

New and Existing Curriculum-Based Writing Measures: Technical Features Within and Across Grades

Kristen L. McMaster
University of Minnesota

Heather Campbell
St. Olaf College

Abstract. The purpose of this study was to examine technical features of new and existing curriculum-based measures of written expression in terms of writing task, duration, and scoring procedures. Twenty-five third-, 43 fifth-, and 55 seventh-graders completed passage-copying tasks in 1.5 min and picture, narrative, and expository writing prompts in 3–7 min. Samples were scored quantitatively. Measures that yielded sufficient alternate-form reliability were examined to determine which had sufficient criterion validity, and those with sufficient criterion validity were examined to determine which detected growth from fall to spring. Different types of tasks yielded varying levels of technical adequacy at each grade, with longer durations having stronger technical adequacy for older students and more complex scoring procedures having stronger technical adequacy for all students. Narrative writing appeared most promising in terms of its technical adequacy across grades. Implications for monitoring progress within and across grades are discussed.

Progress monitoring has long been a hallmark of special education. Individualized Education Program teams use progress monitoring to establish students' present levels of performance, set goals, monitor progress toward those goals, and make instructional changes when progress is insufficient (Deno & Fuchs, 1987). Moreover, progress monitoring is viewed as a way to uphold tenets of the Individuals with Disabilities Education Act (2004) by aligning individual goals and objectives with performance and progress in the

This research was supported in part by Grant H324H030003 awarded to the Institute on Community Integration and the Department of Educational Psychology, College of Education and Human Development, at the University of Minnesota, by the Office of Special Education Programs in the U.S. Department of Education. The paper does not necessarily reflect the position or policy of the funding agency, and no official endorsement should be inferred. The authors thank Professors Stanley L. Deno and Susan Rose for their input to the development of the measures in this study; Cortney Olson, Leah Tanke, Karin Bergstrom, and Lauren Barkmeier for the many hours of scoring writing samples; and the Minneapolis Public School teachers and students who participated in this research.

Correspondence regarding this article should be addressed to Kristen McMaster, University of Minnesota, 250 Education Sciences Building, 56 East River Road, Minneapolis, MN 55455; E-mail: mcmas004@umn.edu

Copyright 2008 by the National Association of School Psychologists, ISSN 0279-6015, who has nonexclusive ownership in compliance with Division G, Title II, Section 218 of P.L. 110-161 and NIH Public Access Policy.

general education curriculum (e.g., Nolet & McLaughlin, 2000). Progress monitoring has also gained increased attention of education policy makers and administrators. Current policies that emphasize standards and accountability (No Child Left Behind Act, 2002) and response to intervention (Individuals with Disabilities Education Act) have illuminated the need for assessment tools that can be used to track student progress and to quickly and accurately identify those at risk of failing to meet critical academic standards.

Educators also have recently focused their attention on assessing and developing students' writing skills. This attention is, in part, in response to reports of high proportions of students who do not meet proficiency levels in writing. In 2002, 72% of 4th-graders, 69% of 8th-graders, and 77% of 12th-graders were performing below a proficient level in writing (National Center for Education Statistics, 2003). Thus, the National Commission on Writing (2003) urged educational policy makers and practitioners to focus on writing in its report, "The Neglected 'R': The Need for a Writing Revolution," and promoted "an integrated system of standards, curriculum, instruction, and assessment" (National Commission on Writing, 2006, p. 19) for ensuring that students achieve excellence in writing.

To document student progress within the curriculum and toward rigorous standards, identify those who are struggling, and inform instruction aimed at improving writing proficiency, technically sound progress monitoring tools are needed. One well-researched progress monitoring approach is curriculum-based measurement (CBM; Deno, 1985). A 30-year program of research has illustrated the capacity of CBM to provide reliable and valid indicators of student performance and progress in core content areas (Marston, 1989; see also Foegen, Jiban, & Deno, 2007; McMaster & Espin, 2007; Wayman, Wallace, Wiley, Ticha, & Espin, 2007) and to effect improvements in student achievement (Stecker, Fuchs, & Fuchs, 2005). Below, we briefly review research on the development of CBM in written expression (CBM-W).

Research at Elementary and Secondary Levels

CBM-W was first developed at the Institute for Research on Learning Disabilities at the University of Minnesota. Institute for Research on Learning Disabilities researchers demonstrated that several simple, countable indices obtained from brief writing samples were valid in relation to standardized writing tests, a developmental scoring system, and holistic ratings (r values = .67 to .88; Deno, Mirkin, & Marston, 1980, 1982). Test-retest, alternate-form, and interscorer reliability and internal consistency coefficients ranged from r values = .50 to .96 (Marston & Deno, 1981; Tindal, Marston, & Deno, 1983). Technically sound indices of writing proficiency for third- to fifth-graders included the number of words written (WW) and words spelled correctly (WSC) when students responded to writing prompts for 3 to 5 min. Correct word sequences (CWS), which incorporates both correct spelling and grammar, also provided a valid index of writing in relation to standardized writing tests and developmental and holistic ratings (Videen, Marston, & Deno, 1982). These measures reliably differentiated among students at different skill levels and were sensitive to growth from fall to spring (e.g., Marston, Deno, & Tindal, 1983).

Several researchers have extended Institute for Research on Learning Disabilities research by examining technical features of existing and new scoring procedures (e.g., correct punctuation, words in complete sentences, simple sentences) in Grades 3–5 (Gansle, Noell, VanDerHeyden, Naquin, & Slider, 2002; Tindal & Parker, 1991). Technical adequacy of writing measures in these later studies was generally weaker than in the Institute for Research on Learning Disabilities studies, with weak to moderate criterion validity (r values = -.02 to .63; Tindal & Parker) and weak to moderate alternate-form reliability (r values = .006 to .62; Gansle et al., 2002; Gansle et al., 2004). These findings raise questions about whether measures studied thus far are appropriate for indexing elementary students' writing.

Extensions of CBM-W to the secondary level indicate that simple indices such as WW and WSC may not be sufficient for assessing older students' writing (Tindal & Parker, 1989; Parker, Tindal, & Hasbrouk, 1991). Researchers have shown that combinations of measures (Espin, Scierka, Skare, & Halverson, 1999) or correct minus incorrect word sequences (CIWS; Espin, Shin, Deno, Skare, Robinson, & Benner, 2000) are more appropriate indices than WW or WSC for older students (e.g., criterion validity coefficients for CIWS ranged from r values = .65 to .75 compared to r values = .34 to .65 for WW and WSC; Espin et al., 2000). Further, researchers have found reliability and validity to be similar for narrative and expository prompts (Espin et al., 2000), and that longer samples (up to 35 min) may increase criterion validity (Espin, De La Paz, Scierka, & Roelofs, 2005).

Results of elementary- and secondary-level research suggest that different scoring indices might be needed at different grades. Weissenburger and Espin (2005) examined this issue directly by administering narrative prompts to 4th-, 8th-, and 10th-graders. Reliability coefficients were stronger for longer writing samples and weaker at higher grades. Criterion validity between CWS and CIWS and a state writing test was moderate at Grades 4 and 8 for CWS and CIWS (r values = .47 to .68), and weak at Grade 10 for CWS and CIWS (r values = .18 to .36).

