8-16-2017

An Exploration of Alternative Scoring Methods Using Curriculum-Based Measurement in Early Writing

Abigail A. Allen

Apryl L. Poch

Erica S. Lembke

An Exploration of Alternative Scoring Methods Using Curriculum-Based Measurement in Early Writing

Abigail A. Allen, PhD CCC-SLP[1], Apryl L. Poch, PhD[2], and Erica S. Lembke, PhD[3]

[1]Clemson University, SC, USA

[2]Duquesne University, Pittsburgh, PA, USA

[3]University of Missouri, Columbia, USA

Abstract

This manuscript describes two empirical studies of alternative scoring procedures used with curriculum-based measurement in writing (CBM-W). Study 1 explored the technical adequacy of a trait-based rubric in first grade. Study 2 explored the technical adequacy of a trait-based rubric, production-dependent, and production-independent scores in third grade. Results of Study 1 suggest that the rubric holds promise as a valid measure of sentence writing ability in first grade and has utility as a supplemental scoring procedure when using CBM-W as a screening tool. Results of Study 2 show that correct word sequences maintained the highest correlation coefficients across time with the trait-based rubric, but the other scoring procedures might offer promise as reliable alternative scoring methods. However, high internal correlations among the text features of the rubric along with highly variable interrater reliability suggest that caution must be taken in interpreting results.

Keywords

writing, assessment

Despite the importance of written communication in class- rooms and the workplace, writing performance of American students is alarmingly poor. On the most recent National Assessment of Educational Progress (NAEP), nearly three quarters of fourth, eighth, and 12th graders scored below the proficient level in writing (National Center for Education Statistics [NCES], 2011; U.S. Department of Education, 2003). For students with disabilities, these percentages increase to approximately 94% to 97% who scored at or below basic on the exam (NCES, 2011). Students with learning and writing disabilities are particularly disadvantaged in writing. Their writing has been characterized by "knowledge telling" in which they do not attempt to con- sider the needs of the reader in relaying (i.e., *telling*) every- thing they know about a topic; they spend little time preparing and developing their writing, and when revising, focus solely on producing accurate spellings (Graham & Harris, 2012). Educators need to be able to identify students who struggle in writing as early as possible in the elementary grades to intervene and adjust instruction to prevent later academic failure.

Models of Early Writing

Many researchers have developed models of early writing development of varying complexity. More streamlined models include Juel, Griffith, and Gough's (1986) model of ideas plus spelling ability. Their study found that word spell- ing and idea generation accounted for about 30% of the variance in writing quality in first and second grade after controlling for IQ and oral language ability. Berninger and Amtmann's (2003) Simple View of Writing advanced Juel and colleagues' model by dividing writing ability into three main skill areas: text generation (turning ideas into words), transcription (spelling and handwriting), and self-regulation skills, with working memory skills constraining all three areas. Other more complex models of writing include revising and editing skills (Hayes, 1996; Hayes & Flower, 1980); productivity, syntactic complexity, and substantive quality (Kim et al., 2011); and macroorganization (Wagner et al., 2011). What these models have in common for young writers is that transcription skills, or spelling and handwriting, are considered critical to later writing development and proficiency. Although these skills do not encompass the entirety of the domain of writing, they are important early skills that relate to later writing proficiency and are the basis for

methods of evaluating students' writing performance in the early elementary grades. One such method of evaluating early writing and identifying students at risk for writing difficulties is curriculum-based measurement in writing (CBM-W).

Research on CBM-W

CBM-W was developed to provide an objective way to assess student writing and identify those in need of intervention. CBM refers to a set of reliable and valid procedures that allow an educator to directly observe and score student performance on standardized tasks that represent indicators of overall proficiency in an academic area. The measures are meant to be quick to administer, easy to score, inexpensive, and standardized across items and scoring procedures to allow for comparisons within and across class- rooms (Deno, 2003). Different writing tasks along with various scoring procedures have been the focus of studies that examine CBM-W.

CBM-W Tasks

Studying CBM-W at the sentence and paragraph levels pro- vides data to use when evaluating students at risk, given that features such as fluency (number of words written), spelling, grammar, punctuation, sentence structure, and semantics can be assessed. Lembke, Deno, and Hall (2003) and Hampton and Lembke (2016) studied both copying and dictation measures at the sentence level and found that sentence dictation tasks had higher criterion-related validity (r = .80–.92) in second grade than other tasks such as sentence copying. McMaster, Du, and Petursdottir (2009) also found that sentence copying was a reliable (r > .70) and valid (r > .50) measure of writing in first grade. Novel sentence writing tasks, where students must write at least one sentence in response to a picture, have demonstrated adequate reliability (r > .70) and criterion-related validity (r > .60) for use in kindergarten through third grade (Coker & Ritchey, 2013; Deno, Mirkin, & Marston, 1980; Deno, Mirkin, Marston, & Lowry, 1982; Lembke et al., 2003; McMaster & Campbell, 2008; McMaster et al., 2009; McMaster et al., 2011). Story prompts are per- haps the most common CBM-W, serving as an indicator of students' ability to write connected text. Although story prompts have technical adequacy at first through third grade, they are considered most appropriate and reflective of

student ability at third grade and above (Deno et al., 1980; Deno et al., 1982; McMaster et al., 2009; McMaster et al., 2011).

Production-Dependent Scoring Methods

Traditionally, CBM-W tasks have been scored with production-dependent methods (e.g., number of words written), meaning that students' scores are dependent upon the quantity of text written. Production-dependent scoring methods that have been studied most frequently with sentence writing and story prompt tasks are (a) words written (WW; any sequence of letters separated by a space from another sequence of letters), (b) words spelled correctly (WSC; any English words that are spelled correctly, regardless of con- text), (c) correct word sequences (CWS; two adjacent words spelled correctly and used correctly in context), and (d) correct minus incorrect word sequences (CIWS; total correct word sequences minus incorrect sequences, which accounts for incorrect sequences directly) (Deno et al., 1982; Parker, Tindal, & Hasbrouck, 1991; Videen, Deno, & Marston, 1982). Research suggests that WW, WSC, CWS, and CIWS are the most reliable and valid scores in the elementary grades compared with other indices like number of long words (e.g., words with seven or more letters; Deno et al., 1980; Lembke et al., 2003; McMaster et al., 2009; McMaster et al., 2011; Ritchey & Coker, 2013). Research by Gansle and colleagues (Gansle, Noell, VanDerHeyden, Naquin, & Slider, 2002; Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006) found questionable reliability (r = .59–.65) but acceptable criterion-related validity (r = .36–.44) for indices such as number of complete sentences (Gansle et al., 2002; Gansle et al., 2006). Gansle et al. (2006) also noted that alternative scoring indices have important social validity for teachers, who state that measures like WW, for example, do not necessarily relate to writing quality and provide little useful information for intervention planning.

Results from studies on production-dependent indices suggest that as students grow older and writing becomes more automatic, measures that tap proficiency, such as CWS and complete sentences, become more useful. This area of research has been expanded with the study of pro- duction-independent indices.

