Alex Smith, PhD[2], Abigail Allen, PhD[3],
Kristen McMaster, PhD[1], Apryl Poch, PhD[2], and Erica Lembke, PhD[2]
[1]University of Minnesota, Minneapolis, USA [2]University of Missouri, Columbia, USA
[3]Clemson University, SC, USA

Abstract
Researchers and practitioners have questioned whether scoring procedures used with curriculum-based measures of writing (CBM-W) capture growth in complexity of writing. We analyzed data from six independent samples to examine two potential scoring metrics for picture word CBM-W (PW), a sentence-level CBM task. Correct word sequences per response (CWSR) and words written per response (WWR) were compared with the current standard metric of correct word sequences (CWS). Linear regression analyses indicated that CWSR predicted scores on standardized norm- referenced criterion measures in more samples than did WWR or CWS. Future studies should explore the capacity of CWSR and WWR to show growth over time, stability, diagnostic accuracy, and utility for instructional decision making.

Keywords
curriculum-based measurement, written expression

Writing is a complex activity that requires the coordination and integration of both receptive and expressive language. Writing is further influenced by the specific task environment and cognitive resources such as attention, long-term memory, short-term memory, and working memory (Hayes & Berninger, 2014). Due to its complexity, writing has proven difficult to measure. To guide assessment and instruction specifically in early writing, researchers have proposed a Simple View of Writing (e.g., Berninger &

Amtmann, 2003), in which transcription skills (e.g., hand- writing, spelling) work in conjunction with self-regulation (e.g., monitoring attention, working toward a goal) to pro- mote text generation (e.g., generating words, sentences, and passages to write). All three processes (transcription, self- regulation, text generation) are constrained by limited attention and memory resources. Using the Simple View as a framework, many researchers have focused on transcription skills, especially among young writers, and studies have shown a significant relationship between transcription skills and overall writing quality (e.g., Abbott, Berninger, & Fayol, 2010; Berninger et al., 1997; Graham, Berninger, Abbott, Abbott, & Whitaker, 1997).

Transcription skills play a significant and critical role in writing development, especially among writers in the early elementary grades, and are thus justified in being incorporated into assessment. One specific approach to assessing writing is curriculum-based measures of writing (CBM-W), which includes scoring procedures that reflect accurate and fluent transcription skills and have evidence of psychometric adequacy related to broader measures of writing proficiency (see McMaster et al., 2011; Ritchey et al., 2016, for a review). However, transcription skills as currently measured using CBM-W do not explain all of the variance in writing quality and student performance on standardized writing assessments.

CBM-W

Curriculum-based measure (CBM) entails a set of global measures of academic performance that are quick and easy to administer, score, and interpret for teachers as well as affordable for schools (Deno, 1985). CBM tasks are standardized across items, scoring procedures, and administration procedures. Standardization allows teachers and schools to interpret results both within and across students, which provides feedback regarding students' responsiveness to instruction. To be most effective, CBM should have evidence that it is technically adequate and sensitive to change across time, allowing it to be used both as a universal screener and as a tool to monitor student progress (Deno et al., 2009). While CBM-W tasks designed to assess pas- sage-level composition (e.g., story, picture, and photo prompts) demonstrate evidence of technical adequacy and face validity for students in third grade and above (McMaster, Du, &

Pétursdóttir, 2009; McMaster et al., 2011; Ritchey & Coker, 2014), research indicates that sentence-level tasks (e.g., picture word, sentence copying, and sentence writing tasks) are more appropriate for students in the early elementary grades (Coker & Ritchey, 2014; Lembke, Deno, & Hall, 2003; McMaster et al., 2009; McMaster et al., 2011; Ritchey & Coker, 2014).

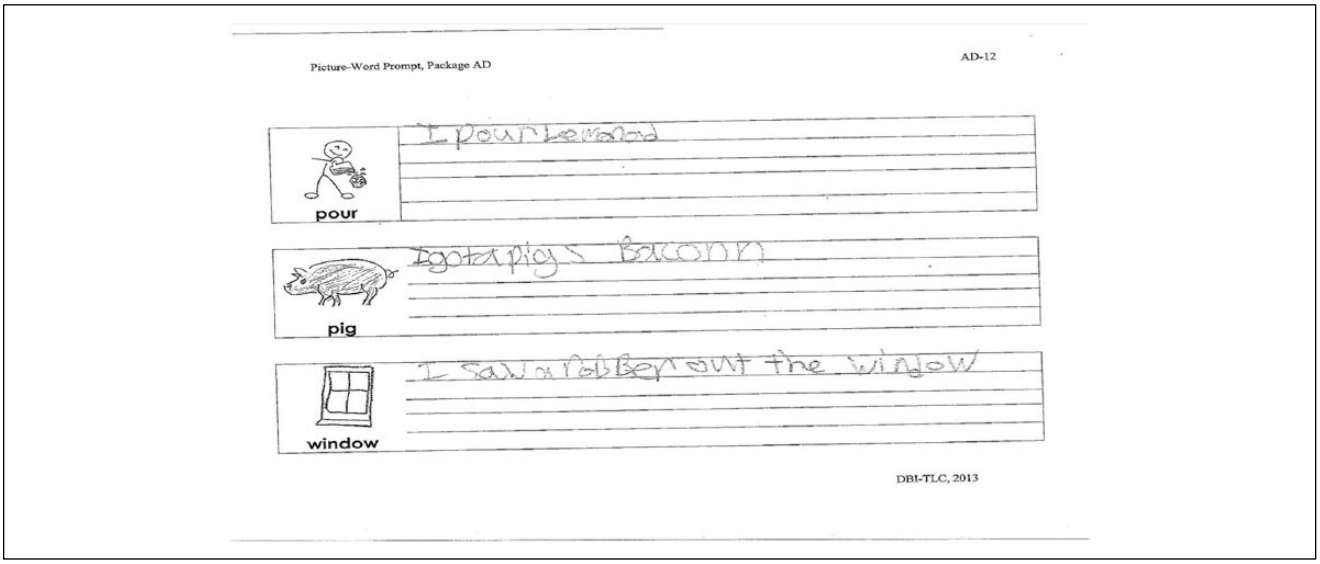

Figure 1. Example of picture word prompt.