Extending CBM Research in Written Expression

CBM-W procedures developed thus far have yielded moderate technical adequacy at best. Continued research is needed to develop ways to accurately index students' writing proficiency and to determine which CBM-W measures are most appropriate at which grades. In the present study, our aim was to replicate and extend research designed to examine CBM-W both within and across elementary and secondary levels. Previous researchers who have examined CBM-W across grades (e.g., Weissenburger & Espin, 2005) administered only one type of writing task

(narrative). These researchers have suggested that different scoring procedures might be needed at different grades, but it is not clear whether the type of task or sample duration should also vary with grade level. We intended to add to the literature by examining three primary features of written expression measures administered to students at different grades: type of task, duration of sample, and scoring procedures.

Type of Task

Researchers have suggested that reliability and validity of CBM-W do not vary depending on type of writing prompt (i.e., narrative vs. expository) for elementary (Deno et al., 1980) and middle school (Espin et al., 2000) students. However, no direct comparison has been made of these measures across grades. Given that (a) students are typically assigned to write in narrative formats at the elementary level and in expository formats at the secondary level (e.g., Deschler, Ellis, & Lenz, 1996), and (b) many low-performing writers struggle to adjust their writing to these different approaches (e.g., Miller & Lignugaris-Kraft, 2002), we explored narrative versus expository writing tasks.

In addition, given that writing prompts have yielded modest reliability and validity coefficients compared to other types of CBM measures (e.g., reading), we wondered whether there are *other* ways to obtain simple, efficient measures of written expression that are viable for progress monitoring. Our search for possible new measures was guided by a theoretical model of developmental constraints on writing acquisition proposed by Berninger and colleagues (Berninger, Mizokawa, & Bragg, 1991; Berninger, Yates, Cartwright, Rutberg, Remy, & Abbott, 1992). According to this model, we should consider three levels of constraints to understand how writing skills develop. At the first level, neurodevelopmental skills such as "rapid coding of orthographic information (sensory input), speed of sequential finger movements (motor output), and rapid, automatic production of alphabet letters" (Berninger et al., 1992, p.

259) may constrain a child's capacity to transcribe ideas into writing. At the second level, linguistic processes at the word, sentence, or discourse level may further constrain composition skills. At the third level, higher order cognitive processes involved in planning, translating, and revising may also constrain the writing process.

According to Berninger et al. (1992), the speed and automaticity with which students integrate letter forms in memory with motor production (e.g., handwriting) may influence their attentional capacity for higher level processes involved in writing. If lower order skills involved in transcription have an effect on higher order writing processes, perhaps a measure involving such transcription processes could serve as a general indicator of students' overall writing proficiency, especially for younger students just beginning to develop writing skills.

Indeed, researchers have found that transcription skills such as spelling and handwriting speed and legibility are strongly related to writing composition (e.g., Graham, 1990; Graham, Berninger, Abbott, Abbott, & Whitaker, 1997; Jones & Christenson, 1999). One way to capture these skills is through passage copying. Graham, Berninger, Weintraub, and Schaffer (1998) examined copying tasks and found that they can successfully discriminate across a wide age range (first through ninth grades), at least in terms of handwriting speed and legibility. Given this evidence, we reasoned that passage copying might be robust enough to serve as an indicator of writing proficiency, and retain administration and scoring ease that are characteristic of CBM.

Sample Duration

Most CBM-W research has been based on 3- to 5-min samples; however, Espin et al. (2005) found that, for older students, longer samples yielded stronger reliability and validity coefficients. A goal of our research was to determine whether longer durations are needed to obtain technically sound indices of students' writing in different grades. We examined 3-, 5-, and 7-min samples for picture,

narrative, and expository prompts. We only used 1.5-min samples for passage copying, following procedures used by Berninger (2001).

Scoring Procedures

WW, WSC, and CWS have been the most commonly used scoring procedures for elementary students. More complex scoring procedures, such as CIWS, appear to have stronger reliability and validity for secondary students (Espin et al., 1999; Espin et al., 2000), but little is known about the technical adequacy of CIWS for elementary students. We included all of these scoring procedures at each grade level. Interestingly, percentage measures (e.g., %WSC and %CWS) have, in some cases, yielded stronger reliability and validity coefficients for elementary and secondary students (Jewell & Malecki, 2005; Parker et al., 1991; Tindal & Parker, 1989); however, percentage measures pose problems for progress monitoring. For example, if a student produced 10 WSC out of 20 WW in fall, and 50 WSC out of 100 WW in spring, %WSC would not reflect this growth. Thus, we did not include percentage measures in this study, because our primary goal was the development of progress monitoring tools.

Research Questions

In the present study, specific research questions included the following: Which measures of writing performance (in terms of *task*, *duration*, and *scoring procedure*) (a) yield sufficient alternate-form reliability for students at Grades 3, 5, and 7; (b) yield sufficient criterion validity for students at Grades 3, 5, and 7; and (c) detect writing growth from fall to spring?

Method

Setting and Participants

This study was conducted as part of the Research Institute on Progress Monitoring (www.progressmonitoring.org), which was created to develop a seamless and flexible system of progress monitoring in reading, writing, and math with students across pre-

school through high school, including students with and without disabilities. The present study took place in a school serving kindergartners through eighth-graders in a large urban Midwestern district. A district administrator who was part of the Research Institute on Progress Monitoring team recruited this school. Because we wished to obtain a cross section of middle- to upper-elementary and middle school students, teachers from Grades 3, 5, and 7 were asked to participate. All teachers at each grade who taught language arts agreed to participate. The school served approximately 750 students; 67% were minorities, 68% received free or reduced-cost lunch, 17% received special education (the school served a high proportion of students who were deaf or hard of hearing), and 24% received English language learner services. School demographics were representative of the district of 40,000 students (73% were ethnic minorities, 69% received free or reduced-cost lunch, 14% received special education, and 23% received English language learner services).

Complete fall data were obtained from 25 third-, 43 fifth-, and 55 seventh-grade participants. Of these students, 40%, 30%, and 53% in Grades 3, 5, and 7 (respectively) were male; 68%, 72%, and 58% were from minority backgrounds (primarily African American, but also Hispanic, Asian, or American Indian); 60%, 65%, and 58% received free or reduced-cost lunch; 8%, 14%, and 9% received special education; and 28%, 28% and 13% were English language learners. Spring data were collected from 21 third-, 32 fifth-, and 41 seventh-graders (representing a 16% to 26% attrition rate; the remaining students moved or had excessive absences during spring testing). There were no reliable differences between those students who dropped out of the study and those who remained in terms of sex ($\chi^2 = 2.82, p = .12$), ethnicity ($\chi^2 = 7.80, p = .10$), free and reduced-cost lunch status ($\chi^2 = 4.46, p = .22$), special education status ($\chi^2 = 0.37, p = .47$), or English language learner status ($\chi^2 = 0.13, p = .80$). In addition, we compared these two groups on fall writing samples, and found no reliable main effects of

attrition or interactions between attrition and grade. Together, these findings suggest that attrition did not affect the representativeness of our sample.