Production-Independent Scoring Methods

Production-independent scores focus on quality over quantity and are averaged over the length of a student's writing. Measures include percent of words spelled correctly (%WSC) and percent of correct word sequences (%CWS). These indices capture the differences between the transcription-focused instruction in the primary grades and the higher level skills required in writing at later grade levels. Parker et al. (1991) and Jewell and Malecki (2005) found that %WSC and %CWS had stronger criterion-related validity coefficients than the production-dependent measures at earlier grades but were less sensitive to growth over time. These results suggest that production-independent scoring procedures may better represent proficiency in spelling and grammar in the elementary grades as opposed to a pure production score. However, more research is needed on these types of scores to determine their utility in earlier grades.

Qualitative Scoring Methods

There is emerging research into using qualitative scoring methods such as holistic ratings and trait-based rubrics as complimentary scoring procedures for use with CBM-W tasks. When utilizing holistic ratings, a writing sample is assigned a number on an arbitrary scale (e.g., from 1 to 7) to indicate the quality and proficiency of the piece as a whole. Although relatively time efficient, holistic scoring does not provide information about where or how the student struggled, which means the score has little instructional utility (Crusan, 2015). In contrast, trait-based rubrics present a series of writing traits deemed important for proficiency (e.g., mechanics, spelling, word choice), and an evaluator must assign points to each individual trait on a given scale (e.g., from 1 to 4). By evaluating individual aspects of writing, trait-based rubrics are intended to increase objectivity, reliability, validity, and instructional utility of scores (Crusan, 2015). To provide more complete information about student writing performance, educators may feel that they need a more qualitative scoring system in addition to quantitative scoring procedures. Using qualitative scoring in isolation is often less reliable (Gansle et al., 2006), thus the combination of qualitative and quantitative scoring methods may prove to be a powerful combination for assessing student writing performance.

In general, studies have found that qualitative and quantitative scoring procedures had very weak to strong correlations with each other across grade levels (r = .35–.76, Coker & Ritchey, 2010; r = .36–.37, Gansle et al., 2002; r = .27, Gansle et al., 2006; r = .06–.67, Lembke et al., 2003; r =.50–.60, McMaster et al., 2009; r = .34–.70, Parker et al., 1991; r = .29–.50, Ritchey & Coker, 2013; r = −.02–.63, Tindal & Parker, 1991). However, a handful of studies found that in the early primary grades, production-dependent measures correlated more strongly with qualitative measures than at upper grade levels. Lembke et al. (2003) found a very strong correlation coefficient between holistic ratings and WSC on a word dictation task in second grade (r = .83). Videen et al. (1982) found that in third through sixth grade, CWS correlated very strongly with holistic ratings on a story-writing task (r = .85). Coker and Ritchey (2010) found a very strong correlation coefficient between trait- based rubric scores and WSC in the spring of kindergarten (r = .82). Tindal and Parker (1991) found that a trait-based rubric correlated with quantitative production-dependent indices (WW, WSC, CWS) more strongly in the spring of third grade (r = .36–.63) than in fourth and fifth grades (r = .10–.47), possibly because students in third grade are still developing their transcription skills.

Although transcription skills are important to the early primary grades, a comprehensive assessment system should also account for writing proficiency, complexity, and quality. Thus, it is possible that the use of a qualitative measure in the form of trait-based rubrics in CBM-W scoring may provide teachers with both a reliable and valid measure of student writing ability when used in combination with pro- duction-dependent and -independent measures. More research is needed to determine exactly what dimensions of writing can and should be evaluated in this manner.

Purpose

The purpose of this study was to investigate the technical adequacy of alternative scoring procedures for CBM-W sentence writing and story prompt tasks in first and third grade. First, the general procedures and methods that are common to both studies will be presented; then, research questions, study-specific methods, results, and discussion for each study will be presented separately. We will conclude with a larger discussion of

implications for practice.

General Method

Participants and Settings

Both first- and third-grade samples were drawn from a larger CBM-W screening study of 338 students that included students in first (*n* = 96), second (*n* = 118), and third grade (*n* = 124). The larger study took place across 27 classrooms in two elementary schools within a school district in a small Midwestern city. A subset of 50 students in each grade level was administered a criterion measure (see “Measures” section). For the current study, a random selection of students from first (*n* = 40) and third grades (*n* = 42) who took the same CBM-W forms was included. First-grade participants were randomly sampled from the subset of 50 students who took the criterion measure. Third-grade participants were randomly sampled across four classrooms; therefore, only 10 third-grade participants took the criterion measure. Demographic information on the students is provided elsewhere.

Measures

Picture word.

Picture Word (PW) CBM-W prompts require students to write one sentence for each picture presented. PW forms contained 12 pictures of common objects and actions paired with the corresponding written word below. Note that a “sentence” was defined as any series of words a student wrote in response to a picture (McMaster et al., 2009). Whether the response constituted a complete sentence was evaluated with the trait-based rubric.

Story prompt.

Story Prompt (SP) CBM-W prompts require students to construct a brief story using a story starter. Each prompt contained an open-ended sentence using simple vocabulary and sentence structure about a topic that most students attending U.S. public schools would have experienced (e.g., “One day, we were playing outside the school and . . .”).Both the PW and SP tasks were scored using the produc- tion-

dependent scores from the larger CBM-W study. Specifically, student responses were scored for WW, WSC, CWS, and CIWS.

Weschler Individual Achievement Test, Third Edition (WIAT-3).

A subsample of 50 students per grade level in the larger screening study was given the *Spelling* and *Sentence Composition* subtests of the WIAT-3 (Weschler, 2009) in May of the academic year. All first-grade participants and 10 third-grade participants took the WIAT-3. Prior to spring data collection, teachers rank-ordered all classroom students 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. Participants in this stratified sample completed the criterion measure after completing all other CBM-W tasks. The *Spelling* subtest has students write letter sounds and single words from dictation. The *Sentence Composition* subtest is made up of two tasks: Sentence Combining, where students are given two simple sentences to combine into one sentence, and Sentence Building, where students create a sentence using a specific word given by the examiner (e.g., "or," "than," etc.).

Procedures

Prompt administration.

Both PW and SP CBM-W were administered at three time points during the 2013–2014 school year: November/December (fall), February (winter), and April (spring). Each classroom was assigned a combi- nation of two out of four possible forms for PW and SP at each time point. Combinations were counterbalanced to control for order effects and were stratified as evenly as possible across grade levels. Classes, not students, were assigned to each form. Both prompts were group administered for 3 min and were completed during students' regularly scheduled writing instruction in the general education classroom. For the PW prompt, students were instructed to "Write a sentence for each picture/word in your packet." The examiner read each word aloud to the class before allowing the students to begin writing. For SP, the students were read the story prompt and asked to think about their story for 30 s before responding for 3

min.

WIAT-3.

The *Spelling* and *Sentence Composition* sub- tests of the WIAT-3 were given individually to all first-grade and 10 third-grade participants in the current study. Six trained graduate students and one project coordinator from the larger study administered and scored all WIAT-3 assessments. The project coordinator was a licensed school psychologist with 30 years of experience and served as the expert scorer and trainer. Interrater reliability (IRR) for scoring was conducted on 20% of each subtest (*Spelling*: 94%–100%; *Sentence Composition*: 92%–100%).

