Technical Features of Curriculum-Based Measurement in Writing

A Literature Review

Kristen McMaster and **Christine Espin**
University of Minnesota

This article reviews research examining technical features of curriculum-based measurement (CBM) in written expression. Twenty-eight technical reports and published articles are included in this review. Studies examining the development and technical adequacy of measures of written expression are summarized, beginning with research conducted at the Institute for Research on Learning Disabilities at the University of Minnesota and followed by extensions of this work. Differences in technical features of writing tasks, sample durations, and scoring procedures employed within and across elementary and secondary levels are highlighted. Gaps in research addressing the technical adequacy of CBM in written expression are identified, and implications for future research and practice are discussed.

Progress monitoring has of late been on the agenda of educational policy decision makers and administrators. With standards-based reform and school accountability at the forefront of educational policy (e.g., No Child Left Behind Act of 2001), it has become clear that if all students are to meet rigorous academic standards, assessment tools are needed to track student progress toward those standards and to quickly and accurately identify students at risk for failing to reach them. Moreover, some have suggested the use of progress monitoring as part of a nondiscriminatory, response-to-intervention approach for special education referral and identification (see Fuchs & Fuchs, 2006; Speece, Case, & Molloy, 2003). For students receiving special education services, progress monitoring is viewed as a way to uphold major tenets of the Individuals with Disabilities Education Improvement Act (IDEIA, 2004) by aligning goals and objectives on Individualized Education Programs with performance and progress in the general curriculum (Nolet & McLaughlin, 2000).

Recently, educators have focused increasing attention on monitoring students' performance and progress in writing. This increased attention is, in part, in response to reports of high proportions of students who do not meet proficiency levels in writing. For example, results of the National Assessment of Educational Progress 2002 writing assessments indicated that 72% of 4th graders, 69% of 8th graders, and 77% of 12th graders were performing below a proficient level (National Center for Education Statistics, 2003). The emphasis on writing performance is also reflected in states' attempts both to introduce or revise standards that represent the multifaceted, complex nature of the writing process and to implement assessment procedures that sufficiently measure critical elements of this construct (Nolet & McLaughlin, 1997).

Technically sound measures of writing progress are needed to ensure that students are progressing toward writing standards, to identify those who struggle, and to inform instruction aimed at improving students' writing proficiency. One of the most extensively researched progress monitoring approaches is curriculum-based measurement (CBM; Deno, 1985). CBM is a procedure in which multiple probes of equivalent difficulty are administered repeatedly, yielding time-series data that reflect student progress. CBM is simple and efficient: Brief samples of behavior, such as the number of words read correctly in 1 min, correlate strongly with critical academic outcomes, such as reading comprehension. Teachers can use such data to quickly and accurately establish baseline performance, set individual goals, graph student progress, and modify instruction when progress is insufficient (Deno, 1985). A 30-year program of research has demonstrated CBM's capacity to provide reliable and valid indicators of student performance and progress in basic skill areas such as reading and mathematics (see Foegen, Jiban, & Deno, 2006; Marston, 1989; Wayman, Wallace, Wiley, Ticha, & Espin, 2006).

Purpose

The purpose of this article is to review the literature on the technical adequacy of CBM in written expression. In doing

Address: Kristen McMaster, University of Minnesota, Educational Psychology, 233 Burton Hall, 178 Pillsbury Dr. SE, Minneapolis, Minnesota 55455. E-mail: mcmas004@umn.edu

so, we pay particular attention to reliability and validity, which are technical qualities required of any measurement tool to be used for educational decision making.

Reliability

Reliability refers to the precision, accuracy, and consistency of a measurement procedure (Thorndike, 2005). With respect to the development of progress measures such as CBM, reliability is important for two specific reasons. First, because such measures are used to discriminate among groups of students (e.g., to identify students at risk), it is important to know that an individual will maintain his or her standing relative to others across testing occasions, alternate forms, and scorers. Second, because CBM is used to make individual decisions based on progress over time, it is important to know the amount of individual variation that can be expected across repeated measurements.

There is not a consensus on criteria by which to judge the reliability of measures. Thus, in reporting study findings, we can, at best, discuss reliability in relative terms. For example, we can compare coefficients to those found for other types of CBM, as well as to other types of writing measures. In reading—the most well established domain in CBM—reliability coefficients have generally been reported as $r > .85$ (Wayman et al., 2006). For standardized writing measures, alternate-form, and test–retest, reliability estimates have ranged from .70 to above .90 (Taylor, 2003). With this information in mind, we consider reliability coefficients of $r > .80$ to be relatively strong, $r =$ .70 to .80 to be moderately strong, $r = .60$ to .70 to be moderate, and $r < .60$ to be weak.