CBM-W Measures

Tasks. Students completed two passage-copying tasks and responded to two picture, two narrative, and two expository prompts. Each prompt was printed at the top of a sheet of paper, followed by lines printed on the same sheet. Additional paper was available if needed. Students wrote in pencil and were encouraged to cross out rather than erase mistakes. We did not specify whether students should print or write in cursive. The passage-copying task was structured based on the format of the passage-copying task in the *Process Assessment of the Learner* (Berninger, 2001). Passages for the copying task came from the district's third-grade reading curriculum (Cooper & Pikulski, 1999) to be consistent with original conceptions of "curriculum-based" measures (Deno, 1985; i.e., measures are drawn from the students' curriculum) but also to be extendable across Grades 5 and 7. The passages were written at a 5.8 to 6.0 grade level according to the Flesch-Kincaid formula. Students were instructed to copy a practice sentence (e.g., "The quick brown fox jumped over the lazy dog."), and then to copy the passage exactly as it appeared, including capitalization and punctuation, for 1.5 min (as in Berninger, 2001).

Picture, narrative, and expository prompts were intended to reflect experiences to which most U.S. public school students would be able to relate, and to be simple in terms of vocabulary and sentence structure. Prompts within a set were designed to tap similar background knowledge, so that background knowledge would be unlikely to interact with students' responses to different prompts. At the same time, the prompts were intended to be sufficiently different so that students would not write the same thing in response to each one. Set 1 prompts were designed to tap background knowledge of school-related travel. The picture prompt con-

sisted of students boarding a bus outside of a school. The narrative prompt was, “On my way home from school, a very exciting thing happened....” The expository prompt was, “Write about a trip you would like to take with the students in your class.” Set 2 prompts were intended to tap background knowledge related to games or free time. The picture prompt was of students playing ball outside a school. The narrative prompt was, “One day, we were playing outside the school and....” The expository prompt was “Write about a game that you would like to play.”

Sample duration. For each copying task, students wrote for 1.5 min. On the remaining prompts, students were stopped after 3 min and instructed to make a slash after the last word they wrote, and were then prompted to continue writing until a total of 5 min passed. Seventh-graders also made a slash at the 5-min mark, and continued to write for a total of 7 min.

Scoring procedures. Writing samples were scored using the following procedures:

1. WW: The total number of words written in the passage. A *word* was defined as at least two letters written in sequence, with the exception of single-letter words such as *I* and *a*, which were counted as words (see Deno et al., 1980).
2. WSC: Words spelled correctly in the context of the sentence.
3. CWS (Videen et al., 1982): Any two adjacent, correctly spelled words that are syntactically and semantically correct within the context of the sample.
4. CIWS (Espin et al., 1999): Correct minus incorrect word sequences.

Criterion Measures

Test of Written Language. The Test of Written Language—Third Edition (TOWL-3; Hammill & Larsen, 1996) Spontaneous Writing subtest (Form A) was group administered to all participants. Students were presented with a picture depicting a futuristic scene of astronauts, spaceships, and construction activity; told to think of a story about the picture;

and then asked to write for 15 min. Writing samples were scored using analytic rubrics for Contextual Conventions (capitalization, punctuation, and spelling), Contextual Language (quality of vocabulary, sentence construction, and grammar), and Story Construction (quality of plot, prose, character development, interest, and other compositional elements). Alternate-form reliability for 8- to 13-year-olds is reported as .80 to .85. The average validity coefficient with the Writing Scale of the Comprehensive Scales of Student Abilities (Hammill & Hresko, 1994) was reported as .50.

Minnesota Comprehensive Assessment. The Minnesota Comprehensive Assessment (Minnesota Department of Children, Families, and Learning & NCS Pearson, 2002) is a state test; only fifth-graders were required to take the writing subtest at the time of this study. The writing subtest is untimed and requires students to respond to one of four prompts designed to elicit problem/solution, narrative, descriptive, or clarification writing formats. The student composition is scored on five domains (composing, style, sentence formation, usage/grammar, and mechanics/spelling). Two readers trained by NCS Pearson scored each composition independently using a whole number rubric of 1 to 4 for each domain (a score of 4 indicates that a student demonstrates consistent control of that domain, whereas a score of 1 indicates little or no control). The scores given to each domain by the two raters are summed and weighted based on the importance of that domain (for example, composing is given more weight than mechanics). The sum of the weighted scores for each domain is the score for the composition; the maximum is 88. Reliability and validity data are not provided in the technical manual.

Language arts grade-point average (GPA). In Grade 7 only, students’ end-of-year language arts GPAs were available from district records. GPAs were based on three strands from the English/language arts grade-level expectations (Minneapolis Public Schools, 2005), including reading and litera-

ture (word recognition and fluency, vocabulary, comprehension, and appreciation of literature); writing (informative, expressive, and persuasive writing; elements of composition; spelling, grammar, and usage; using reference materials; and handwriting and word processing); and speaking, listening, and viewing (communicating effectively, using electronic and print media). GPA is reported based on a 4-point scale.

Procedures

CBM administration. CBM-W measures were administered in November 2004 and May 2005. The second author group-administered measures to each third- and fifth-grade class during their media time in the library, and to seventh-grade students during their language arts classes. The measures were administered in three sessions, each 1 week apart. In Sessions 1 and 2, students completed a copying task followed by two prompts. In Session 3, students responded to the remaining two prompts. In November, we administered two probes for each prompt so that alternate-form reliability could be examined. We counterbalanced writing prompts such that students in different classes responded to two picture, two narrative, or two expository prompts, with one from Set 1 and one from Set 2 in counterbalanced order in each session.

Scoring training and agreement. A doctoral student in special education experienced in the development, administration, and scoring of CBM-W was designated as the expert scorer. She met with four other scorers (all graduate students in educational psychology) for a 2-hr session to describe, demonstrate, and practice the scoring procedures. Each scorer then scored a writing packet that included all prompts and copying tasks produced by one student. The expert compared each scorer's results with her own, and calculated the percentage of agreement for each scoring procedure by dividing the smaller score by the larger score and multiplying by 100.

For each scorer, the expert randomly selected 1 of every 10 packets, scored them

independently, and compared the scorer's results with her own. If agreement for each score was not at least 80%, the expert and the scorer met to discuss the discrepancy. If there were only a few discrepancies, the two came to agreement on the correct score. If there were several discrepancies, the entire packet was rescored and the scorer had to reach 80% agreement with the expert again. In only one case did a scorer have to rescore an entire packet. Interscorer agreement ranged from 86% to 98%, with a mean of 93%. WW yielded the highest levels of agreement; CIWS yielded the lowest as this procedure involves more subjectivity in judging grammaticality.

To score the TOWL-3 Spontaneous Writing samples, the first author and a doctoral student in special education met for 1 hr to review and practice scoring procedures. We then scored 10% of the writing samples. The number of agreements was divided by the number of agreements plus disagreements and multiplied by 100 to obtain percentage agreement. All discrepancies were discussed until we agreed on the appropriate score. We scored common samples until at least 85% agreement was obtained on 10% of the samples. We then divided and independently scored the remaining samples. Interscorer agreement across the different scoring procedures ranged from 86% to 99%, with a mean of 93%.