Of these sentence-level measures, picture word CBM (PW) has an emerging research base supporting its use with writers in the early elementary grades (McMaster et al., 2009; McMaster et al., 2011; McMaster, Brandes, Herriges, & Jung, 2014). PW is designed to assess both transcription and text generation skills at the sentence level. PW includes 12 picture–word combinations (see Figure 1) in which students are given 3 min to write their best sentence for as many picture–word combinations as they can. PW has the potential to serve as a CBM-W that captures more than transcription skills alone, yet is more accessible to young and/or struggling writers than passage- level CBM-W. However, most PW scoring procedures still rely to a large extent upon transcription skills.

Three common classes of scoring procedures used with CBM-W, including PW, are (a) simple production, (b) accurate production, and (c) production independent. Simple production measures rely on the quantity a student writes and include the number of words written (WW). Accurate production measures rely upon both quantity and accuracy, such as number of words spelled correctly (WSC) and number of correct word sequences (CWS), defined as two adjacent words spelled correctly that are also used correctly in the context of the sentence, inclusive of capitalization and punctuation (Videen, Deno, & Marston, 1982). Production-independent measures focus on accuracy rather than quantity, and are generally either averaged over the length of a student's writing (e.g., average WSC) or calculated as percentage measures, such as percentage WSC or percentage CWS (Jewell & Malecki, 2005; Parker, Tindal, & Hasbrouck, 1991).

Each type of scoring procedure has strengths and weak- nesses. Many production-dependent measures (e.g., WW) are quick and easy to score, but accurate production measures (e.g., CWS) are more difficult to score but have more evidence of technical adequacy than simple quantity metrics (McMaster et al., 2009; McMaster et al., 2011). Those that only measure quantity are also less instructionally useful to educators. Production-independent measures capture aspects of quality and do not penalize slow writers, but are prone to ceiling effects (e.g., only spelling one word but spelling it correctly results in 100% of WSC; Jewell & Malecki, 2005; Parker et al., 1991). Furthermore, these metrics do not explicitly assess growth or performance in

lexical, syntactic, or discursive complexity, which are important features in writing development that may affect the ongoing performance in accuracy and production of transcription skills.

Capturing Complexity

Some researchers have criticized current CBM-W scoring procedures for not capturing the full range of skills needed for quality writing (Coker & Ritchey, 2010; Tindal & Parker, 1991). Current measures predominantly quantify mechanical aspects of writing (transcription skills) with a focus on fluency and accuracy, but do not adequately capture students' abilities to take risk and experiment with their writing or incorporate newly learned skills. Although transcription skills are critical to writing proficiency in the early elementary grades (Graham et al., 1997; Jones & Christensen, 1999), spelling and handwriting alone do not explain all of the variance in writing quality, and these skills become less central to writing proficiency in third grade and beyond (Berninger et al., 1997; Jewell & Malecki, 2005; Parker et al., 1991). Thus, an overreliance upon transcription-level skills may negatively affect the long-term predictive validity of CBM-W. Furthermore, production and accuracy share a nonlinear and dynamic relation with complexity, defined as the use of longer, more obscure, or newly learned vocabulary and/or sentence structures, that McCutchen (2006) calls a "paradox in the development of writing skill" (p. 126). Within this paradox, more skilled writers appear less fluent than less skilled writers as the more skilled writers begin to attend more closely to the quality of their writing, such as generating new and com- plex ideas, choosing sophisticated words to express those ideas, and attending to text structure and genre.

Relatedly, research in the field of second language acquisition has conceptualized the multicomponential development of writing in a model called CAF, wherein C stands for complexity, A for accuracy, and F for fluency (Housen & Kuiken, 2009). Second language acquisition research has shown that, as writing skills develop, speed and accuracy of composition may temporarily suffer as the student explores and attempts to compose more complex writing (Bulté & Housen, 2014; Housen & Kuiken, 2009). The CAF model may also be an effective way to assess the writing development of all

writers, even those writing in their native language. Furthermore, quality writing, the ultimate goal of writing instruction, is an amalgamation of each aspect of CAF, and each aspect should be accounted for in assessment because the three components have dynamic relations (Bulté & Housen, 2014; Housen & Kuiken, 2009). In other words, growth in one aspect (complexity) may not immediately translate to growth in one or both of the other aspects (accuracy or fluency). For example, as students begin to use newly learned vocabulary and/or sentence structures (i.e., increase in complexity), they are likely to slow down their production and make more errors initially (i.e., decrease in production and accuracy). Using typical scoring strategies with PW, such as CWS, may therefore show a lack of growth or even negative growth as students' grammar and spelling temporarily suffer while they attempt to write more complex sentences using more complex vocabulary. Applying the CAF model, CWS accounts for accuracy and fluency but not complexity, at least not within the context of the PW prompt. When calculating *total* CWS across a series of unrelated sentences in the PW task, a student who is slow at producing text would essentially be penalized with a lower CWS score because of the lack of total word sequences he or she produced, regardless of the syntactic complexity of those individual sentences. For example, a slow writer could feasibly produce two complex sentences and end up with a lower CWS score than a student who writes 12 simple sentences. The second student could achieve a higher total CWS score because he or she produced more but not necessarily *better* sentences and word sequences. Total CWS is therefore limited in its capacity to measure individual sentence complexity, particularly with the PW task. Compare this with writing narrative text, where even a slow writer could feasibly produce a short story containing more sophisticated elements of syntax and story grammar (e.g., dialogue, quotations, multiple com- plex sentences). Therefore, *total* CWS may better capture a developing writer's use of advanced syntax and mechanics in the context of an extended story writing task while being limited in capturing individual sentence-level complexity with PW prompts. A metric that incorporates production averaged across sentences may be a better estimate of sentence-level complexity with the PW task by attempting to capture the degree to which each sentence is syntactically sophisticated rather than the volume of word sequences produced. CAF then supports the argument