Validity

Reliability is a necessary, but not sufficient, feature of measures to be used for educational decision making. The *validity* of a measure—how well it measures what it purports to measure—is critical (Thorndike, 2005). The complexity of the writing process poses a particular challenge for establishing validity. Writing involves several major activities, including generating and organizing ideas, translating those ideas into written form, and revising the written product (Hayes & Flower, 1980). These activities require the coordination of a variety of processes, including lexical knowledge and retrieval, phonological and semantic coding, use of syntactic structures (e.g., Berninger, 1994), self-monitoring (McCutchen, 1996), and ortho-motor skills (Jones & Christensen, 1999). Researchers must demonstrate that a brief measure designed for repeated administration can serve as a valid indicator of students' overall writing proficiency, which presumably encompasses all of the above processes.

Criterion validity—how well a measure relates to other measures in the same domain—is often the focus of technical adequacy studies. But criterion validity is only one aspect of validity. In judging the adequacy of a measure to be used for educational decision making, the overall construct validity of a measure should be considered. Messick (1995) provided a useful framework for judging construct validity, stating that construct validity should be viewed as a unified concept comprising (a) content validity (representativeness of the domain being sampled), (b) substantive validity (reflecting the theoretical rationale underlying the measure), (c) structural validity (how well the scoring structure fits with the construct being measured), (d) external (convergent and discriminant) validity, (e) generalizability (how well scores and interpretations generalize across populations, settings, and tasks), and (f) consequential validity (implications for educational decision making).

To demonstrate the first four aspects of validity, writing tasks should (a) represent the multifaceted nature of the writing process (content validity); (b) reflect the variety of cognitive processes that writing theorists have indicated are important (substantive validity); (c) be scored using procedures that are not too narrow or too broad such that relevant information is overlooked or irrelevant information is included (structural validity); and (d) correlate well with comprehensive writing measures that assess multiple writing domains and *not* correlate well with measures of other constructs, such as mathematical problem solving (external validity).

To demonstrate the generalizability and consequential aspects of validity, CBM writing measures should be *seamless* (useful across a variety of students in general and special education and students of different ages) and *flexible* (useful for monitoring progress across a variety of curricula found in states, districts, and schools). A seamless and flexible progress monitoring system allows systematic comparison of growth rates under a variety of instructional conditions and allows the progress of students to be followed from one year to the next, from one setting to the next, and from one curriculum to the next. Because we believe seamlessness and flexibility to be important goals in the development of progress monitoring tools, we pay particular attention to how writing measures function for students of different ages and skill levels, how well writing measures reflect growth, and whether these functions vary by different types of measures.

Method

The search for studies of CBM in written expression was part of a literature search by the Research Institute on Progress Monitoring at the University of Minnesota. Electronic databases including ERIC, Science Citation Index Expanded, PsycInfo, and Expanded Academic Index were searched using the following terms: *curriculum-based measurement* (or *measure*), *general outcome measure,* and *progress monitoring.* This yielded 578 articles and reports. Titles and abstracts were screened to confirm that they related to CBM, and Methods sections were screened to identify empirical studies, yielding 160 articles. Articles were grouped by subject (reading, math, spelling, and writing); 9 of these addressed writing. In addition, 17 technical reports on writing were accessed from the Institute for Research on Learning Disabilities (IRLD) at the University of Minnesota. An ancestral search of identified studies in the initial search yielded 12 additional articles. Studies were included if they reported information regarding reliability and/or any of the six aforementioned aspects of validity (*n* = 28).

Results and Discussion

In this section, we begin by summarizing studies conducted by the IRLD, as these studies provided the foundation for later work. Then, we summarize extensions of this work conducted within and across elementary and secondary levels. Studies reporting validity and reliability correlations are summarized in Table 1 by section (IRLD studies, elementary studies, secondary studies, and studies across grade levels) and then in chronological order. Note that only ranges of coefficients are reported in Table 1, along with all criterion measures. Because of space limitations, we describe in more detail below those correlations that are most useful for understanding the technical adequacy of the CBM measures.