Data Analysis

Fall writing data were analyzed to determine a subset of measures (those with sufficiently reliable scores) to be administered again in spring. Because of the large amount of data collected, we conducted the first set of analyses on all measures, and then narrowed the focus of each subsequent set of analyses to measures deemed sufficient in the previous analyses. Because there is no consensus on criteria by which to judge reliability and validity of measures, sufficient reliability was determined based on (a) the general rule that reliability coefficients of at least $r = .80$ are desirable for group-administered tests (e.g., Salvia & Ysseldyke, 2001), and (b) coefficients found for other types of CBM and other

types of writing measures. In reading, sufficient reliability coefficients have generally been reported as $r > .85$ (Wayman et al., 2007). For standardized writing measures, reliability estimates have ranged from .70 to above .90 (Taylor, 2003). With this information in mind, we considered reliability coefficients of $r = .70$ to be sufficient.

Similarly, to judge the strength of validity coefficients, we relied on (a) the general rule that correlations of $r < .60$ should be interpreted with caution, and (b) that writing measures have historically yielded modest criterion validity coefficients (Taylor, 2003). Because we wished to be as inclusive as possible in identifying promising measures, we considered correlations above .50 to be sufficient for inclusion in further analyses.

We began our analyses by examining distributions of each writing task, scoring procedure, and duration. Measures with relatively normal distributions (i.e., skewness and kurtosis were less than 2.58, which is acceptable for small samples) were examined to determine which had sufficient alternate-form reliability, by calculating Pearson r correlation coefficients between forms. Those with sufficient reliability were then examined to determine which had sufficient criterion validity by calculating Pearson's r or Spearman's rho coefficients with criterion measures. Measures with sufficient reliability *and* criterion validity were examined to identify whether they detected fall to spring growth, using paired samples t tests.

Results

Before addressing the three research questions, we created histograms for each measure administered at each grade. Scores on all measures were normally distributed. Complete descriptive information for each measure at each grade can be obtained from the first author.

Alternate-Form Reliability

Pearson's r correlation coefficients were calculated for each type of writing task, duration, and scoring procedure. A Bonferroni ad-

justment was made to reduce the risk of Type I error. One hundred ninety-two correlation coefficients were computed; thus, a significant p value of $< .001$ was set (.05 was divided by the number of coefficients computed). Fall and spring alternate-form reliability coefficients are listed in Table 1; sufficient coefficients are in boldface.

Passage copying. Alternate-form reliability for all scoring procedures except CIWS were sufficient at Grade 3 (r values = .79 to .95), but not consistently so for Grade 5 or 7.

Picture prompts. Picture prompts yielded sufficient alternate-form reliability coefficients for all scoring procedures on 3- and 5-min samples for Grades 3 and 5 (r values = .74 to .93). More complex scoring procedures (CWS and CIWS) applied to longer samples (7 min) were needed to yield consistently sufficient alternate-form reliability at grade seven (r values = .75 to .81).

Narrative prompts. All scoring procedures applied to 3- and 5-min narrative prompts yielded sufficient alternate-form reliability for Grade 3. Longer samples (5 min) were needed to yield consistently sufficient alternate-form reliability at Grade 5 (r values = .71 to .88). Longer samples (5 to 7 min) and more complex scoring procedures (CWS and CIWS) were needed to yield consistently sufficient alternate-form reliability at Grade 7 (r values = .70 to .85).

Expository prompts. Alternate-form reliability was not consistently sufficient for any of the scoring procedures at Grade 3. At Grade 5, 3- and 5-min prompts yielded sufficient alternate-form reliability for all scoring procedures (r values = .70 to .89). At Grade 7, longer samples (5 to 7 min) and more complex scoring procedures (CWS and CIWS) were needed to yield consistently sufficient alternate-form reliability (r values = .75 to .82).

Criterion Validity

We examined criterion validity of measures using those that had sufficient alternate-

Table 1
Alternate-Form Reliability of Writing Tasks and Scoring Procedures by
Grade Level (Fall and Spring)

Task	WW		WSC		CWS		CIWS	
	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
Passage copying								
Grade 3	.95	.85	.91	.79	.86	.84	.60	.67
Grade 5	.65	.68	.45	.69	.69	.63	.71	.61
Grade 7	.65	.58	.63	.61	.71	.65	.64	.58
Picture (3 min)								
Grade 3	.82	.91	.83	.88	.87	.93	.88	.85
Grade 5	.86	<i>na</i>	.87	<i>na</i>	.86	<i>na</i>	.74	<i>na</i>
Grade 7	.51	.53	.56	.55	.62	.54	.68	.55
Picture (5 min)								
Grade 3	.85	.90	.88	.89	.90	.91	.91	<i>ns</i>
Grade 5	.87	<i>na</i>	.90	<i>na</i>	.88	<i>na</i>	.83	<i>na</i>
Grade 7	.60	.67	.63	.65	.68	.65	.74	<i>ns</i>
Picture (7 min)								
Grade 7	.73	.65	.75	.67	.80	.75	.81	.77
Narrative (3 min)								
Grade 3	.73	.70	.78	.70	.86	.76	.86	.72
Grade 5	.60	.76	.54	.78	.58	.80	.69	.84
Grade 7	.70	<i>ns</i>	.69	<i>ns</i>	.69	.57	.67	.69
Narrative (5 min)								
Grade 3	.79	.74	.86	.73	.90	.78	.89	.80
Grade 5	.74	.76	.71	.84	.71	.86	.77	.88
Grade 7	.78	.64	.79	.67	.80	.70	.78	.77
Narrative (7 min)								
Grade 7	.73	.63	.76	.66	.81	.75	.85	.82
Expository (3 min)								
Grade 3	.71	<i>ns</i>	.62	<i>ns</i>	.67	<i>ns</i>	.70	<i>ns</i>
Grade 5	.70	.74	.75	.77	.74	.88	.63	.84
Grade 7	.55	.74	.57	.74	.62	.79	.69	.82
Expository (5 min)								
Grade 3	.60	<i>ns</i>	.52	<i>ns</i>	.62	<i>ns</i>	.75	<i>ns</i>
Grade 5	.73	.76	.79	.78	.82	.89	.77	.85
Grade 7	.66	.72	.69	.74	.75	.79	.82	.81
Expository (7 min)								
Grade 7	.74	.70	.78	.72	.82	.76	.87	.78

Note. All correlations are significant, $p < .001$, unless indicated as *ns* (nonsignificant) or *na* (not applicable; these correlations were not computed because of the non-normality of the score distributions or because they were not administered in spring). Boldface coefficients met our criterion for "sufficient" reliability. Grade 3 Fall $n = 25$, Spring $n = 21$; Grade 5 Fall $n = 43$, Spring $n = 32$; Grade 7 Fall $n = 55$, Spring $n = 41$. WW = words written; WSC = words spelled correctly; CWS = correct word sequences; CIWS = correct minus incorrect word sequences.

form reliability ($r > .70$). Spring CBM-W scores were correlated with spring TOWL-3 raw scores, the Minnesota Comprehensive Assessment writing subtest (Grade 5 only), and

end-of-year language arts GPA (Grade 7 only). Means and standard deviations for these measures are provided in Table 2. Validity coefficients are provided in Table 3 (measures

Table 2
Descriptive Statistics for Criterion Measures

Measure	Grade 3		Grade 5		Grade 7		Total	
	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
TOWL-3	21.65	(11.73)	24.12	(12.38)	32.98	(11.94)	27.97	(12.94)
MCA writing subtest	—	—	33.88	(29.64)	—	—	—	—
Language arts GPA	—	—	—	—	2.40	(1.42)	—	—

Note. TOWL-3 = Test of Written Language—Third Edition; MCA = Minnesota Comprehensive Assessment (writing subtest); GPA = grade point average.

for which there were no statistically significant correlations are not included). For all analyses, we used $p < .001$ because of the large number of correlations (112) calculated.