that a measure including aspects of production, accuracy, and complexity should perform better than previously explored scoring procedures (e.g., CWS).

As illustrated above, scoring procedures measuring only production and/or accuracy may not account for improved writing quality when students begin to produce more com- plex compositions. Therefore, growth in complexity may result in highly variable data as students learn and develop as writers, which can affect the face validity and instructional utility of the measure for teachers. Assessment that does not take into account all aspects of CAF may even appear to work against a teacher's instructional objectives at times by effectively penalizing students with lower CWS scores as they attempt to produce more complex writing and apply new skills. The incorporation of complexity may improve both instructional utility and face validity; how- ever, few researchers have examined how to quantify complexity with CBM-W.The incorporation of complexity may improve both instructional utility and face validity; however, few researchers have examined how to quantify complexity with CBM-W.

Prior Complexity Studies With CBM-W

Among those studies that have addressed complexity, few measures have shown promise: One measure of complexity reported in the literature, the terminal unit (T unit), is often used in linguistics (Berman, 2014), but has not shown promise as a measure with CBM-W, and could prove difficult for many elementary school teachers to score (Campbell, Espin, & McMaster, 2013; McMaster & Espin, 2007). A T unit is an independent clause and any related dependent clauses. In previous studies, the typical unit of measurement was mean total number of T units students produced (Deno et al., 1982), which means that there would be essentially no difference between a series of simple sentences containing one independent clause (e.g., "I like hats") and a series of complex sentences with dependent clauses (e.g., "When it's cold, I wear a hat"). This could explain why the T unit has not shown promise thus far.

Other measures of complexity, such as holistic and trait- based rating scales and rubrics, generally have correlated weakly to moderately across grade levels (r = .35–.76, Coker & Ritchey, 2010; r = .36–.37, Gansle, Noell, VanDerHeyden, Naquin, & Slider, 2002; r = .27, Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006; r = .06–

.67, Lembke et al., 2003; r = .50–.60, McMaster et al., 2009; r = −.02 to .63, Tindal & Parker, 1991). Some studies suggest a stronger relation between accurate production measures (e.g., WSC) and qualitative measures in early elementary grades (e.g., Coker & Ritchey, 2010; Lembke et al., 2003), but these correlations appear to decrease by upper elementary grades (Gansle et al., 2006).

A study by Allen, Poch, and Lembke (2017) explored the use of qualitative rubrics with CBM-W in two studies. Study 1 used PW. Allen et al. (2017) administered PW to first- grade students (n = 40) and used a sentence writing rubric adapted from Coker and Ritchey (2010) to score the writing measures. Allen et al. (2017) found that the total rubric scores had a weak to moderate concurrent correlation with the *Wechsler Individual Achievement Test-3* (WIAT-3; Psychological Corporation, 2009), Spelling subtest (r = .41, p < .01), and Sentence Composition subtest (r = .30, p > .05). These findings show promise for the use of a rubric with PW, but there were some concerns regarding interrater reliability, specifically for grammatical structure (82% interrater reliability), with a total interrater reliability of 89% and an internal reliability Cronbach's alpha of .64. The use of the rubric requires a different scoring modality by teachers in addition to traditional CBM-W scoring methods. This requirement for multiple scoring modalities (using traditional CBM-W scoring methods and then scoring the sample again using a rubric) may reduce CBM-W's feasibility for practicing teachers, especially for regular progress monitoring. Rubrics may be better suited for diagnostic, summative, and perhaps screening assessment, but may not be useful for frequent progress monitoring because of the amount of time they take to complete and questionable reliability.

Development of a New Scoring Procedure

In previous studies, PW has been scored using WW, WSC, and CWS (McMaster et al., 2009; McMaster et al., 2011). All three metrics have the benefit of being relatively easy for teachers to score rapidly and reliably, and can be used to efficiently track progress for ongoing monitoring. Researchers have provided evidence that CWS has accept- able technical adequacy and sensitivity to growth compared with other CBM

metrics (McMaster et al., 2009; McMaster et al., 2011). However, CWS relies primarily upon accuracy and fluency, and may not adequately measure growth in complexity, a key feature of quality as students begin to compose longer texts and move beyond the early elementary grades (Berman, 2014).

Because PW is structured as a sentence-level measure, syntactic complexity is the focus of this study. Although syn- tactic complexity represents only one aspect of complexity, this study represents a first step in exploring the potential of incorporating complexity within classic CBM-W scoring procedures. Literature in linguistics debates the exact definition of and grammatical structures underlying syntactic complexity, and most measures corresponding to the competing definitions are not likely to be reliably applied by practicing teachers, at least not until sophisticated scoring software is readily available or teachers are required to take several years of linguistic coursework (Bulté & Housen, 2014). However, Bulté and Housen (2014) state that syntactic complexity is basically just writing longer sentences. Although this definition may seem overly simplistic, the reliable identification of what is and what is not a sentence has proven difficult.