IRLD Studies

Criterion Validity. The first IRLD studies of written expression focused on the criterion validity of a number of different tasks and scoring procedures (Deno, Mirkin, & Marston, 1980; summary published as Deno, Marston, Mirkin, 1982). Writing tasks included story prompts, topic sentences, and picture stimuli to which students responded for 1 to 5 min. Responses to each type of task were scored for the number or length of *T*-units (one main clause plus any attached subordinate clauses; Hunt 1965), large words (words with seven or more letters), mature words (words not commonly used, as measured by the *Standard Frequency Index;* Finn, as cited in Deno et al., 1980), number of words written (WW), words spelled correctly (WSC), and correct letter sequences (CLS; any two adjacent letters that are correct according to the spelling of the word).

Across studies, validity coefficients were strongest for 3- to 5-min samples of writing. Validity coefficients were strongest between the *Test of Written Language* (TOWL; Hammill & Larsen, 1978) raw total score and mature words (*r*s = .76–.88), WW (*r*s = .69–.82), and WSC (*r*s = .71–.88; Deno et al., 1980, Studies 1 and 2) and between the Developmental Scoring System (DSS; Lee & Canter, 1971) and WW (*r*s = .84–.88), WSC (*r*s = .76–.84), and CLS (*r*s = .78–.86). Correlations were similar for each type of prompt (story, picture, topic sentence). From this work, it appeared that the number of letters and words produced in 3- to 5-min samples provided valid indices of writing performance—at least as measured by the TOWL, which assesses multiple dimensions of writing using an analytic rubric, and the DSS, a measure of syntactic maturity that also uses an analytic rubric.

Videen, Deno, and Marston (1982) extended this work by introducing correct word sequences (CWS; any two adjacent, correctly spelled words that are acceptable within the context of the sample). Videen et al. wondered whether students might begin generating words that would not add meaning to their writing but would improve their writing scores if only WW and WSC were used to monitor progress. Thus, they suggested that CWS might better reflect improvement but still maintain ease and efficiency of scoring. Samples from Deno et al. (1980) were selected randomly and scored for CWS. Weak to moderate correlations for CWS were found with the DSS (*r* = .49) and TOWL (*r* = .69). Correlations between CWS and holistic ratings of the samples were relatively strong (*r* = .85). Correlations were weak (*r*s = –.03 to .20) between CWS, mean *T*-units, and Poteet's checklist (cited by Videen et al.), on which samples were rated according to penmanship, spelling, grammar, and ideation.

Reliability. IRLD researchers examined several types of reliability of written expression measures, including test–retest and alternate-form reliability, and internal consistency. Most studies also reported *interscorer reliability*, which was generally strong, with coefficients above .90 for most measures (Deno et al., 1982; Marston & Deno, 1981, Study 4; Marston et al., 1983; Marston, Lowry, Deno, & Mirkin, 1981; Tindal, Marston, & Deno, 1983; Videen et al., 1982).

In terms of *test–retest reliability*, Marston and Deno (1981, Study 1) found that WW and CLS written in 5 min had relatively strong test–retest correlations over a 1-day interval (*r*s = .91 for WW, .81 for WSC, and .92 for CLS) and moderate correlations over a 3-week interval (*r*s = .64 for WW, .62 for WSC, and .70 for CLS). Deno, Marston, et al. (1982) examined what they termed "growth stability" (reliability from fall to spring). Coefficients for first-graders were weak (*r*s = .20–.47). Coefficients for WSC and CLS were moderate to strong for second- through sixth-graders (*r*s = .60–.86), except for WSC in Grade 3 (*r* = .37). Tindal, Germann, and Deno (1983) reported fall to spring coefficients of *r* = .56 for fifth-graders for both WW and CLS.

With respect to *alternate-form reliability*, Marston and Deno (1981, Study 2) found that reliability between two 5-min story prompts was strong for WW (*r* = .95), WSC (*r* = .95), and CLS (*r* = .96). Tindal, Marston, and Deno (1983) obtained reliability coefficients ranging from *r* = .72 for WSC to .93 for CLS. Shinn, Ysseldyke, Deno, and Tindal (1982) obtained weaker coefficients (*r*s = .51–.71 for WW), as did Tindal, Germann, and Deno (1983), who reported reliabilities for fourth- and fifth-graders of *r* = .71 for WW and .70 for number of letters.