Sufficient correlations are displayed as boldface in Table 3. Passage copying yielded sufficient criterion validity with the TOWL-3 for WSC and CWS in Grade 3 only. For picture prompts, CWS and CIWS written in 3 min, and WSC and CWS written in 5 min, yielded sufficient criterion validity with the TOWL-3 for Grade 3. At Grade 7, CIWS written in 7 min in response to picture prompts yielded sufficient criterion validity with the TOWL-3, and both CWS and CIWS written in 7 min yielded sufficient criterion validity with language arts GPA.

With respect to narrative prompts, across criterion measures, CWS and CIWS written in 3, 5, and 7 min yielded sufficient criterion validity at each grade, except for 3-min CWS for Grade 5. Expository prompts were examined only at Grades 5 and 7 (sufficient reliability was not observed in Grade 3). Longer samples (5 to 7 min) and more complex scoring procedures (CWS and CIWS) yielded the most consistently sufficient criterion validity coefficients in these grades.

Growth from Fall to Spring

To examine which indicators of writing performance detected growth from fall to spring, we identified measures that met the criteria mentioned in the previous section for sufficient reliability and validity. Paired-sam-

ples t tests were conducted for each measure at each grade. Means, standard deviations, and average weekly growth rates for each measure that showed statistically significant growth are displayed in Table 4. In Grade 3, none of the scoring procedures on any measures identified as having sufficient reliability and criterion validity for third-graders showed reliable fall to spring growth. In Grade 5, for the 3-min narrative prompt, students gained, on average, .54 CWS per week. For the 5-min narrative prompt, students gained, on average, .72 CWS and .70 CIWS per week. For the 5-min expository prompt, students gained, on average, .91 CWS and .89 CIWS per week. In Grade 7, for the 5-min expository prompt, students gained, on average, .43 CIWS per week.

Complete results are available from the first author.

Discussion

In this study, we examined technical features of measures of written expression within and across Grades 3, 5 and 7. In Table 5, measures that met our criteria for sufficient reliability, validity, and capacity to show fall-to-spring growth are summarized. Measures deemed sufficient in at least one area are in boldface; those deemed sufficient in all three areas are in boldface and italicized. In the next section, we discuss measures that appear most promising as well as implications for further research and practice using CBM-W to monitor progress within and across grades.

Table 3
Criterion Validity Correlations Between Spring CBM and TOWL-3, MCA,
and Language Arts GPA

Task		WW	WSC	CWS	CIWS
Passage copying					
Grade 3	TOWL-3	<i>ns</i>	.55	.66	<i>ns</i>
Grade 7	TOWL-3	<i>ns</i>	<i>ns</i>	<i>ns</i>	<i>ns</i>
Picture (3 min)					
Grade 3	TOWL-3	<i>ns</i>	<i>ns</i>	.63	.62
Grade 7	TOWL-3	<i>ns</i>	<i>ns</i>	<i>ns</i>	<i>ns</i>
	GPA	<i>ns</i>	<i>ns</i>	<i>ns</i>	<i>ns</i>
Picture (5 min)					
Grade 3	TOWL-3	<i>ns</i>	.60	.70	<i>ns</i>
Grade 7	TOWL-3	<i>ns</i>	<i>ns</i>	<i>ns</i>	<i>ns</i>
	GPA	<i>ns</i>	<i>ns</i>	<i>ns</i>	<i>ns</i>
Picture (7 min)					
Grade 7	TOWL-3	<i>ns</i>	<i>ns</i>	<i>ns</i>	.55
	GPA	<i>ns</i>	<i>ns</i>	.57	.67
Narrative (3 min)					
Grade 3	TOWL-3	<i>ns</i>	<i>ns</i>	.63	.70
Grade 5	TOWL-3	<i>ns</i>	<i>ns</i>	<i>ns</i>	.62
	MCA	<i>ns</i>	<i>ns</i>	.54	.62
Grade 7	GPA	<i>ns</i>	<i>ns</i>	.55	.65
Narrative (5 min)					
Grade 3	TOWL-3	<i>ns</i>	<i>ns</i>	.66	.68
Grade 5	TOWL-3	<i>ns</i>	<i>ns</i>	<i>ns</i>	.65
	MCA	<i>ns</i>	<i>ns</i>	.56	.68
Grade 7	GPA	<i>ns</i>	<i>ns</i>	.59	.71
Narrative (7 min)					
Grade 7	GPA	<i>ns</i>	0.47	.57	.72
Expository (3 min)					
Grade 5	MCA	<i>ns</i>	<i>ns</i>	<i>ns</i>	.54
Grade 7	TOWL-3	<i>ns</i>	<i>ns</i>	<i>ns</i>	<i>ns</i>
	GPA	<i>ns</i>	.53	.61	.67
Expository (5 min)					
Grade 5	TOWL-3	<i>ns</i>	<i>ns</i>	<i>ns</i>	.57
	MCA	<i>ns</i>	.45	.55	.60
Grade 7	TOWL-3	<i>ns</i>	<i>ns</i>	<i>ns</i>	<i>ns</i>
	GPA	<i>ns</i>	.48	.59	.66
Expository (7 min)					
Grade 7	TOWL-3	<i>ns</i>	<i>ns</i>	<i>ns</i>	.52
	GPA	<i>ns</i>	.52	.62	.68

Note. All correlations are significant, $p < .001$, unless indicated as *ns* (nonsignificant). Boldface coefficients met our criterion for “sufficient” validity. WW = words written; WSC = words spelled correctly; CWS = correct word sequences; CIWS = correct minus incorrect word sequences; TOWL-3 = Test of Written Language—Third Edition; MCA = Minnesota Comprehensive Assessment (writing subtest); GPA = language arts grade point average.

Table 4
Descriptive Statistics for Measures Showing Statistically Significant
Fall-to-Spring Growth

Measure	Fall		Spring		Mean Difference	<i>t</i> Value	Gain per Week
	Mean	(<i>SD</i>)	Mean	(<i>SD</i>)			
Grade 5							
Narrative 3 min							
CWS	29.67	(15.55)	40.43	(19.73)	10.76	-3.73*	.54
Narrative 5 min							
CWS	46.22	(25.81)	60.59	(31.73)	14.37	-3.55*	.72
CIWS	22.38	(33.00)	36.28	(37.53)	13.91	-3.04*	.70
Expository 5 min							
CWS	46.63	(21.32)	64.88	(29.94)	18.25	-4.54*	.91
CIWS	20.91	(29.27)	38.75	(38.69)	17.84	-5.13*	.89
Grade 7							
Expository 5 min							
CIWS	41.69	(30.99)	50.21	(28.47)	8.51	-6.55*	.43

Note. CWS = correct word sequences; CIWS = correct minus incorrect word sequences.