The clause (phrase with a subject and a verb) has been demonstrated to be a valid measure of syntactic complexity in both spoken and written discourse (Berman, 2014). Berman (2014) discourages the use of sentence-level measures, which contain more than one clause, because the term “sentence” is an ambiguous construct, whereas the clause is more easily defined and identified. Definitions of “sentence” may vary across scorers and lead to poor reliability. However, PW allows for the easy identification of an unambiguous syntactical unit similar to a sentence, which we call an item response (see also Allen et al., 2017). Each written response to a picture–word combination with at least two scored sequences (i.e., two CWS, two IWS, or one CWS and one IWS) counts as an item response, and is treated as a sentence regardless of whether it has correct grammar and/or punctuation. Because teachers do not have to identify the ambiguous sentence or splice a student’s writing into clauses, item responses on PW might serve as a more reliable, efficient, and feasible scoring mechanism than clauses or T units. The average number of words per item response (length) can be considered a measure of syntactic complexity similar to the average length of a sentence. This measure was termed words written per item response (WWR). WWR should capture aspects of

syntactic complexity related to sentence length as well as production due to the timed nature of PW, but does not capture accuracy. Prior CBM-W research underscoring the importance of accuracy along with a desire to incorporate all three components of CAF led

Table 1. Examples of Scoring Metrics.

Example Responses	CWS	CWSR	WWR
1. I like cats. 2. I like pants. 3. I like paper.	12	4	3
1. I walked out of my house and saw five cats.	11	11	10

Note. CWS = correct word sequences; CWSR = correct word sequences per response; WWR = words written per response.

us to also create a CWS per item response (CWSR) measure that attempted to incorporate syntactic complexity and accuracy by calculating the average number of CWS per item response. CWSR should capture elements of accuracy and syntactic complexity, as well as production due to the timed nature of PW, and thus be a better indicator of overall writing performance than measures that do not account for all aspects of CAF. See Table 1 for clarification on these scoring metrics.

Purpose

The purpose of this study was to report the interscorer reliability and concurrent criterion validity of two new scoring procedures for PW that were designed to measure a student's growth in syntactic complexity: WWR and CWSR. These two new metrics were compared with CWS, which has evidence of sensitivity to growth and technical adequacy for students in first through third grades (Lembke et al., 2003; McMaster, Brandes, Herriges, & Jung, 2014). This study represents initial attempts at creating and

validating a metric that will provide quantifiable data regarding a student's growth in syntactic complexity that is both technically adequate and feasible for elementary school teachers. Our research questions included the following:

Research Question 1: What is the reliability and validity of CWSR and WWR?
Research Question 2: How does CWSR predict writing performance in comparison with CWS and WWR?
Specifically, we hypothesized that both WWR and CWSR may be scored more reliably than other complexity measures (i.e., rubrics and T units), and that CWSR better predicts writing performance than either WWR or CWS because it incorporates some elements of each of the three aspects of CAF.

Method

Participants

We used data from two benchmarking studies conducted from Fall 2013 through Spring 2015 in two sites (Site 1 and Site 2) across Grades 1 to 3. Participants in each grade in each state were treated as a unique sample. The original sample was n = 274 students from two elementary schools in a large urban district in Site 1 and n = 338 students from two elementary schools from a small city school district in Site 2. For Site 1, researchers asked the classroom teachers to rank order students according to writing proficiency within each class. A stratified subset of participants, selected according to high, middle, and low writing performance, completed the Spelling, Writing Samples, and Sentence Writing Fluency subtests from the *Woodcock–Johnson Tests of Achievement-IV* (WJ IV-ACH; Houghton Mifflin Harcourt, 2014) in addition to CBM-W measures in the spring. Eighty-six students from Site 1 had complete data for the WJ IV-ACH, and were used in this study. For Site 2, researchers asked the classroom teachers to rank order students according to writing proficiency within each class, and the middle 50 students in each grade level were administered the WIAT-3. A total of 142 students had complete data from the WIAT-3 as a criterion measure after the spring CBM-W administration, and were used for this study. Our analysis, as described below, treated this subset as six independent samples, one sample from each grade at each site.

Table 2 describes the demographics for each of the six samples we used in our analysis.

Table 2. Demographics of Samples.

	Site 1			Site 2		
Grade	1	2	3	1	2	3
N	22	33	32	50	49	43
Male (%)	62	43	54	52	53	51
Female (%)	38	67	46	48	47	49
SpEd (%)	14	9	13	4	6	7
FRL (%)	57	30	50	64	55	37
ELL (%)	24	3	9	0	0	0

Note. Percentages rounded to nearest whole number. SpEd = Students receiving special education services; FRL = students receiving free or reduced-price lunch; ELL = English language learner.

Measures

Nationally normed assessments, the WJ IV-ACH (Houghton Mifflin Harcourt, 2014) for Site 1 and WIAT-3 (Psychological Corporation, 2009) for Site 2, were used as criterion measures to calculate the concurrent validity of PW for these studies. The subtests for the WJ IV-ACH were sufficient to obtain a Broad Written Language cluster score. Interrater reliability across all subtests of the WJ IV-ACH was above 90%. Spelling and Sentence Composition subtests were used for WIAT-3. Interrater reliability was above 90% across all subtests of the WIAT-3. The data used in this analysis were taken from the spring administration of both bench- marking studies. The students had been administered alternate forms of PW in the fall and winter. During the spring administration, they were given the PW prompt along with another writing CBM in a classroom setting. They were later administered the WJ IV-ACH or WIAT-3 individually.