Fuchs, Deno, and Marston (1982) aggregated scores across alternate forms in an attempt to reduce error associated

(text continues p. 76)

TABLE 1. Characteristics of Studies Examining Technical Adequacy of Curriculum-Based Measurement in Written Expression

	Sample			Writing measure			Criterion validity		Reliability				
Study	*N*	Grade (s)	Level	Type of prompt	Time (min)	Scoring procedures	Criterion measure	Correlation coefficient	Test–retest	Alternate form	Internal	Growth stability	Inter-scorer
						IRLD Studies							
Deno, Mirkin, & Marston (1980, Study 1; see also Deno, Marston, & Mirkin, 1982)	28	3–6	GE, LD	Story prompt, picture stimulus	5	Mean T-unit Large words Mature words WW WSC	TOWL	.03–.33 .29–.72 .41–.79 .41–.82 .45–.88					
Deno et al. (1980, Study 2)	28	3–6	GE, LD	Story prompt, picture stimulus, topic sentence	1–5	T-unit length Large words Mature words WW WSC	TOWL, SAT	.02–.58 .50–.75 .60–.88 .57–.81 .60–.80					
Deno et al. (1980, Study 3)	82	3–6	GE, LD	Story prompt	1–5	T-unit length Large words Mature words WW WSC CLS	DSS	.29 .23–.35 .54–.74 .65–.88 .67–.84 .64–.86					
Marston & Deno (1981, Study 1)	28	1–6	LD	Story prompts	5	Mature words WW WSC CLS			.50–.57 .64–.91 .62–.81 .70–.92				
Marston & Deno (1981, Study 2)	161	1–6	GE	Story prompts Story prompts Picture stimulus Topic sentence	5 5	WW, WSC, CLS Mature words WW WSC				.95–.96 .74–.79 .79–.85 .81–.87			
Marston & Deno (1981, Study 3)	105	1–6	GE	Story prompts	5	Mature words WW WSC CLS					.74–.98 .87–.99 .70–.97 .87–.99		
Marston & Deno (1981, Study 4)	20	1–6	GE	Story prompts	5	Mature words WW WSC CLS							.90–.94 .98–.99 .98–.99 .98–.99

(Table continues)

(Table 1 continued)

	Sample			Writing measure			Criterion validity		Reliability				
Study	*N*	Grade (s)	Level	Type of prompt	Time (min)	Scoring procedures	Criterion measure	Correlation coefficient	Test–retest	Alternate form	Internal	Growth stability	Inter-scorer
Videen, Deno, & Marston (1982)	50	3–6	GE	Story prompt Topic sentence	5	CWS	DSS TOWL Holistic rating Mean T-unit Poteet Checklist	.49 .69 .85 .18 –.03 to .20					.90
Deno, Marston, Mirkin, Lowry, Sindelar, & Jenkins (1982)	566	1–6	GE	Story prompt	3	WSI WW WSC CLS						.04–.58 .27–.72 .20–.78 .36–.86	.96–.99
Shinn, Ysseldyke, Deno, & Tindal (1982)	71	1–5	LD, LA	Story prompt	3	WW				.51–.71			
Fuchs, Deno, & Marston (1982)	78	3–6	LA	Story prompts	3	WSC				.55–.89			
Marston, Deno, & Tindal (1983)	785	3–6	LA	Story prompt	3	WW, WSC, CLS							.91–.96
Tindal, Germann, & Deno (1983)	60	4–5	GE	Story prompt	3	WW CLS Letters written				.71 .70		.56 .56	
Tindal, Marston, & Deno (1983)	566	1–6	GE	Story prompt	3	WW WSC CLS				.73 .72 .93			.98 .98 .98
						Elementary Studies							
Tindal & Parker (1991)	89 91 80	3 4 5	LD, Ch. 1, LA, GE	Story prompt	3–10	WW WSC CWS	Analytic scoring system	–.02 to .63					.92–.99
Parker, Tindal, & Hasbrouk (1991a, Study 1)	1,917	2–5	GE	Story prompt	6	WW WSC CWS %WSC %CWS	Holistic rating	.36–.49 .43–.64 .58–.61 .48–.67 .43–.70					

(Table continues)

(Table 1 continued)