Production-dependent scoring.

Scorers were advanced doctoral students and were trained during the larger screening study by a professor in special education. In the larger screening study, 20% of student assessments were scored for IRR (mean: 98%.)

Study 1: First Grade

Research Questions

Research Question 1: What is the reliability and criterion-related validity of scores obtained from a trait-based sentence writing rubric in first grade?

Research Question 2: To what extent do scores obtained from a trait-based rubric demonstrate student growth in first grade?

Method

Participants and setting.

Participants included 40 first-grade students who were randomly selected from the 50 students given the WIAT-3. The sample was drawn from seven classrooms across two K-5 elementary schools. Demo- graphics of this sample were as follows: 21 males (53%), 19 females (47%), 60% White, 28% Black, 8% Hispanic, and 3% Asian. In total, 58% were eligible for free/reduced-price lunch, 8% received services for gifted students, 3% received services for special education, and no students received ser-

vices for English Language Learners (ELL).

Measures

CBM-W.

For this study, one form of PW was analyzed for each student at each time point, resulting in 120 writing samples.

Trait-based rubric.

The sentence writing rubric (see Figure 1) was adapted from Coker and Ritchey's (2010) rubric used with kindergarten students. Each sentence was scored from 0–3 on each of three dimensions: (a) sentence type, (b) spelling, and (c) grammatical structure. A fourth dimension, mechanics, was scored 0–2. Students could score a total of 11 points on the rubric for each sentence written. The final rubric score for a prompt was averaged across all sentences written. Rubric adaptations included changing "Response Type" to "Sentence Type" to better fit the PW task and changing the total points in mechanics. The original rubric awarded 3 points for capitalizing proper nouns but because PW only uses common nouns, the points were altered to avoid penalizing students for words not used in the PW task.

	3	**2**	**1**	**0**
Sentence Type	Compound or complex sentence	Complete simple sentence	At least one word to several words	No legible words
Spelling	All words spelled correctly	51-99% of words spelled correctly	≤ 50% pf words are spelled correctly	No words spelled correctly
Mechanics	n/a	Initial capital letter AND correct punctuation	Initial correct capital letter OR punctuation	No use of capitals or punctuation
Grammatical Structure	Sentence is grammatically correct	One grammatical error; meaning is not changed (e.g. plurals, verb tense)	2 or more errors; OR errors change meaning (e.g. fragments)	Multiple errors OR sentence meaning is unknown

Figure 1. First grade trait-based rubric.

Source. Adapted with permission from Coker and Ritchey (2010).

WIAT-3.

The sample of first graders for this study (*n* = 40) was drawn from the subsample of first graders (*n* = 50) who took the WIAT-3 in May of the academic year.

Procedures.

All writing samples were previously scored using production-dependent scores (WW, WSC, CWS, CIWS) from the larger screening study. The original pro- duction scores from the larger study were used for this analysis. Rubric scoring was completed by two advanced doctoral candidates in special education trained by the first author. In total, 25% of the writing samples were scored for IRR by the two doctoral students (see "Results" section).

Data analysis.

To answer the research question, "What is the reliability and criterion-related validity of scores obtained from a trait-based sentence writing rubric in first grade?" IRR was calculated as a percentage of agreements (number of agreements divided by total agreements and disagreements) on both the overall rubric score and the individual trait scores. Cronbach's alpha was calculated to measure internal consistency reliability of the rubric measure. To evaluate criterion-related validity, a series of Pearson product–moment correlations was calculated with the fall, winter, and spring rubric scores; the production-dependent scores; and the spring administration of the WIAT-3.

To answer the research question, "To what extent do scores obtained from a trait-based rubric demonstrate student growth in first grade?" paired-sample *t* tests were calculated for fall–winter, winter–spring, and fall–spring rubric score differences to determine whether student growth on the rubric between time points was statistically significant.

Descriptive statistics (mean, *SD*) were also calculated for all measures. All data were scored and entered into SPSS (v. 22.0) for analysis.

Results

Descriptive data.

Visual examination of data histograms and skewness and kurtosis values indicated that the distribution of the PW data was approximately normal. Descriptive statistics are reported in Table 1. Participants showed growth on all scoring procedures and the trait-based rubric over time. The average number of responses per PW prompt, also increased over time (fall: *M* = 5.34, *SD* = 2.31, range = 2–11; winter: *M* = 6.28, *SD* = 2.64, range = 1–12; spring: *M* = 7.43, *SD* = 2.99, range = 2–12).

Reliability

Interrater reliability.

Interrater reliability (IRR) was con- ducted on a random sample of 25% of samples by doctoral students trained in rubric scoring (see “Procedures” section). Results were 91% agreement for response type, 90% for spelling, 93% for mechanics, and 82% for grammatical structure, for a total interrater reliability of 89%. There was some disagreement between raters on the grammatical structure trait regarding whether to score an incomplete sentence consisting of a few words as a fragment, which receives 1 point, or as a sentence with multiple errors or an unknown meaning, which receives 0 points.

Table 1. First-Grade Descriptive Data and Criterion-Related Validity Coefficients.

	Fall			Winter			Spring		
Measure	*M*	*SD*	Validity coefficient	*M*	*SD*	Validity coefficient	*M*	*SD*	Validity coefficient
Rubric	7.59	0.89	—	7.71	1.13	—	8.35	1.06	—
WW	22.00	8.98	.45**	24.88	9.09	.35*	30.43	11.05	.45**
WSC	18.00	7.92	.55**	20.10	8.41	.50**	26.85	11.21	.56**
CWS	14.73	9.20	.74**	17.40	9.64	.68**	27.50	15.53	.73**
CIWS	2.75	10.93	.81**	4.40	12.51	.77**	16.75	18.54	.83**
WIAT-3[a] *Spelling*	—	—	.16	—	—	.41**	103.50	10.43	.41**
WIAT-3[a] *Sentence Composition*	—	—	.21	—	—	.38*	101.45	16.15	.30

Note. WW = words written; WSC = words spelled correctly; CWS = correct word sequences; CIWS = correct minus incorrect word sequences; WIAT = *Weschler Individual Achievement Test*.

[a]WIAT-3 age-normed scores were used.

*$p < .05$. **$p < .01$.

Internal consistency reliability.

Cronbach's alpha was calculated for the rubric dimensions at each of the three time points. Alphas were .29 in fall, .50 in winter, and .64 in spring. The accepted minimum criterion level for CBM reliability is α = .80 (National Center on Response to Intervention, 2010). The spring coefficient was closest to this criterion, but at no time point did the rubric demonstrate evidence of adequate internal reliability, indicating it may be measuring multiple related constructs.

Criterion-related validity

CBM-W production scoring. Validity coefficients are reported in Table 1. All coefficients were statistically significant; however, only the correlations between the rubric, CWS, and CIWS were above the generally acceptable $r \geq .60$ criterion for criterion-related validity (McMaster et al., 2009). The rubric, CWS, and CIWS had moderate to strong correlations at each time point (r = .68–.83).

Criterion measure.