* $p < .01$.

Promising Measures Within and Across Grades 3, 5, and 7

Passage copying. Passage copying produced sufficiently reliable scores on alternate forms for students in all three grades, but only produced *valid* scores (with respect to the TOWL-3) for Grade 3. These results provide further evidence of the relation between transcription skills and overall writing proficiency, at least for younger students, as described by researchers such as Graham et al. (1998) and Jones and Christenson (1999). This finding is promising because copying tasks are easy to administer and score compared to other CBM-W measures.

Unfortunately, passage copying did not show statistically significant fall to spring growth within Grade 3, raising questions about its utility for progress monitoring. Passage copying may not have shown growth for several reasons. First, whereas it may have been a good initial indicator of writing proficiency, perhaps growth in writing proficiency simply does not translate to growth in copying skills. A second possibility is that the third-

graders in this study simply did not make much progress in writing, and thus no measure would reflect growth. Alternatively, the particular passages that were included in the two copying probes may have been inappropriate for measuring growth. The passages were written at a sixth-grade level (according to Flesch-Kincaid) and may have been too difficult to reflect third-graders' writing growth. It is possible that an easier copying task would be more sensitive to third-graders' growth. There is other preliminary evidence of the utility of copying tasks for beginning writers (e.g., Lembke, Deno, & Hall, 2003). Such tasks should be examined further for young students.

Picture prompts. Generally, our findings suggest that picture prompts are promising, but require further examination. CWS and CIWS produced in 3 min yielded sufficient alternate-form reliability and criterion validity, as did WSC and CWS produced in 5 min. Picture prompts were also administered to seventh-graders in spring, and a handful of scoring procedures showed sufficient criterion va-

Table 5
Measures With Sufficient Alternate-Form Reliability and Criterion Validity, and Reliable Fall-to-Spring Growth

Task/Duration	Scoring	Grade 3			Grade 5			Grade 7		
		Alternate-Form Reliability?	Criterion Validity?	Fall to Spring Growth?	Alternate-Form Reliability?	Criterion Validity?	Fall to Spring Growth?	Alternate-Form Reliability?	Criterion Validity?	Fall to Spring Growth?
Passage copying	WW	Yes	No	—	No	No	—	No	No	—
	WSC	Yes	TOWL-3	No	No	No	—	No	No	—
	CWS	Yes	TOWL-3	No	No	No	—	Yes	No	—
	CIWS	No	—	No	Yes	No	—	No	No	—
Picture, 3 min	WW	Yes	No	—	Yes	—	—	No	No	—
	WSC	Yes	No	—	Yes	—	—	No	No	—
	CWS	Yes	TOWL-3	No	Yes	—	—	Yes	No	—
	CIWS	Yes	TOWL-3	No	Yes	—	—	Yes	No	—
Picture, 5 min	WW	Yes	No	—	Yes	—	—	Yes	No	—
	WSC	Yes	TOWL-3	No	Yes	—	—	Yes	No	—
	CWS	Yes	TOWL-3	No	Yes	—	—	Yes	No	—
	CIWS	Yes	No	—	Yes	—	—	Yes	No	—
Picture, 7 min	WW	—	—	—	—	—	—	Yes	No	—
	WSC	—	—	—	—	—	—	Yes	No	—
	CWS	—	—	—	—	—	—	—	GPA	No
	CIWS	—	—	—	—	—	—	Yes	TOWL-3, GPA	No
Narrative, 3 min	WW	Yes	No	—	Yes	No	—	Yes	No	—
	WSC	Yes	TOWL-3	No	Yes	No	—	Yes	No	—
	CWS	Yes	TOWL-3	No	Yes	<i>MCA</i>	<i>Yes</i>	Yes	GPA	No
	CIWS	Yes	TOWL-3	No	Yes	<i>MCA, TOWL-3</i>	No	No	—	—
Narrative, 5 min	WW	Yes	No	—	Yes	No	—	Yes	No	—
	WSC	Yes	No	—	Yes	No	—	Yes	No	—
	CWS	Yes	TOWL-3	No	Yes	<i>MCA</i>	<i>Yes</i>	Yes	GPA	No
	CIWS	Yes	TOWL-3	No	Yes	<i>MCA, TOWL-3</i>	<i>Yes</i>	Yes	GPA	No
Narrative, 7 min	WW	—	—	—	—	—	—	Yes	No	—
	WSC	—	—	—	—	—	—	Yes	No	—
	CWS	—	—	—	—	—	—	Yes	GPA	No
	CIWS	—	—	—	—	—	—	Yes	GPA	No
Expository, 3 min	WW	Yes	—	—	Yes	No	—	Yes	No	—
	WSC	No	—	—	Yes	No	—	Yes	GPA	No
	CWS	No	—	—	Yes	No	—	Yes	GPA	No
	CIWS	Yes	—	—	Yes	<i>MCA</i>	No	Yes	GPA	No
Expository, 5 min	WW	No	—	—	Yes	No	—	Yes	No	—
	WSC	No	—	—	Yes	No	—	Yes	No	—
	CWS	No	—	—	Yes	<i>MCA</i>	<i>Yes</i>	Yes	GPA	No
	CIWS	Yes	—	—	Yes	<i>MCA, TOWL-3</i>	<i>Yes</i>	Yes	GPA	Yes
Expository, 7 min	WW	—	—	—	—	—	—	Yes	No	—
	WSC	—	—	—	—	—	—	Yes	GPA	No
	CWS	—	—	—	—	—	—	Yes	GPA	No
	CIWS	—	—	—	—	—	—	Yes	TOWL-3, GPA	No

Note. Boldface measures met at least one of our criteria for “sufficient” reliability and criterion validity. Boldface italic measures met all three criteria (reliable, valid, and show growth). WW = words written; WSC = words spelled correctly; CWS = correct word sequences; CIWS = correct minus incorrect word sequences; TOWL-3 = Test of Written Language—Third Edition; MCA = Minnesota Comprehensive Assessment (writing subtest); GPA = Language arts grade point average. Dashes indicate that conclusions were not determined in this study.

lidity. However, like passage copying, this measure did not show growth for third- or seventh-graders. The explanations offered in the previous section for passage copying may apply here: Picture prompts generally did not reflect growth; students’ responses to the specific stimuli in this study did not reflect growth; or students made no real progress in

writing. Further research should explore these possible explanations.

Narrative prompts. CBM-W researchers have frequently examined narrative prompts, and they are commonly administered in practice. However, a recent review of research has indicated weak to moderate reli-

ability and validity of scores on narrative prompts, especially when simple scoring procedures (WW and WSC) are used, and especially at higher grades (McMaster & Espin, 2007). Our findings suggest that more complex scoring procedures might yield more accurate results across grades. CWS and CIWS yielded sufficient reliability and criterion validity for students in Grades 3, 5, and 7 on 5-min samples; WW and WSC did not. These findings are consistent with previous research supporting the use of more complex scoring procedures for elementary- and secondary-level students (Espin et al., 2000; Jewell & Malecki, 2005; Parker et al., 1991; Tindal & Parker, 1989).