Study	Sample			Writing measure			Criterion validity		Reliability				
	N	Grade (s)	Level	Type of prompt	Time (min)	Scoring procedures	Criterion measure	Correlation coefficient	Test–retest	Alternate form	Internal	Growth stability	Inter-scorer
Gansle, Noell, VanDerHeyden, Naquin, & Slider (2002)	83 96	3 4	GE	Story prompts	3	(see text)	Teacher ratings ITBS LEAP	−.14 to .37 −.24 to .36 −.12 to .33		.006–.62			
Lembke, Deno, & Hall (2003)	15	2	LA	Word copying Sentence copying Word dictation Sentence dictation	2 3 3 3	WW WW WSC CLS CIWS	Atomistic and holistic	(see text)					
Gansle, Noell, Vanderheyden, Slider, Hoffpauir, Whitmarsh, & Naquin (2004)	22 23 15	3 4	GE	Story prompt	3	WW Punctuation marks Correct punctuation Words in complete sentences CWS Simple sentences	*Woodcock-Johnson–Revised* Writing Samples subtest	.23 .42 .34 .35 .36 −.05					
Secondary Studies													
Tindal & Parker (1989a)	172	6–8	LA, LD	Story prompt	3, 6	WW WSC CWS LW ML/CWS % WSC % CWS % LW	Teachers' holistic ratings	.10 .24 .31 .45 .59 .42 .73 .75					
Tindal & Parker (1989b)	95 97 75	6 8 11	GE	Written retell		WW, holistic rating, # passage-related idea units	Maze Story prompt	.17–.19 .02–.05					

(Table continues)

(Table 1 continued)

	Sample			Writing measure			Criterion validity		Reliability				
Study	*N*	Grade (s)	Level	Type of prompt	Time (min)	Scoring procedures	Criterion measure	Correlation coefficient	Test–retest	Alternate form	Internal	Growth stability	Inter-scorer
Parker, Tindal, & Hasbrouck (1991a)	36	6–8	LD	Story prompt	6	WW LW WSC CWS ML/CWS % LW % CSW	Teachers' holistic ratings, TOWL	.39, .16 .45, .56 .54, .25 .64, .27 .63, .18 .60, .56 .53, .28			.77 .81 .78 .75 .69 .79 .77	.69–.83 .69–.83 .68–.79 .49–.77 .26–.66 .17–.76 .45–.75	
Parker, Tindal, & Hasbrouk (1991b, Study 2)	243	6–11	GE	Story prompt	6	WW WSC CWS % WSC % CWS	Holistic ratings	.39–.41 .43–.52 .48–.56 .34–.46 .36–.42					
Watkinson & Lee (1992)	52	6–8	LD, GE	Story prompt	3	WW LW WSC CWS % LW % WSC % CWS	Intercorrelations with production-dependent measures Intercorrelations with production-independent measures	.78–.99 .79–.92					.80–.99
Espin, Scierka, & Skare (1999)	147	10	LD, LA, GE	Story prompt	3	WW WSC CWS Characters Sentences ML/CWS	English GPA, Teacher ratings, CAT	.05–.22 .08–.41 .18–.52 .16–.48 .30–.63 .20–.40					
Espin, Shin, Deno, Skare, Robinson, & Benner (2000)	112	7–8	GE, LD, EBD	Story & expository prompts	3, 5	WW WSC Characters Sentences CWS CIWS	Teacher ratings and district test	.34–.47 .38–.51 .40–.51 .54–.77 .54–.65 .65–.75		.73–.77 .72–.76 .78–.81 .61–.82 .75–.80 .72–.78			

(Table continues)

(Table 1 continued)

	Sample			Writing measure			Criterion validity		Reliability				
Study	*N*	Grade (s)	Level	Type of prompt	Time (min)	Scoring procedures	Criterion measure	Correlation coefficient	Test–retest	Alternate form	Internal	Growth stability	Inter-scorer
Fewster & MacMillan (2002)	430	6–7	GE	Story prompts	3	WSC	EN8	.34					
	422						EN9	.29					
	212						EN10	.28					
	428						SS8	.24					
	415						SS9	.16					
	214						SS10	.21					
Espin, De La Paz, Scierka, & Roelofs (2005)	22	7–8	LD, GE	Expository prompts	35	WW CWS CIWS	Functional elements Holistic ratings	(see text)					
Espin, Wallace, Campbell, Lembke, Long, & Ticha (2005)	183	10	GE, LD	Story prompt	3	WW	Minnesota Basic Standards Test	.26–.29		.64–.84			
					5	WSC		.26–.29		.64–.82			
					7	CWS		.43–.48		.66–.85			
					10	CIWS		.56–.60		.66–.84			
Studies Across Grade Levels													
Jewell & Malecki (2005)	87	2	GE	Story prompt	3	WW, WSC, CWS, %WW, %WSC, CIWS	THASS	.34–.58					
	59	4					SAT	.34–.67					
	57	6					Language Arts grade	.45–.61					
Weissenburger & Espin (2005)	184	4	GE	Story prompt	3	WW	WKCE	.26–.28		.55–.84			
	137	8			5	CWS	Language	.39–.40		.59–.84			
	152	10			10	CIWS	Arts test	.46–.48		.61–.82			