The coefficients of rubric scores and the WIAT-3 subtests are reported in Table 1. The predictive criterion-related validity of the total rubric scores to both subtests of the WIAT-3 was weak in fall (*Spelling*: r = .16; *Sentence Composition*: r = .21; $p > .05$) and weak to moderate in winter (*Spelling*: r = .41, $p \leq .01$; *Sentence Composition*: r = .38, $p \leq .05$). The concurrent criterion-related validity of the rubric with the WIAT-3 was again weak to moderate (*Spelling*: r = .41, $p \leq .01$; *Sentence Composition*: r = .30, $p > .05$).

Growth.

A series of paired-sample t tests was conducted to determine whether student growth on particular measures over time was significant. Results of the t tests indicate statistically significant growth on rubric scores from winter to spring, $t(39) = -3.74$, p = .001, and fall to spring $t(37) = -4.99$, $p < .001$, but not from fall to winter.

Discussion: Study 1

The purpose of the current study was to determine the technical adequacy of a

trait-based sentence writing rubric for use with CBM-W in first grade. To investigate this, interrater and internal consistency reliability, criterion-related validity coefficients, and growth were examined.

Although the mean number of total responses may seem high for a first-grade sample, ranging from approximately five in the fall to seven in the spring, it is important to note two things. First, on the PW task, a "sentence" is defined as any series of words written in response to a picture. While students are instructed to "write a sentence for each picture," their writing is scored as, whether the response is a complete sentence or a series of unrelated words (McMaster et al., 2009). Therefore, a first grader writing five "sentences" means she or he wrote something about five of the pictures included in the PW probe. This does not necessarily mean that the responses were long (e.g., "I like paper." "I like pants." "I see a dog.") or even complete sentences (e.g., "blue hat," "ride horse"), but this illustrates why a trait-based rubric can contribute important instructional information in concert with production-dependent scores. It allows educators to look more deeply at a student's responses to a PW probe and determine whether they are using complete sentences that are coherent, which the pro- duction-dependent scoring methods used in CBM-W do not adequately address (Tindal & Parker, 1991).

Second, the sample was drawn from the first-grade participants who were given the WIAT-3, which was a stratified sample of high-, middle-, and low-performing writers. It is likely that the mean number of total responses per PW probe is higher than one might expect from a sample of solely low-achieving students.

This study contributes to the emerging literature supporting using trait-based rubrics with CBM-W. The rubric demonstrated evidence of adequate interrater reliability (89%) but not internal consistency reliability (.29–.64). Reliability results were lower than in previous studies (IRR: 98%, α = .84–.89, Coker & Ritchey, 2010). This could first be due to a small sample. The current study included 40 participants while previous studies of trait-based rubrics had sample sizes of more than 200 (Coker & Ritchey, 2010; Tindal & Parker, 1991). It could also be due to the length of the rubric. A maximum total score of 11 points leaves little room for variance across students, which could explain the weaker reliability coefficients. Perhaps a longer or more complex rubric might show evidence of greater reliability. Although the current study lends some additional

interrater reliability evidence to the literature on trait-based writing rubrics, more research is needed to obtain stronger reliability and draw conclusions about the utility of the rubric.

The criterion-related validity results were mixed. In terms of criterion-related validity with the CBM-W production scores, the rubric showed evidence of moderate to strong validity with CWS and CIWS scoring procedures across all time points (r = .68–.83) which is consistent with findings from previous studies (r = .35–.76, Coker & Ritchey, 2010; r = .36–.54, McMaster et al., 2009; r = .36– .63; Tindal & Parker, 1991). However, this could simply be because CWS and CIWS measure the same things as Spelling and Mechanics on the rubric.

Criterion-related validity coefficients with the standardized writing test were lower. The rubric demonstrated evidence of a moderate relationship at best with the WIAT-3 *Spelling* subtest in winter (r = .41) and spring (r = .41) and a weak relationship with the *Sentence Composition* subtest at all time points (fall: r = .21; winter: r = .38; spring: r = .30). This means that the rubric appears to be tapping some element of transcription ability but not text generation. This could be related to how the rubric and *Sentence Composition* subtest are scored. Both the Sentence Combining and Sentence Building tasks score student responses according to semantics; that is, part of a student's score is related to the meaning of their sentences. Either they retained the original meaning of the two sentences (in Sentence Combining) or used the given word correctly in their sentence (in Sentence Building). The rubric and the CBM-W production scores, however, do not require that a student use the target word in their sentence or that they respond in any particular way to the prompt. Although the WIAT-3 and the rubric both assess elements of grammar, syntax, and spelling, the WIAT-3 explicitly assesses meaning while the rubric does not, which may have contributed to the weaker relationship. The more moderate relationship with the *Spelling* subtest however indicates that the rubric is indeed tapping some construct related to transcription ability. For now, we must interpret the criterion-related validity evidence of the rubric with caution. The rubric may hold promise as a valid measure of sentence writing ability in first grade and may have utility as a supplemental scoring procedure when using CBM-W as a screening tool, but more research is needed. Future iterations

should look into incorporating a semantic trait into the rubric to further investigate the criterion-related validity of the measure and its representation of text generation ability.

The rubric scores also showed potential as a way to measure growth over an academic year. Statistically significant growth was shown on the rubric scores from winter to spring and fall to spring. In this sample there was not significant growth from fall to winter. This could be because the fall (November/December) and winter (February) time points were too close together. Additional research into rubric scoring with more evenly-spaced time points would be warranted. It is worth noting that no student received the maximum score possible on the rubric, indicating no ceiling effect and potential use as a progress monitoring measure to track student growth.

Study 2: Third Grade

Research Questions

Research Question 1: What is the reliability, criterion-, and construct-related validity of scores obtained from a trait-based story-writing rubric in third grade?

Research Question 2: To what extent do scores obtained from a trait-based scoring rubric demonstrate student growth in third grade?

Method

Participants and setting.

Participants were 42 third-grade students from four classrooms, selected at random from the larger screening study, who completed the same SP task (see "General Method" section). There were 25 female students (60%) and 17 male students (40%); 62% received a free/reduced-price lunch; 55% were White, 36% were African American, and 1% were Hispanic or multiracial. To exert an added level of control over the data given the large number of scoring indices, students who received special education, ELL, or gifted/talented services were excluded.

Measures

Story prompt (SP).

For this study, data from only one form (Form A) was analyzed for each student at each time point, resulting in 126 writing samples. The SP read, “One day, we were playing outside the school and . . .” (McMaster & Campbell, 2008; McMaster et al., 2009).

Alternative scoring methods.

Several production-dependent and production-independent measures were examined. Each is described further in Table 2.