Narrative prompts did not reflect fall to spring growth for third or seventh graders, but statistically significant growth was detected for CWS and CIWS on fifth-graders' 5-min samples. Further research should examine whether narrative prompts can be used to monitor fifth-graders' progress on an ongoing and frequent basis (e.g., weekly) and whether such progress data can be used to enhance instructional decision making. Further research is also needed to determine what, if any, conditions improve the capacity of narrative prompts to detect growth.

Expository prompts. Expository prompts did not appear to be sufficiently reliable for third-graders, but several scoring procedures yielded sufficient reliability and validity for fifth- and seventh-graders. CWS and CIWS from 5-min samples reflected reliable fall to spring growth for fifth-graders. CIWS on 5- and 7-min samples reflected reliable growth for seventh-graders. These findings are consistent with those of previous researchers who have found reliability and validity of narrative and expository prompts to be similar for secondary students (e.g., Espin et al., 2000). Findings extend previous work by suggesting that, at least for seventh-graders, expository prompts may be more sensitive to fall-to-spring growth.

Limitations

The following study limitations should be considered. We had a relatively small sam-

ple, especially at Grade 3. The small sample size could have led to Type II error because of a possible restricted range of performance levels of students in the sample, leading to smaller correlation coefficients that led us to reject measures that might have stronger technical adequacy than they appeared to have. On the other hand, we also introduced the risk of Type I error by computing a large number of correlation coefficients; this risk was reduced by using a Bonferroni adjustment. In addition, although our sample appeared to be representative of the school and urban district in which this study was conducted, findings cannot necessarily be generalized to other groups with different demographic characteristics.

With respect to our specific measures, another limitation is that it is not clear whether our findings would generalize to prompts that tap different background knowledge (i.e., different pictures or prompts with different content), or other kinds of expository writing (e.g., analytic or persuasive writing). Our criterion measures also present limitations; particularly, the state standards test and language arts GPA are not ideal given their lack of technical information. However, there are few available writing measures and we wished to be as informative as possible. Given that state standards tests and GPAs are given much weight in educational decisions, we considered them important to include.

A final limitation has to do with our process of reducing measures from fall to spring. We eliminated measures in the fall for two reasons: (a) sufficient reliability was not demonstrated and (b) our time was restricted during spring administration, and so measures were selected that appeared most promising. However, our elimination of some of the measures may have been premature. For example, picture prompts showed sufficient reliability for fifth-graders in fall; had we administered these measures in spring, we would have additional information regarding their technical adequacy within and across grades. Future research should include further examination of these measures.

Implications for Research and Practice: Using CBM-W Across Grades

What we know. Across Grades 3, 5, and 7, our results support the use of 5-min narrative writing prompts, at least for screening purposes. Educators should consider using more complex scoring procedures (CWS and CIWS as opposed to WW or WSC) to ensure reliability and validity. Whereas administration and scoring of writing prompts are more time-consuming than copying tasks, it appears that narrative prompts have the capacity to extend across a wider range of grades. Finally, at least for Grades 5 and 7, CWS and CIWS obtained from expository writing prompts appear promising for monitoring progress across grades. Educators interested in doing so should consider using the same scoring procedures and sample durations across grades to ensure that scores from one grade to the next are comparable.

What we still need to learn. Further research should examine the utility of the measures used in this study to connect progress to grades not included in this study (e.g., third to fourth, fourth to fifth, and so on) as well as to younger writers. More research is needed to examine measures eliminated from the study after the fall administration (picture prompts in Grade 5).

Implications for Research and Practice: Using CBM-W Within Grades

What we know. Within Grade 3, our results support the use of 1.5-min passage-copying tasks scored for WSC or CWS, at least for screening. The advantage to using copying tasks is that they are quick and easy to administer and score. Students' responses to 3-min picture prompts scored for CWS, or 5-min picture prompts scored for CWS or CIWS, seem most viable in terms of reliability and criterion validity. Within Grade 5, CWS and CIWS for 3- and 5-min narrative and expository prompts seem viable in terms of reliability, validity, and capacity to show growth. Within Grade 7, 5- to 7-min expository prompts scored for CIWS seem most

promising for progress monitoring. In other words, within each grade, measures appear to vary in their utility for progress monitoring in terms of *type* of task (picture or narrative in middle-elementary, narrative or expository in late-elementary to middle school) and *duration* (i.e., longer durations at higher grades). Interestingly, CWS and CIWS appeared to be the most appropriate scoring procedures across grade levels. Perhaps these more complex scoring procedures better reflect writing quality (i.e., CWS requires that the writing has meaning whereas WW and WSC do not), thereby increasing the criterion validity of this scoring procedure.

What we still need to learn. Within each grade, future researchers should examine the utility of measures identified as promising for progress monitoring. For example, do the measures show growth over relatively brief periods (e.g., weekly or monthly) such that teachers can use them for instructional decision making? How many data points are needed, and how often must measures be administered, to establish stable growth trajectories? When teachers use CBM-W for instructional decision making, does students' writing performance improve?

In addition, whereas some of the measures did not show promise in this study, we believe it would be premature to eliminate them from the universe of possible CBM-W tools. For example, the narrative prompts did not show substantial fall-to-spring growth for third-graders, but we do not suggest disregarding this measure as a possible progress monitoring tool. Given reliability and validity of narrative prompt scores for third-graders, and its widespread use, further examination of this measure is needed. It is possible that the third-graders in our study simply did not make much growth in writing from fall to spring. Future researchers should examine the sensitivity of CBM-W to growth when strong writing interventions are in place, as well as to changes in progress when changes in instruction are made. Future researchers could also compare growth on CBM-W to growth on other writing measures. Moreover, it is possi-

ble that *other* narrative prompts, or aggregated scores of several narrative prompts, will yield a different picture of growth over time.

Conclusions

In this study, we identified measures that show promise for indexing students' writing proficiency within and across Grades 3, 5, and 7. Whereas our results lead us closer to identifying a set of measures for monitoring students' writing progress within and across grades, we should emphasize (as have others; e.g., Tindal & Hasbrouck, 1991) that writing is a complex and multidimensional process, and that few (if any) measures of written expression are likely to capture all of the critical dimensions of writing. A number of CBM-W measures listed in Table 5 are likely to be sufficient for screening (i.e. those with sufficient reliability and validity) and progress monitoring (i.e. those that show reliable growth), but are not likely to be sufficient for more detailed diagnostic purposes. Thus, educators should carefully consider the purpose of assessing written expression when selecting which measures they will use.