Note. GE = general education; LD = learning disabilities; LA = low achieving; Ch. 1 = Chapter 1; EBD = emotional/behavioral disorders; WW = words written; WSC = words spelled correctly; CLS = correct letter sequences; CWS = correct word sequences; CIWS = correct minus incorrect word sequences; LW = legible words; ML/CWS = mean length correct word sequences; TOWL = *Test of Written Language* (Hammill & Larsen, 1978); SAT = *Stanford Achievement Test* (Madden, Gardner, Rudman, Karlsen, & Merwin, 1978); DSS = *Developmental Scoring System* (Lee & Canter, 1971); ITBS = *Iowa Test of Basic Skills* (Hoover, Hieronymus, Frisbie, & Dunbar, 1996); LEAP = *Louisiana Educational Assessment Program* (Mitzel & Borden, 2000); CAT = *California Achievement Test* (CTB/McGraw-Hill, 1985); *Woodcock-Johnson–Revised* (WJ-R; Woodcock & Johnson, 1989); *Minnesota Basic Standards Test* (Minnesota Department of Children, Families, and Learning and NCS Pearson, 2002); GPA = grade point average; EN8, EN9, EN10 = English GPA for 8th, 9th, & 10th grades; SS8, SS9, SS10 = social studies GPA for 8th, 9th, & 10th grades; THASS = *Tindal & Hasbrouck Analytic Scoring System* (Tindal & Hasbrouck, 1991); WKCE = *Wisconsin Knowledge and Concepts Examination* (CTB/McGraw-Hill, 1996).

with a particular stimulus or with a particular point in time (e.g., a student might have more to say about one topic than another or might be more tired, hungry, anxious, etc., during one session than another). Participants responded weekly to different story prompts for 10 weeks. Correlations were calculated between scores on adjacent measures (WSC in Week 1 and Week 2), across 4 sessions (mean WSC from Weeks 1 and 3 with mean WSC from Weeks 2 and 4), then across 6, 8, and 10 sessions. Aggregations across more days resulted in stronger reliability (*r* = .55 across 2 days, .72 across 4 days, .85 across 6 days, .88 across 8 days, and .89 across 10 days).

Marston and Deno (1981, Study 3) examined *internal consistency* of mature words, WW, WSC, and CLS written in five 1-min intervals. Split-half reliability coefficients ranged from *r*s = .96 to .99, and Cronbach's alpha ranged from *r*s = .70 to .87. Thus, it appears that students' writing performance remained stable within a sample.

Sensitivity to Growth. In addition to criterion validity and reliability, IRLD researchers examined sensitivity of writing measures to growth, an important indicator of the validity of CBM measures if they are to be used for monitoring progress. One approach to examining sensitivity to growth is to compare scores for students of different ages and skill levels. Deno et al. (1980) found that the number of mature words, WW, WSC, and CLS successfully differentiated among students at different grades, as well as between students with and without learning disabilities (LD). Similarly, Shinn et al. (1982) found that students with low achievement reliably outperformed students with LD on WSC, but that students with LD reliably outperformed students with low achievement in growth from fall to spring. Shinn et al. cautioned that these results were complicated by moderate test–retest reliability (*r*s = .51–.71).

Marston et al. (1981) and Deno et al. (1982) found that WW, WSC, and CLS increased from first to sixth grade and from fall to spring within each grade, although not dramatically, especially at fifth and sixth grades. Similarly, Marston, Deno, and Tindal (1983) found significant growth in WW, WSC, and CLS across the first through sixth grades, as well as within-grade gains across 10 weeks. Growth was not evident on the *Stanford Achievement Test* (SAT; Madden, Gardner, Rudman, Karlsen, & Merwin, 1978) Language subtest over the same 10-week period. Marston et al. (1983) argued that direct measures of written expression were more appropriate than were standardized tests for monitoring progress over short intervals.