Table 2. Definition of Production-Dependent and Production-Independent Scoring Indices

Scoring indices	Abbreviation	Definition
Words written	WW	The total number of words written; a “word” is a sequence of letters separated by a space from another sequence of letters.[a]
Words spelled correctly	WSC	The number of correctly spelled words regardless of context.[a]
Correct word sequences	CWS	Any two adjacent words that are spelled and used correctly in context.[a]
Incorrect word sequences	IWS	Any two adjacent words that are not spelled and used correctly in context.
Correct-incorrect word sequences	CISW	The number of correct word sequences minus incorrect word sequences.
Total punctuation marks	TPM	The number of total punctuation marks used (e.g., period, exclamation point).
Correct punctuation marks (context)	CPM_Context	The number of correct punctuation marks that are used correctly in context.
Incorrect punctuation marks (context)	IPM_Context	The number of incorrectly used punctuation marks in the context of the sample.
Correct-incorrect punctuation marks (context)	CIPM_Context	The number of correct punctuation marks minus incorrect punctuation marks in context.
Correct punctuation marks (no context)	CPM_No Context	The number of correct punctuation marks minus incorrect punctuation marks in the sample.
Complete sentences	CS	The number of complete sentences in the student’s response. A complete sentence must start with a capital letter,

		have a subject (i.e., noun), have a predicate (i.e., verb), and end with punctuation[b].
Words in complete sentences	WiCS	The total number of words in all sentences that were scored as complete sentences. Words did not have to be spelled or used correctly[b].
Number of error words	#EW	The total number of words written minus the number of words used in complete sentences[c].
Number of incorrect words	#IW	The total number of words spelled or used incorrectly in context.
Percent of correct punctuation marks (context)	%CPM_Context	The number of correct punctuation marks divided by the total number of punctuation marks (correct plus incorrect punctuation marks) multiplied by 100. Placement of punctuation was consistent with the context of the writing sample.
Decimal percent of correct punctuation marks (context)	Dec% CPM_Context	The decimal conversion of the percentage.
Percent of correct punctuation mark (no context)	%CPM_No Context	The number of correct punctuation marks divided by the total number of punctuation marks (correct plus incorrect punctuation marks) multiplied by 100. Context of the sample did not influence scoring of punctuation.
Decimal percent of correct punctuation marks (no context)	Dec% CPM_No Context	The decimal conversion of the percentage.
Mean length of correct word sequences	ML_CWS	The "average length of all continuous strings of CWSeq [correct word sequences][a]." ML_CWS is derived by dividing the number of correct word sequences by the number of sets of correct word sequences.

[a] Definition consistent with Parker, Tindal, and Hasbrouck (1991).

[b] Definition consistent with Gansle, Noell, VanDerHeyden, Naquin, and Slider (2002). cRecommended by Gansle et al. (2002).

Trait-based rubric.

A 4-point (score 0–3) trait-based writing rubric (see Figure 2), that was meant to mimic what a teacher might use, was used as the criterion measure in this study. The rubric was researcher-developed (by the second author) and was influenced by trait-base from CBM-W literature (Coker & Ritchey, 2010), the 6 + 1 Traits Scales (Northwest

Regional Educational Laboratory [NWREL], 2000), the writing criteria from the National Writing Project, and the New York State Education Department's rubric from the Regents Examination in English Language Arts.

Figure 2. Third-grade trait-based rubric.

Qualitative Score Rating				
Text Features	**3**	**2**	**1**	**0**
Sentence Fluency / Structure	Includes multiple sentences using a variety of sentence types OR a single sentence with clauses and phrases, OR a compound sentence.	Includes a complete sentence OR multiple thoughts that create a run-on sentence. Sentences begin the same way.	Relies on short, choppy sentences OR includes at least one word to several words. Reader may have to reread.	Has no legible words OR an unclear response OR no response.
Spelling and Word Choice	All words are spelled correctly OR only makes errors when using sophisticated language.	More than half of the words are spelled correctly; errors reflect phonetic spelling.	Half or fewer than half of the words are spelled correctly; attempts semi- phonetic spelling.	No words are spelled correctly; letter strings pre-phonetic, OR no response.
Mechanics and Conventions	Includes an initial capital letter and correct punctuation OR there are multiple sentences and all include initial capital letter and punctuation. Proper nouns are capitalized.	Includes an initial capital letter and correct punctuation for most sentences. Some errors in conventions. Most proper nouns are capitalized.	Includes only initial correct capital letter or punctuation, OR uses random capitalization and punctuation. Few, if any, proper nouns (if used) are capitalized.	No use of capital letter or punctuation for sentences OR response is not a sentence OR capitalization and punctuation intermittent OR mixed upper and lowercase letters.
Grammatical Structure	Sentence or sentences are grammatically correct.	No more than 1 grammatical error that does not change sentence meaning OR minimum errors that do not impede understanding.	Includes multiple grammatical errors or errors that change sentence meaning.	Includes multiple grammatical errors; sentence meaning is unknown OR response is not a sentence.
Relationship to and Completeness of Prompt	Directly and appropriately linked to the prompt and includes multiple elaborations.	Directly answers the question without elaboration (e.g., the student finished the story prompt	Linked to at least one idea or the general theme of the prompt, but response may only contain a few words.	Not related to prompt, OR unclear response, OR no response.
Ideas	Ideas narrowed, focused, and tells a story.	Ideas broad and generally on topic	General idea is understandable, but some ideas	Ideas, if evident, are unclear or unrecognizable.

			are unclear.	
Organization	Response is highly organized, easy to follow, and uses sophisticated transitions.	Organizational structure is evident; transitions are attempted but relies on “First,” “Second,” etc.	Organization is somewhat evident (e.g., may have a beginning but ends with “The End.”).	Organization not apparent. No clear beginning and ending.
Voice	Takes risks to say more than what is expected, writes with clear sense of audience, point of view is evident.	Individual perspective becoming evident, some awareness of audience, but primarily writes to complete task.	Reader has limited connection to writer, moments of sparkle, but it’s quickly hidden.	Response is unclear which limits creating a connection with the reader, and/or aware- ness of audience not evident.
Development	Interesting, important details provide support.	Details lack development – may appear list-like.	Some details clear while others are fuzzy.	Little if any development using details provided.

The trait-based rubric for this study utilized nine different text features. Each text feature was scored based on the corresponding criteria provided, and the student’s score for each text feature was summed together for a composite score. The maximum score a student could receive was 27.

Social validity.

In addition, a convenience sample of third-grade teachers from two elementary schools within a rural school district in Western New York was asked to review the trait-based rubric. Teachers responded to three questions: (a) What do you like about the rubric? (b) What changes would you make to the rubric? and (c) Would such a rubric be helpful in your writing instruction to evaluate student performance? Why or why not? Although this data were not iteratively used to develop the rubric for the cur- rent study, the data provide valuable information about educator perceptions and will be used in future iterations of this work.

Procedures.

All SP assessments were rescored for production-dependent (e.g., WW, WSC, CWS, CIWS), production-independent (see Table 2), and rubric scores by the second

author. IRR was completed on 26% of the sample across all scoring methods with an advanced doctoral student who was trained in production-dependent scoring from the larger screening study and on production-independent and rubric scoring by the second author.

Data analysis.

Descriptive statistics (mean, SD) were also calculated for all measures. All data were entered into SPSS (v. 22.0) for analysis. To answer the research question, “What is the reliability, criterion-, and construct-related validity of scores obtained from a trait-based story-writing rubric in third grade?” Pearson product–moment correlations were calculated to explore the strength of the relation- ship between the rubric, production scores, and WIAT-3 scores and the individual traits of the rubric with the total rubric score. Cronbach’s alpha was calculated to examine the internal consistency reliability of the rubric. IRR was assessed by calculating the total number of agreements in scoring divided by the total number of agreements plus total disagreements. To answer the research question, “To what extent do scores obtained from a trait-based scoring rubric demonstrate student growth in third grade?” paired-samples t tests were calculated across time points.