The WJ IV-ACH is an individually administered battery of achievement tests designed for children, adolescents, and adults from ages 2 to 90 years. The WJ IV-ACH

contains subtests that measure five curricular areas: reading, mathematics, written language, oral language, and academic knowledge. The subtests are combined to form cluster scores. The following descriptions of subtests and clusters are derived from the manual for the WJ IV-ACH. All subtest and cluster standard scores have a mean of 100 and *SD* of 15.

In the Spelling subtest, the administrator reads a target word, reads a sentence that includes the word, and reads the target word again. The student is asked to spell the word. The words increase in difficulty with each number. The sub- test has a reported median reliability of .91 in the 5–19 age range. In the Writing Samples subtest, students are asked to write a variety of sentences. Items increase in difficulty, and are scored for passage length, vocabulary, and sophistication according to scoring standards outlined in the testing manual. The subtest has a reported median reliability of .90 in the 5–19 age range. In the Sentence Writing Fluency sub- test, students are asked to write simple sentences that contain three target words. This subtest has a 5-min time limit. The target words require higher levels of sentence complexity as the item numbers increase. The subtest has a reported median reliability of .83 in the 7–11 age range. Finally, the Broad Written Language cluster includes the Spelling, Writing Samples, and Sentence Writing Fluency subtests. It is a comprehensive measure of written language. The reported median reliability in the 5–19 age range is .95.

The WIAT-3 is a comprehensive assessment of student academic achievement designed for children in Grades pre-K through 12 (or ages 4–19 years 11 months; Pearson, 2009). The reported age-based reliability coefficients (ages 6–10 years) for the Spelling subtest are r = .94–.95, and those for the Sentence Composition subtest are r = .84–.90. For use as the criterion measure, the Spelling and Sentence Composition subtests were administered to 50 students in each grade level, ranked as being in the middle of their class according to teacher rating.

PW was used to assess transcription and text generation (e.g., translating ideas into coherent writing) skills at the sentence level. The administration of PW prompt involves providing one or more students with a packet containing 12 pictures with their accompanying words. Students are instructed to write one sentence for each item and to write as much as possible for 3 min. Each PW prompt was scored for WW, WSC, and

CWS. The stimuli were simple nouns or verbs presumed to be known by most first to third graders. There are 20 alternate forms of the picture word prompt; four forms were used in the current study. Each student completed two forms of PW, and mean performance was used for analysis.

Procedures

After the spring CBM-W administration, a subset of students from each site was administered the criterion measures. At Site 1, teachers selected subgroups of students who had high, average, and low writing ability to take the WJ IV-ACH. In Site 2, students were rank ordered within each class by their teacher, and the middle 50 of each grade level were given the WIAT-3.

WJ IV-ACH administrators were graduate students, including one nationally certified school psychologist who acted as an expert administrator and scorer for training and fidelity purposes. Administrators were trained by the expert. All administrators passed a mock administration in which the expert acted as a student to score administrators on adherence to basal, ceiling, and reversal rules, as well as idiosyncrasies of item administration.

A subset of administrators were trained as scorers in another training session. After the training session, all scorers were given a set of protocols to score. Criterion for interrater agreement (International Reading Association [IRA]) with the expert scorer was set at 85% of items scored for each subtest. This level was met in the first round of scoring by all scorers on the Spelling and Sentence Writing Fluency subtests. The Writing Samples IRA was below the threshold for all scorers. A second round of training commenced, after which all scorers were above 90%. After scoring, a sample of 10% of protocols scored for each scorer was examined for IRA. All scorers were above 90%. This study utilized the Written Language cluster, which includes the Spelling, Writing Samples, and Sentence Writing Fluency subtests. Form A was used for this study.

For the WIAT-3, Students were individually administered the Spelling and Sentence Composition subtests of the WIAT-3. Trained graduate students and one of the project coordinators from the larger study administered and scored all WIAT-3 assessments. Interrater reliability for scoring on the Spelling subtest ranged from 94% to

100% and on the Sentence Combining subtest from 92% to 100%.

PW probes were administered in the Fall, Winter, and Spring. For the purposes of this analysis, we used the probes that were administered concurrently with the WJ IV-ACH and WIAT-3. To obtain values for WWR and CWSR, the authors gathered the existing PW probe data (which had already been scored for WW and CWS), and counted the number of PW items to which students responded. WW and CWS were divided by the number of item responses to obtain WWR and CWSR. Any item attempted by a student was counted as a response, unless the student wrote only one word on the final item of a form as time ran out, in which case this item was not counted as a response. To measure interrater reliability, 62% of the PW forms were scored by one of three scorers. The three scorers had been previously trained in classic PW metrics, and were trained in scoring item responses. The training consisted of explaining the rules, modeling one form, and providing independent practice on additional forms until 100% agreement with the two lead authors was obtained. Training took no more than one practice form for 100% agreement to be obtained. To calculate interrater reliability, total agreements were divided by agreements plus disagreements and multi- plied by 100. Interrater reliability was high at 98%. Interrater reliability was 99.6% to 100% for WW and 94.7% to 97% for CWS (Lembke et al., 2014; McMaster et al., 2014). Thus, interrater reliability for WWR was 97.6% to 98%, and that for CWSR was 92.7% to 95%.