References

- Berninger, V. (2001). *Process assessment of the learner (PAL) test battery for reading and writing*. San Antonio, TX: The Psychological Corporation.
- Berninger, V., Mizokawa, D., & Bragg, R. (1991). Theory-based diagnosis and remediation of writing disabilities. *Journal of School Psychology, 29*, 57–79.
- Berninger, V., Yates, C., Cartwright, A., Rutberg, J., Remy, E., & Abbott, R. (1992). Lower-level developmental skills in beginning writing. *Reading and Writing: An Interdisciplinary Journal, 4*, 257–280.
- Cooper, D., & Pikulski, J. (1999). *Invitations to literacy*. Boston, MA: Houghton Mifflin.
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219–232.
- Deno, S. L., & Fuchs, L. S. (1987). Developing curriculum-based measurement systems for data-based special education problem solving. *Focus on Exceptional Children, 19*, 1–16.
- Deno, S. L., Mirkin, P., & Marston, D. (1980). *Relationships among simple measures of written expression and performance on standardized achievement tests* (Vol. IRLD-RR-22). University of Minnesota, Institute for Research on Learning Disabilities.
- Deno, S. L., Mirkin, P., & Marston, D. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children Special Education and Pediatrics: A New Relationship, 48*, 368–371.
- Deschler, D. D., Ellis, E., & Lenz, B. K. (1996). *Teaching adolescents with learning disabilities* (2nd ed.). Denver, CO: Love Publishing.
- Espin, C. A., De La Paz, S., Scierka, B. J., & Roelofs, L. (2005). The relationship between curriculum-based measures in written expression and quality and completeness of expository writing for middle school students. *Journal of Special Education, 38*, 208–217.
- Espin, C. A., Scierka, B. J., Skare, S., & Halverson, N. (1999). Criterion-related validity of curriculum-based measures in writing for secondary school students. *Reading and Writing Quarterly: Overcoming Learning Difficulties, 15*, 5–27.
- Espin, C., Shin, J., Deno, S. L., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *Journal of Special Education, 34*, 140–153.
- Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *Journal of Special Education, 41*, 121–139.
- 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 alternative measures for curriculum-based measurement in writing. *School Psychology Review, 31*, 477–497.
- Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Slider, N. J., Hoffpauir, L. D., Whitmarsh, E. L., et al. (2004). An examination of the criterion validity and sensitivity to brief intervention of alternate curriculum-based measures of writing skill. *Psychology in the Schools, 41*, 291–300.
- Graham, S. (1990). The role of production factors in learning disabled students' compositions. *Journal of Educational Psychology, 82*, 781–791.
- 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.
- Graham, S., Berninger, V., Weintraub, N., & Schafer, W. (1998). Development of handwriting speed and legibility in grades 1–9. *Journal of Educational Research, 92*, 42–52.
- Hammill, D. D., & Hresko, W. P. (1994). *Comprehensive scales of student abilities*. Austin, TX: PRO-ED.
- Hammill, D. D., & Larsen, S. C. (1996). *Test of Written Language—Third Edition*. Austin, TX: PRO-ED.
- Individuals with Disabilities Education Improvement Act. (2004). P. L. 108–446 U.S.C.
- 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.
- Jones, D., & Christensen, C. A. (1999). Relationship between automaticity in handwriting and students' ability to generate written text. *Journal of Educational Psychology, 91*, 44–49.
- Lembke, E., Deno, S., & Hall, K. (2003). Identifying an indicator of growth in early writing proficiency for elementary school students. *Assessment for Effective Intervention, 28*, 23–35.

- Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18–78). New York: Guilford.
- Marston, D., & Deno, S. (1981). *The reliability of simple, direct measures of written expression* (Vol. IRLD-RR-50). University of Minnesota, Institute for Research on Learning Disabilities.
- Marston, D., Deno, S. L., & Tindal, G. (1983). *A comparison of standardized achievement tests and direct measurement techniques in measuring pupil progress* (Vol. IRLD-RR-126). University of Minnesota, Institute for Research on Learning Disabilities.
- McMaster, K. L., & Espin, C. A. (2007). Technical features of curriculum-based measurement in writing: A literature review. *Journal of Special Education, 41*, 68–84.
- Miller, T. L., & Lignugaris-Kraft, B. (2002). The effects of text structure discrimination training on the writing performance of students with learning disabilities. *Journal of Behavioral Education, 11*, 203–230.
- Minneapolis Public Schools. (2005). *Secondary English/language arts grade level expectations*. Retrieved September 23, 2007, from http://ci.mpls.k12.mn.us/6-12_ELA_State_Standards_Grade_Level_Expectations.html
- Minnesota Department of Children, Families and Learning & NCS Pearson. (2002). *Minnesota comprehensive assessments, Grades 3 & 5, technical manual*. Retrieved January 25, 2005, from <http://education.state.mn.us/>
- National Center for Education Statistics. (2003). *Nation's report card: Writing*. Retrieved April 23, 2006, from <http://nces.ed.gov/nationsreportcard/writing/>
- National Commission on Writing. (2003). *The neglected "R": The need for a writing revolution*. Retrieved March 16, 2007, from <http://www.writingcommission.org/>
- National Commission on Writing. (2006). *Writing and school reform*. Retrieved March 16, 2007, from <http://www.writingcommission.org/>
- No Child Left Behind Act (2002). P.L. 107–110, 115 Stat. 1425, 2002 U.S.C.
- Nolet, V., & McLaughlin, M. J. (2000). *Assessing the general curriculum: Including students with disabilities in standards-based reform*. Thousand Oaks, CA: Corwin Press.
- Parker, R. I., Tindal, G., & Hasbrouck, J. (1991). Countable indices of writing quality: Their suitability for screening-eligibility decisions. *Exceptionality, 2*, 1–17.
- Salvia, J., & Ysseldyke, J. (2001). *Assessment in special and remedial education* (8th ed.). Boston: Houghton Mifflin.
- Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42*, 795–819.
- Taylor, R. L. (2003). *Assessment of exceptional students: Educational and psychological procedures*. (6th ed.). Boston, MA: Allyn & Bacon.
- Tindal, G., & Hasbrouck, J. (1991). Analyzing student writing to develop instructional strategies. *Learning Disabilities Research and Practice, 6*, 237–245.
- Tindal, G., Marston, D., & Deno, S. L. (1983). *The reliability of direct and repeated measurement* (Vol. IRLD-RR-109). University of Minnesota, Institute for Research on Learning Disabilities.
- Tindal, G., & Parker, R. (1989). Assessment of written expression for students in compensatory and special education programs. *Journal of Special Education, 23*, 169–183.
- Tindal, G., & Parker, R. (1991). Identifying measures for evaluating written expression. *Learning Disabilities Research and 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). University of Minnesota, Institute for Research on Learning Disabilities.
- Wayman, M. M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education, 41*, 85–120.
- Weissenburger, J. W., & Espin, C. A. (2005). Curriculum-based measures of writing across grades. *Journal of School Psychology, 43*, 153–169.

Date Received: April 26, 2007

Date Accepted: July 18, 2008

Action Editor: Sandra Chafouleas ■

Kristen L. McMaster, PhD, is an assistant professor of Special Education at the University of Minnesota. Her research interests include promoting teachers' use of data-based decision making and evidence-based instruction, and developing classroom-based and individualized interventions for students who struggle with reading and written expression.

Heather Campbell, PhD, is an assistant professor of Education at St. Olaf College, Northfield, Minnesota. Her interests include expanding educational opportunities for low-income and first-generation students and assisting teachers to develop instructional methods and evaluation tools for English language learners.