Results

Descriptive data. Based on visual inspection of data histograms, the distribution of the data was approximately nor- mal. However, some scores exhibited slightly or extremely elevated skewness and kurtosis values. Participants grew on all scores across time points (see Table 3).

Reliability

Interrater reliability (IRR).

Across all time points, mean IRR for production-dependent scores was 90% (99% WW, 98% WSC, 87% CWS, 74% CIWS), 84% for the remaining production scores (mean range: 79%–89%), and 79% for the rubric (due to space constraints, full results are available upon request.).

Table 3. Third-Grade Descriptives and Correlations Between Alternative Scoring and Total Rubric Scores.

	Fall 13			Winter 14			Spring 14		
Scoring index	*M*	*SD*	*r*	*M*	*SD*	*r*	*M*	*SD*	*r*
Words written	35.64	14.03	.49***	14.03	35.64	.35*	39.26	15.10	.58***
Words spelled correctly	31.14	13.47	.53***	13.47	31.14	.36*	34.48	14.71	.64***
Correct word sequences	22.83	12.11	.65***	12.11	22.83	.54***	25.98	13.41	.76***
Incorrect word sequences	16.00	9.77	.01	9.77	16.00	−.09	16.90	8.60	−.10
Correct minus incorrect word sequences	6.83	15.85	.50***	15.85	6.83	.47**	9.17	15.83	.70***
Total punctuation marks	2.07	2.35	.49***	2.35	2.07	.49***	1.88	2.44	.50***
Correct punctuation marks in context	1.10	2.12	.48***	2.12	1.10	.43**	1.26	2.37	.44**
Incorrect punctuation marks in context	0.93	1.42	.24	1.42	0.93	.16	0.64	0.76	.25
Correct minus incorrect punctuation marks in context	0.14	2.75	.30	2.75	0.14	.25	0.59	2.44	.36*
Correct punctuation marks without context	1.64	2.26	.46**	2.26	1.64	.52***	1.64	2.35	.50***
Incorrect punctuation marks without context	0.40	0.67	.31*	0.67	0.40	−.02	0.26	0.50	.11
Correct minus incorrect punctuation marks without context	1.24	2.37	.37*	2.37	1.24	.50***	1.38	2.23	.51***
Complete sentences	0.52	1.02	.42**	1.02	0.52	.47**	0.76	1.48	.43**
Words in complete sentences	5.45	10.43	.44**	10.43	5.45	.43**	7.26	13.05	.47***
Number of error words	30.19	16.91	.14	16.91	30.19	.03	32.00	17.63	.15
Number of incorrect words	7.45	5.17	−.10	5.17	7.45	−.18	7.67	4.34	−.19

Percent of correct punctuation marks in context	26.53	40.04	.48***	40.04	26.53	.42**	30.97	42.13	.46**
Decimal percent of correct punctuation marks in context	0.27	0.40	.48***	0.40	0.27	.42**	0.31	0.42	.46**
Percent of correct punctuation marks without context	53.14	45.79	.31*	45.79	53.14	.48***	52.59	47.64	.49***
Decimal percent of correct punctuation marks without context	0.53	0.46	.31*	0.46	0.53	.48***	0.53	0.48	.49***
Mean length of correct word sequences	4.74	4.26	.51***	4.26	4.74	.25	4.46	2.37	.63***
Structure	1.93	0.56	.42**	0.56	1.93	.70***	1.86	0.61	.75***
Spelling	2.10	0.37	.11	0.37	2.10	.11	2.14	0.35	.27
Mechanics	1.17	0.85	.45**	0.85	1.17	.37*	1.00	0.91	.63***
Grammar	2.12	0.59	.47**	0.59	2.12	.54***	2.12	0.55	.55***
Prompt	2.45	0.74	.84***	0.74	2.45	.77***	2.57	0.59	.85***
Ideas	2.38	0.80	.86***	0.80	2.38	.88***	2.40	0.83	.88***
Organization	2.33	0.72	.87***	0.72	2.33	.86***	2.26	0.70	.89***
Voice	2.12	0.80	.87***	0.80	2.12	.92***	2.14	0.93	.90***
Development	2.05	0.99	.84***	0.99	2.05	.81***	2.07	0.95	.87***
Total rubric	18.38	4.17	1.00	18.64	4.49	1.00	18.55	4.96	1.00

*Correlation is significant at the .05 level (two-tailed). **Correlation is significant at the .01 level (two-tailed). ***Correlation is significant at the .0016 level (two-tailed, with Bonferroni correction).

Internal consistency reliability.

To measure whether the dimensions of the rubric assess a single construct of early writing quality, Cronbach's alpha was calculated for the rubric dimensions at each time point. Cronbach's alphas at fall, winter, and spring were .84, .85, and .90, respectively; these values met the accepted criterion level of α = .80 (National Center on Response to Intervention, 2010).

Validity

Criterion-related validity.

Pearson product–moment correlations were calculated across the alternative scoring measures and the trait-based rubric (see Table 3). To control for the number of correlations run (31) and to help diminish Type I error, a Bonferroni correction was used (.05/31 = .0016). Moderate positive correlations were found between some of the alternative scoring methods and the trait-based rubric at fall (r = .48–.65), winter (r = .48–.54), and spring (r = .47–.76). Although CWS maintained the strongest correlation coefficients with the trait-based rubric across the three time points, only the correlations between the rubric and CWS were above the general $r > .60$ criterion for acceptable validity at fall, and correlations between WSC, CWS, CIWS, and mean length of correct word sequences (ML_CWS) with the trait-based rubric met this threshold at spring (McMaster et al., 2009). Ten students in this sample also completed the WIAT-3 *Spelling* and *Sentence Composition* subtests from the larger study. No statistically significant ($p \leq .05$) correlations of rubric scores with the WIAT-3 subtests were found.

Construct-related validity.

Pearson-product moment correlations revealed moderate to moderately strong correlations between the text features (except spelling) of the trait-based rubric and the final rubric score (see Table 3). At fall, correlations with structure, spelling, and mechanics were very weak (range r = −.17–.37), whereas prompt was highly correlated with ideas, organization, voice, and development (range r = .71–.86). Similar patterns for prompt with ideas, organization, voice, and development were evident at winter (range r = .65–.78) and spring (range r = .74–.76). Structure exhibited moderate

correlations with all but spelling at winter (range r = .33–.58; spelling r = −.08) and spring (range r = .34–.64; spelling r = .10).

Social validity.

A convenience sample of third-grade teachers reviewed the trait-based rubric. Teachers indicated that the rubric was "student-friendly" and comprehensive and that they liked the quality indicators on Organization and Relationship to and Completeness of the Prompt. Teachers recommended weighting content over grammar/ mechanics and suggested that the rubric contains a section on analysis and using text-based responses.

Growth.

Paired-sample t tests revealed no statistically significant growth for the total rubric scores: fall–winter: $t(41)$= − .58, p = .57; winter–spring: $t(41)$ = .14, p = 89; fall– spring: $t(41)$ = − .28, p = .78.