Data Analysis

One goal of this analysis was to compare the technical adequacy and concurrent validity of WWR and CWSR with those of CWS. To that end, we constructed regression models to determine how much variance in the criterion scores was accounted for by CWS, WWR, or CWSR. The models took the form of Equation 1, where the criterion metric score is predicted by the complexity metric:

negative correlations). We used a Fisher *r*-to-z transformation to account for potential negative skew in sampling distribution. Fisher's *r*-to-z transformations of each correlation were averaged and transformed back into *r* coefficients. The mean correlation between each metric and all criterion measures used across all sites and grades is depicted in Table 5 along with Fisher's *r*-to-z transformations and 95% confidence intervals.
The first research question is related to the interrater reliability and ease of scoring WWR and CWSR as compared with previous studies, and descriptions of other complexity measures such as trait-based rubrics. Assuming that the individual was already scoring PW for WW and CWS, the additional time for counting item responses and calculating either WWR or CWSR using Excel was negligible, adding about 3 to 5 s for each PW form scored. Scorer error associated with scoring CWSR is approximately 2% in addition to any error associated with scoring CWS. Thus, IRA for WWR was approximately equal to that for WW and CWSR to that for CWS. CWSR IRA ranged from 92.7% to 95%, which can be compared with 89% IRA and suspect internal reliability (alpha < .70) found by Allen et al. (2017) when applying a rubric to PW. Thus, both WWR and CWSR

$$\hat{Y}_{Criterion} = \beta_0 + \beta_1 X_{CBMW\ Metric}.$$

Results

Our first research question addressed the concurrent validity of our new complexity metrics. A Bonferroni correction was applied to account for the three analyses run per criterion measure for each data set; thus, alpha was set at .0167. CWSR accounted for a significant amount of variance in criterion measures in 12 of 12 analyses. WWR explained a significant amount of variance in the criterion measures in only three of the 12 analyses. Regarding our second research question, CWSR was the only metric that explained a significant amount of variance in all of the criterion measures used across all six data sets. CWSR also accounted for more variance than WWR or CWS in the criterion measures in all but two cases (third-grade WJ IV-ACH Broad Writing and third-grade WIAT-3 Spelling). CWSR was the most robust metric for predicting the criterion measure scores across the six independent data sets and four

criterion measures. CWS accounted for significant variance in only six of the analyses. Table 3 shows the results of the regression analyses. Table 4 shows a correlation matrix for all metrics and measures across both sites.

Given that correlations have maximum absolute values of 1, there is potential for negative skew when several samples show a positive correlation (or a positive skew for

Table 3. R^2 Values for Linear Regression Models Predicting Criterion Measures With CBMW Metrics.

	Site 1				Site 2		
Grade	CWS	WWR	CWSR	Grade	CWS	WWR	CWSR
Grade 1				Grade 1			
WJ Broad Writing				WIAT Spelling			
R^2	.08	.30*	.32*	R^2	.24**	.00	.30**
β_1	.32	7.11	6.5	β_1	.42	−.10	5.42
WJ Spelling				WIAT Sentence Composition			
R^2	.1531	.23	.39**	R^2	.06	.00	.16*
β_1	.46	6.87	7.95	β_1	.28	.26	5.06
Grade 2				Grade 2			
WJ Broad Writing				WIAT Spelling			
R^2	.16	.14	.36**	R^2	.46**	.28**	.52**
β_1	.36	4.86	6.62	β_1	.60	7.64	7.17
WJ Spelling				WIAT Sentence Composition			
R^2	.14	.10	.30**	R^2	.36**	.20**	.42**
β_1	.42	5.06	7.70	β_1	.60	7.44	7.08
Grade 3				Grade 3			
WJ Broad Writing				WIAT Spelling			
R^2	.46**	.14	.44**	R^2	.44**	.00	.28**
β_1	.58	3.92	6.30	β_1	.40	.74	4.00
WJ Spelling				WIAT Sentence Composition			
R^2	.40**	.14	.44**	R^2	.12	.03	.25**
β_1	.68	4.78	7.74	β_1	.34	2.16	6.00

Note. All values are rounded to two places. CBM-W = curriculum-based measures of writing; CWS = correct word sequences; WWR = words written per response; CWSR = correct word sequences per response; WIAT = *Wechsler Individual Achievement Test.*
*$p \leq .0166$. **$p \leq .0033$.

Given that correlations have maximum absolute values of 1, there is potential for negative skew when several samples show a positive correlation (or a positive skew for negative correlations). We used a Fisher *r*-to-z transformation to account for potential negative skew in sampling distribution. Fisher's *r*-to-z transformations of each correlation were averaged and transformed back into *r* coefficients. The mean correlation between each metric and all criterion measures used across all sites and grades is depicted in Table 5 along with Fisher's *r*-to-z transformations and 95% confidence intervals.

Table 4. Correlations of Metrics Across Sites and Grades.

Site 1						Site 2					
Grade 1	WWR	CWS	CWSR	WJ-BWL	WJ-Sp	Grade 1	WWR	CWS	CWSR	WIAT-Sp	WIAT-SC
WWR	1					WWR	1				
CWS	.14	1				CWS	−.11	1			
CWSR	.77	.52	1			CWSR	.58	.58	1		
WJ-BWL	.54	.30	.57	1		WIAT-Sp	−.01	.49	.54	1	
WJ-Sp	.48	.39	.63	.93	1	WIAT-SC	.02	.26	.40	.68	1
Grade 2	WWR	CWS	CWSR	WJ-BWL	WJ-Sp	Grade 2	WWR	CWS	CWSR	WIAT-Sp	WIAT-SC
WWR	1					WWR	1				
CWS	.26	1				CWS	.41	1			
CWSR	.86	.52	1			CWSR	.75	.78	1		
WJ-BWL	.39	.40	.60	1		WIAT-Sp	.52	.68	.73	1	
WJ-Sp	.32	.38	.55	.92	1	WIAT-SC	.46	.60	.64	.73	1
Grade 3	WWR	CWS	CWSR	WJ-BWL	WJ-Sp	Grade 3	WWR	CWS	CWSR	WIAT-Sp	WIAT-SC
WWR	1					WWR	1				
CWS	.17	1				CWS	.11	1			
CWSR	.88	.41	1			CWSR	.73	.55	1		
WJ-BWL	.37	.67	.67	1		WIAT-Sp	.52	.66	.52	1	
WJ-Sp	.37	.63	.66	.94	1	WIAT-SC	.50	.35	.50	.56	1

Note. WWR = words written per response; CWS = correct word sequences; CWSR = correct word sequences per response; WJ-BWL = *The WJ Broad Written Language subtest*; WJ-Sp = *WJ Spelling subtest*; WIAT-Sp = *WIAT Spelling subtest*; WIAT-SC = *WIAT Sentence Composition*.