Discussion: Study 2

The purpose of the current study was to explore a series of alternative scoring methods and their relationship to a trait- based writing rubric and to examine the extent to which student performance on the trait-based rubric demonstrated evidence of reliability, criterion- and construct-related validity, and sensitivity to growth over the academic year in third grade. To investigate this, a series of correlations and paired-sample t tests were calculated. Results indicated that the writing rubric was moderately correlated with select scoring indices across the time period, but that CWS maintained the strongest correlation across fall, winter, and spring, and that WSC, CIWS, and ML_CWS also demonstrated evidence of acceptable reliability at spring. Although previous literature supports the technical adequacy of WSC and CIWS (Jewell & Malecki, 2005), limited research exists on the use of ML_CWS as a reliable and valid measure of student writing in the early grades (Gansle et al., 2002; Gansle et al., 2006). This study suggests that ML_CWS may also be another reliable and valid measure of student writing in third grade.

The total punctuation marks (TPM) also demonstrated consistent moderate correlations across time points (r = .49, .49, .50, respectively), suggesting that this measure should continue to be explored as a possible method of understanding students' writing. Furthermore, the number of correct punctuation marks (CPM; either within or outside of con- text) also demonstrated consistent moderate correlations. Future research should continue to explore this measure; however, a consistent definition must first be established. Gansle et al. (2002) similarly found promise for the use of CPM. In this study, CPM without context is consistent with the definition provided by Gansle et al. Moreover, Gansle et al. recommend the use of the number of error words (#EW) as a possible scoring method; however, this study finds no support for this measure.

Unfortunately, high variability in IRR on individual student samples existed across the alternative scoring methods and traits on the rubric (0%–100%), suggesting that more research is needed, despite acceptable internal consistency reliability. Even with well-established writing procedures, IRR thresholds are often set lower in writing research, typically around 80% or 85%; therefore, it is possible that promising scoring methods may exhibit evidence of lower total IRR compared with measures in other areas such as reading (McMaster et al., 2009). However, the wide variability in the IRR in this study means that we must interpret the consistency of these exploratory measures with caution. Given the timed nature of the CBM task and that struggling writers tend to produce less text, even disagreements on one word or sequence can immediately influence reliability. In the future, more extensive training may be conducted to help increase the reliability of multiple scorers. Alternatively, the scorers could meet immediately following reliability scoring to come to agreement on scoring procedures. However, because many of these measures are exploratory, it would be of benefit to thoroughly review the extent to which the current research both utilizes, defines, and calculates such measures. It might also be necessary for authors to be more explicit in their definitions so that they can be more easily utilized across research teams.

Although students' performance on the trait-based rubric exhibited moderately strong correlations across time, student growth was not statistically significant, suggesting that the rubric is not sensitive to growth. Indeed, students' total rubric scores

were essentially unchanged across fall, winter, and spring. Strong construct-related validity coefficients among the components of this rubric appear to indicate that multiple text features may be measuring the same construct. Similarly, weak construct-related validity coefficients suggest that other items may be measuring a different construct and may not be appropriate. Given that the SP CBM-W is timed and that students in third grade are at a unique developmental stage in which they are beginning to write to learn, the nine text features may not be appropriate for this type of task. Based on this analysis, spelling appears to be measuring a different feature and should be eliminated as an item on the rubric. This is potentially interesting given that the current literature often supports the use of spelling as an indicator of students' writing performance (Deno et al., 1982; Parker et al., 1991) and that models of early writing similarly posit that spelling and/or transcription skills are necessary for producing connected text (e.g., Berninger & Amtmann, 2003; Juel et al., 1986). Earlier research by Fulk and Stormont-Spurgin (1995) had also suggested that spelling may be a prominent skill by which to discriminate students with learning disabilities from their low-performing peers, which makes exclusion of spelling from the rubric theoretically troublesome if using these measures with students with learning or writing disabilities.

In addition, it is likely that the items prompt, ideas, organization, voice, and development share too much variance. It may be that some of these items need to be eliminated from the rubric or that some (or all) of these items should be combined into a single item. Future research should explore an abbreviated version of this trait-based rubric. Nonetheless, it is worth noting that no student received the maximum score possible on the rubric (27 points) nor did a student receive the minimum score possible on the rubric (0 points), indicating that no ceiling or floor effects exist. Teachers, overall, also indicated that they believed the researcher-developed trait-based rubric demonstrated potential for learning more about their students' writing. Indeed, because student growth often takes time, teachers are interested in ways to better display gradual changes in student writing, especially for learners with the most intensive writing needs, including learners with learning and writing disabilities. To do so, they may turn to rubrics (Gansle et al., 2006) as an alternative and/or supplement to more common production scores like CWS. Thus, though this researcher-developed trait-

based rubric shows limited potential in its current format, this is an exciting avenue for additional research so that teachers can have access to tools that exhibit technical reliability and validity.

Finally, no statistically significant correlations were found between the total trait-based rubric scores and students' performance on the *Spelling* and *Sentence Composition* subtests of the WIAT-3. However, given the small number of students who completed these subtests in the present study, it is possible that there is not enough data to provide a reliable and valid estimation of the relation- ship. A standardized writing task should be administered to all students in future research rather than only using a researcher-developed trait-based rubric as the criterion measure.

Implications for Practice

Writing in the schools is typically scored using some form of holistic rating or trait-based rubric (e.g., 6 + 1 Traits, etc.; Gansle et al., 2006; NWREL, 2000). Today, even state assessments evaluate student writing performance using a rubric. Although these methods can be broadly informative, they are often unreliable in isolation. Aligning objective quantitative components with qualitative components into writing rubrics may improve their consistency and utility in the classroom. The results of these two small exploratory studies suggest that alternative scoring methods used with CBM-W tasks provide a more comprehensive assessment of student writing ability in the early elementary grades compared with traditional methods of writing assessment.

Rubrics

The trait-based rubrics demonstrated questionable interrater and internal consistency reliability, which reflects past research and a common problem of practice (Gansle et al., 2006). The rubrics also demonstrated some evidence of criterion-related validity. The rubrics had stronger criterion- related validity coefficients with CWS compared with other production scores, which is similar to results from past research (Coker & Ritchey, 2010; Tindal & Parker, 1991). However, the rubrics demonstrated evidence of weak to moderate criterion-related validity coefficients with the WIAT-3. Taken as a whole, the technical adequacy of the rubrics from these two studies

indicates that it may be possible to capture important aspects of writing using trait-based rubrics complimentary to CBM-W scoring, but additional research with larger samples is needed, including samples that include a larger number of students with learning and writing disabilities. The validity evidence indicated that the rubrics measured some aspects of writing that are important for transcription and sentence construction in developing writers, but they should be used with caution and not in isolation to make educational decisions.

In terms of growth, the first-grade rubric demonstrated significant growth over time while the third-grade rubric did not. It could be that in first grade, students are still learning how to write and the rubric traits involved mostly transcription abilities, so the scores were better able to capture growth over an academic year. In third grade, students have transitioned from learning to write to writing to learn. It is possible that the components of writing could not be sufficiently demonstrated given that SPs are administered for only 3 min; moreover, it is possible that the limited findings are an artifact of having too small of a sample to score using a rubric. Future iterations of this work may include a modified version of the third-grade rubric to include different writing traits (e.g., one trait that captures components of the current rubric items of prompt, ideas, organization, voice, and development) or measure the existing rubric with different writing tasks (e.g., essay tasks of at least 10 min). However, teachers indicated that the third-grade rubric had social validity and potential for student use, meaning that educators believed the rubric was measuring critical aspects of writing. This could indicate a disconnect between what the research says are important mid-grade writing skills and what educators believe is important for measuring writing skills.

When considering the utility of rubrics in a school set- ting, we must acknowledge the reality of teachers' time. Although we did not measure the amount of time it took to score the CBM-W tasks, we must consider the efficiency of each procedure. The available evidence suggests that CWS or CIWS are the most technically sound scoring methods with CBM-W and should be used when screening and progress monitoring. If further research on rubrics as a complimentary CBM-W scoring procedure shows promise, it would be worth considering whether that method is more time efficient and whether a rubric can supplant, rather than supplement, more traditional scoring methods.

At this point, the technical adequacy of rubrics is not as strong as their production scoring counterparts; therefore, teachers should use the procedures that are the most reliable and valid, not necessarily the fastest.

No student scored the minimum or the maximum scores possible on either rubric, indicating an absence of floor and ceiling effects. The rubrics have potential as screeners for various ability levels. It was still difficult to obtain reliable scores across raters while including enough dimensions of writing to capture the entirety of a student's writing ability. The more traits included in a rubric means the more potential for rater disagreement in scoring. A scoring rubric needs to be robust enough to capture important skills but feasible enough to be reliable and consistent across students and raters. A longer rubric may also require more time from teachers to score. This study did not measure the length of time it took raters to use the rubrics for scoring; future studies should investigate the most efficient use of rubrics and alternative scoring procedures while also providing accurate and useful information for educators.

Alternative Production Scores

Punctuation marks (correct punctuation and total punctuation regardless of accuracy) had moderate correlations over time with other production-dependent scoring procedures. This indicates that use of punctuation taps an important aspect of writing quality and complements information obtained from production-dependent scores. Current scoring procedures for SP probes often do not take punctuation as an independent variable into account. Even though CWS involves an element of punctuation, the sequence score is a composite representation, meaning that even if the punctuation is correct, if the words are spelled incorrectly or are not used grammatically, the sequence is scored as an error. Although students in third grade are still young, the use of accurate punctuation is often emphasized in instruction. This is also a period in which students begin to experiment with different forms of punctuation, especially the use of quotation marks to denote dialogue. It may be that students' ability to use punctuation, whether accurately or to experiment with its use, provides information about students' general writing abilities.

Limitations and Future Research

The third-grade rubric had evidence of low IRR. Although teachers found that it had utility, its lack of consistency across scorers indicates that more work is needed in capturing critical writing skills in third grade. Potential future research could investigate different traits or using the existing rubric with a different CBM-W task to better align the evaluation with task demands. The first-grade rubric was highly correlated with CWS; however, CWS is based entirely on spelling and grammar and two out of the four traits on the rubric were spelling and grammar skills. Although the rubric measures writing constructs that are important for transcription, it may not provide enough detailed or useful additional information beyond the CWS scores for educators. Future research should include adding new traits or domains to the rubric and examining its use with other CBM-W tasks. Moreover, this study included a limited number of students who received special education services (only 3% of the sample in Study 1). Because students with disabilities—and more specifically students with learning disabilities or students with writing disabilities—compose a very restricted range, any coefficients reported potentially overestimate the writing performance of students with learning and writing disabilities. Unfortunately, given the small sample sizes of the studies reported here, it was not possible to create a subsample of low-performing students and conduct parallel analyses for this group. Replication of this study with a more targeted group of writers with disabilities is needed.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

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

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. (2013). Universal screening for writing risk in kindergarten. *Assessment for Effective Intervention*, *39*, 1–12. doi:10.1177/1534508413502389

Crusan, D. (2015). Dance, ten; looks, three: Why rubrics matter. *Assessing Writing*, *26*, 1–4. doi:10.1016/j.asw.2015.08.002 Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, *37*, 184–192. doi:10.1177/00224669030370030801

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: The 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: The University of Minnesota Institute for Research on Learning Disabilities.

Fulk, B. M., & Stormont-Spurgin, M. (1995). Spelling interventions for students with disabilities: A review. *The Journal of Special Education*, *28*, 488–513. doi:10.1177/002246699502800407

Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Moving beyond total words written: 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., & Harris, K. R. (2012). *Writing better: Effective strategies for teaching students with learning difficulties*. Baltimore, MD: Paul H. Brookes.

Hampton, D. D., & Lembke, E. S. (2016). Examining the technical adequacy of progress monitoring using early writing curriculum-based measures. *Reading & Writing Quarterly*, *32*, 336–352. doi:10.1080/10573569.2014.973984

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1–27). Mahwah, NJ: Lawrence Erlbaum.

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

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

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*, 243–255. doi:10.1037/0022-0663.78.4.243

Kim, Y.-S., Al Otaiba, S., Puranik, C., Folsom, J. S., Greulich, L., & Wagner, R. K. (2011). Componential skills of beginning writing: An exploratory study. *Learning and Individual Differences*, *21*, 517–525. doi:10.1016/j.lin- dif.2011.06.004

Lembke, E., Deno, S. L., & Hall, K. (2003). Identifying an indicator of growth in early writing proficiency for elementary school students. *Assessment for Effective Intervention*, *28*, 23–35. doi:10.1177/073724770302800304

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. L., Du, X., & Petursdottir, A.-L. (2009). Technical features of

curriculum-based measures for beginning writers. *Journal of Learning Disabilities*, *42*, 41–60. doi:10.1177/0022219408326212

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

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

National Center on Response to Intervention. (2010). *Users guide to universal screening tools chart*. Washington, DC: National Center on Response to Intervention, Office of Special Education Programs, U.S. Department of Education.

Northwest Regional Educational Laboratory. (2000). *6+1 Trait® writing: A model that works*. Greensboro, NC: Carson- Dellosa.

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

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

Tindal, G., & Parker, R. (1991). Identifying measures for evaluating written expression. *Learning Disabilities Research & Practice*, *6*, 211–218. doi:10.1111/ldrp.12030

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2003). *The Nation's Report Card: Writing 2002* (NCES 2003–529). Washington, DC: Author.

Videen, J., Deno, S., & Marston, D. (1982). *Correct word sequences: A valid indicator of proficiency in written expression* (No. IRLDRR-84). Minneapolis: The University of Minnesota Institute for Research on Learning Disabilities.

Wagner, R., Puranik, C., Foorman, B., Foster, E., Wilson, L., Tschinkel, E., . . .Kantor, P. (2011). Modeling the development of written language. *Reading and Writing*, *24*, 203–220. doi:10.1007/s11145-010-9266-7

Weschler, D. (2009). *Weschler Individual Achievement Test* (3rd ed.). Upper Saddle

River, NJ: Pearson.