Discussion

The first research question is related to the interrater reliability and ease of scoring WWR and CWSR as compared with previous studies, and descriptions of other complexity measures such as trait-based rubrics. Assuming that the individual was already scoring PW for WW and CWS, the additional time for counting item responses and calculating either WWR or CWSR using Excel was negligible, adding about 3 to 5 s for each PW form scored. Scorer error associated with scoring CWSR is approximately 2% in addition to any error associated with scoring CWS. Thus, IRA for WWR was approximately equal to that for WW and CWSR to that for CWS. CWSR IRA ranged from 92.7% to 95%, which can be compared with 89% IRA and suspect internal reliability (alpha < .70) found by Allen et al. (2017) when applying a rubric to PW. Thus, both WWR and CWSR require less time to score than either rubrics or T units, if the teacher is already scoring PW using WW and CWS, and both WWR and CWSR can be scored more reliably than rubrics previously examined with PW.

The second research question is related to the validity of CWSR as compared with WWR and CWS. Namely, CWSR should be more indicative of overall writing ability

than either WWR or CWS because CWSR incorporates some aspect of each component of CAF. We found that CWSR was a better predictor of norm-referenced criterion measures than WWR or CWS in all but two of our analyses. The only metric that significantly predicted each criterion measure across all samples was CWSR. This finding lends support to our hypothesis that a measure including some aspect of each component of CAF, such as CWSR, would be a more robust measure of overall writing performance than those that do not. In comparison, the study using a rubric with PW found a moderate concurrent correlation with the *WIAT-3 Spelling subtest* (r = .41, p < .01; Allen et al., 2017), while we found moderate strong correlation of CWSR with the same subtest at the same grade level (r = .54, p < .01). CWSR also had a moderate correlation with the *WIAT-3 Sentence Composition subtest* (r = .40, p < .01), while the rubric had a weak correlation with the same subtest (r = .30, p > .05). CWSR had higher correlations with the same subtests and grade level as the rubric employed by Allen et al. (2017). Thus, CWSR shows promise for demonstrating both stronger reliability and validity than the previously examined rubric.

It is not within the scope of this article to comment on what factors influence the variable predictive power of CWS and WWR, though for subtests such as Writing Samples in the WJ IV-ACH, it is possible that the variability may be influenced by items in which sentence complexity is an explicit scoring criterion. Exploring the characteristics that underlie the variability may be an interesting line of future study. Our findings complement previous work (Allen et al., 2017; Deno et al., 1982; Housen & Kuiken, 2009; Jewell & Malecki, 2005; Parker et al., 1991) by pro- viding evidence that complexity is a dimension of writing that can be accurately measured with PW CBM-W, and that CAF is a promising framework for future studies. Furthermore, CWSR can be quickly and reliably scored, provided teachers are scoring for CWS already. We can reasonably conclude that CWSR is a useful metric to consider when administering, scoring, and interpreting PW, particularly when considering the relative ease with which this metric can be collected. Future research should examine whether CWSR has utility for measuring change within the context of intensive writing intervention, and how that utility might compare with the established metric of CWS.

Limitations

This study represents an initial stage in a program of research examining complexity measures such as CWSR and WWR. We have no evidence for, and make no claims about, measurement of growth within the context of response to intervention paradigms. Furthermore, our sample from Site 2 was drawn from the median writers in their classrooms. The lack of variance due to the restricted range in those samples may limit the inferences that can be made. To increase the variability of the sample, future research on CWSR will need to include larger samples. It will also be important to examine rates of improvement and student growth on CWSR and WWR measures. Furthermore, considering that one of the primary objectives of CBM is to identify risk and inform instruction for students who are at risk, future studies should explore the reliability, validity, sensitivity to growth, and utility of CWSR and other measures incorporating all elements of CAF specifically for at-risk writers. This study provides evidence that each component of CAF should be carefully considered as we develop and validate new CBM-W scoring metrics, but the scope of our definition of complexity was limited. Specifically, while CWSR captures more complexity than simple CWS, sentence length does not capture depth of vocabulary or nuance in the author's passage. Future studies should consider employing lexical complexity as well as syntactic complexity because vocabulary knowledge has been shown to be a vital component to writing as well as reading comprehension.

As with any measure or metric, there are challenges implicit in scoring for CWSR. We mentioned that CWSR does not add appreciably to the difficulty, or detract appreciably from the reliability, of scoring for CWS. However, reliable and accurate scoring for CWS is in itself a skill that can take time and effort for practitioners to master. As researchers and practitioners are considering writing measures, these factors should be included in their decision- making process.

Finally, although CWSR shows promise as a measure of overall student performance in writing, it is not intended to replace the diagnostic information a rubric provides. Although rubrics may not be ideal for regular progress monitoring (weekly or biweekly), they can provide valuable diagnostic information as teachers select appropriate interventions, or decide that a change in instruction is needed. The use of

CWSR should be integrated into evaluations using rubrics to inform teacher decision-making and formative assessment.

Implications for Practice

Our analysis provides some evidence for the reliability and concurrent validity of CWSR as a metric within the context of PW. While this research program is in its early phases, we can make some claims with confidence. CWSR is a relatively good predictor of performance on nationally norm- referenced writing assessments. With the administration of a brief probe that can be scored quickly and reliably, teachers can obtain a measure of student writing that can be used for screening, and will continue to be useful as the student's writing increases in complexity. CWSR has the potential to be a useful tool for teachers of beginning writers.

Practitioners and researchers who are looking at CWSR through the lens of the CAF model and the Simple View of Writing can see that within the context of the PW prompt, the CWSR metric can be used to inform on some dimensions of complexity, accuracy, and fluency. That is, CWSR provides a measure of syntactic complexity as indexed by sentence length, but does not assess other forms of complexity such as those indexed by clausal density. CWSR, then, might provide a way for teachers to think about students' word choices in an effort to begin supporting their later text generation.

This study has provided some evidence that the incorporation of syntactic complexity can affect the validity of PW. Furthermore, CWSR was the most consistent and significant predictor of student performance on standardized writing assessments. That stated, there is no evidence thus far that CWSR or WWR is sensitive to growth, or could be used as effective progress monitoring measures. Thus, CWSR could help teachers by providing supplemental information in addition to CWS. Namely, if students are not showing growth in CWS but the teacher deems their sentences to be improving in complexity or they are showing growth in CWS but are producing more simplified sentences, then WWR or CWSR could provide additional information to teachers by indicating growth, or lack thereof, in syntactic complexity. The evidence provided here lends support to CWSR as being a valid measure of overall writing ability, and therefore

a worthy measure to inform instruction.

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 sup- port for the research, authorship, and/or publication of this article: The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A130144 to the University of Minnesota. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Abbott, R. D., Berninger, V. W., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. *Journal of Educational Psychology*, *102*, 281–298. doi:10.1037/a0019318

Allen, A. A., Poch, A. L., & Lembke, E. S. (2017). An exploration of alternative scoring methods using curriculum-based measurement in early writing. *Learning Disability Quarterly*, *41*, 85–99.

Berman, R. A. (2014). Linguistic perspectives on writing development. In B. Arfe, J. Dockrell & V. Berninger (Eds.), *Writing development in children with hearing loss, dyslexia, or oral language problems: Implications for assessment and instruction* (pp. 16–32). New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780199827282.003.0002

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.

Berninger, V. W., Vaughan, K. B., Abbott, R. D., Abbott, S. P., Rogan, L. W., Brooks, A., . . . Graham, S. (1997). Treatment of handwriting problems in beginning

writers: Transfer from handwriting to composition. *Journal of Educational Psychology*, *89*, 652–666. doi:10.1037/0022-0663.89.4.652

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65. doi:10.1016/j. jslw.2014.09.005

Campbell, H., Espin, C. A., & McMaster, K. (2013). The technical adequacy of curriculum-based writing measures with English learners. *Reading and Writing*, *26*, 431–452.

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

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

Deno, S. L., Marston, D., Mirkin, P., 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* (Research Report No. 87). Minneapolis: Institute for Research on Learning Disabilities, University of Minnesota.

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

Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Moving beyond total words writ- ten: The reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *School Psychology Review*, *31*, 477–497.

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., Berninger, V. W., Abbott, R. D., Abbott, S. P., & Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology*, *89*, 170–182. doi:10.1037/0022-0663.89.1.170

Hayes, J. R., & Berninger, V. W. (2014). Cognitive processes in writing: A framework. In B. Arfe, J. Dockrell & V. Berninger (Eds.), *Writing development in children with hearing loss, dyslexia, or oral language problems: Implications for assessment and instruction* (pp. 3–15). New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780199827282.003.0001

Houghton Mifflin Harcourt. (2014). *WJIV: Woodcock-Johnson Tests of Achievement*. Boston, MA: Author.

Housen, A., & Kuiken, F. (Eds.). (2009). Complexity, Accuracy and Fluency (CAF) in second language acquisition research [Special issue]. *Applied Linguistics*, 30, 461–473. doi:10.1093/applin/amp048

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*(1), 27–44.

Jones, D., & Christensen, C. (1999). The relationship between automaticity in handwriting and students' ability to generate written text. *Journal of Educational Psychology*, *91*, 44–49. doi:10.1037/0022-0663.91.1.44

Lembke, E., Allen, A., & Poch, A. (2014). Screening using *c*urriculum *b*ased *m*easurement in *w*riting in Grades 1-3: Missouri. Columbia: University of Missouri.

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.

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: Guilford Press.

McMaster, K., Brandes, D., Herriges, M., & Jung, P. (2014). *Screening using Curriculum*

Based Measurement in Writing in Grades 1-3: Minnesota. Minneapolis: University of Minnesota.

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

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.

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

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

Pearson. (2009). *Wechsler Individual Achievement Test* (3rd ed.). San Antonio, TX: Psychological Corp.

Psychological Corporation. (2009). *WIAT III: Wechsler Individual Achievement Test*. San Antonio, TX: Author.

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

Ritchey, K. D., McMaster, K. L., Al Otaiba, S., Puranik, C. S., Kim, Y., Parker, D. C., & Ortiz, M. (2016). Indicators of fluent writing in beginning writers. In K. Cummings & Y. Petcher (Eds.), *The fluency construct* (pp. 21-66). New York, NY: Springer.

Tindal, G., & Parker, R. (1991). Identifying measures for evaluating written expression. *Learning Disabilities Research & Practice*, *6*, 211–218.

Videen, J., Marston, D., & Deno, S. L. (1982). *Correct word sequences: A valid indicator of proficiency in written expression* (Vol. IRLD-RR-84, p. 61). Minneapolis: Minnesota University, Minneapolis Inst for Research on Learning Disabilities.