Summary and Discussion of IRLD Studies. The IRLD studies laid the groundwork for developing technically sound written expression measures in the following ways. First, across third through sixth grades, moderate to strong criterion validity coefficients were found for countable indices of writing. Coefficients were strongest between mature words, WW, WSC, and both the TOWL and DSS (*r*s = .67–.88). Moreover, coefficients did not differ substantially among different writing tasks (story prompts vs. topic sentences vs. picture stimuli) or for 3- to 5-min samples. Such findings indicated that valid measures of written expression could be obtained with brief writing samples and relatively efficient, objective scoring procedures.

IRLD researchers also examined reliability of measures across first through sixth grades. Reliability was not as strong within grade levels as it was across grade levels (e.g., Tindal, Germann, & Deno, 1983). Reliability coefficients were also lower for students with LD and for students with low achievement, possibly reflecting a range restriction (Marston & Deno, 1981; Shinn et al., 1982). Poor reliability is problematic, especially if the measures are used to identify struggling writers. There is some evidence that aggregating scores across sessions improves reliability (Fuchs et al., 1982), but aggregation is also problematic if measures are to be given on a frequent basis. If six measures are needed to obtain reliable information, it might take weeks or even months to determine whether a student is making progress, limiting timely instructional decisions.

IRLD researchers demonstrated that several scoring procedures are sensitive to growth (Deno et al., 1980; Deno et al., 1982; Marston et al., 1981; Shinn et al., 1982). However, most examinations of growth were conducted cross-sectionally (across grades) or from fall to spring; no one examined the technical adequacy of measures for monitoring progress on a frequent (e.g., weekly) basis. This remains an important area for future development if measures of written expression are to be used for progress monitoring.

Extensions of Research on CBM in Written Expression: Elementary Studies

Several researchers have further examined the technical adequacy of written expression measures for elementary students by examining (a) measures for students at different skill levels, (b) measures to be used for screening, (c) new scoring procedures, and (d) measures for beginning writers.

Students at Different Skill Levels. Tindal and Parker (1991) noted that a better understanding of how writing measures function for students at different skill levels was needed. They examined criterion validity and sensitivity to growth of writing assessments for elementary students of a range of skill levels. We found it interesting that in contrast to IRLD studies, correlations among WW, WSC, CWS, and analytic scores applied to the same writing sample (1 to 5 on story idea, organization-cohesion, and conventions-mechanics) were weak to moderate (*r*s = −.02 to .63).

Statistically significant differences were found between students with LD and general education students on all measures, between Chapter 1 (at-risk) and general education students on most measures, and between students with low performance and general education students on some measures—indicating that the measures effectively differentiated among students of different skill levels. Students also made significant gains in WW, WSC, and CWS from fall to spring. However, the authors noted that some students' writing improved in quantity but not quality, while others' improved in quality but not quantity. They concluded that evaluating writing for educational decision making would likely require a multifaceted approach. Tindal and Hasbrouck (1991) further examined writing samples of students of different ages and skills and suggested that whereas quantitative scoring procedures appeared more useful for monitoring progress, a qualitative scoring approach provided further, complementary information that could be used for making a diagnosis and for tailoring instruction to address specific needs.

Screening. To identify suitable screening measures for identifying struggling writers, Parker, Tindal, and Hasbrouck (1991b, Study 1) scored responses to a story prompt administered in fall and spring quantitatively (WW, WSC, CWS, %WSC, and %CWS) and qualitatively (using a 7-point holistic rating of communicative effectiveness). Correlations between quantitative and qualitative scores were weak to moderate at each grade (see Table 1).

To assess further the utility of quantitative scores for screening, Parker et al. (1991b) examined dispersions in the bottom 30% to 40% of the score distributions and Standard Error of measurement (SEm) bands on percentile line graphs. These analyses indicated that %WSC was suitable for second-graders; %WSC and %CWS were suitable for third-graders; and %WSC, %CWS, and WW were suitable for fourth-graders. The remaining indices were not deemed suitable for screening due to their failure to distinguish among low performers. The authors concluded that across second through fifth grades, %WSC was the most viable screening tool, given its moderate correlation with holistic ratings and suitable distribution in the lower ranges. They emphasized that the lack of sensitivity to student differences of the other scoring approaches could lead to false negatives and recommended caution in their use.

New Scoring Procedures. Gansle, Noell, VanDerHeyden, Naquin, and Slider (2002) cited teachers' dissatisfaction with using WW as a primary index of writing, a practice that was occurring in schools in which they conducted their research. To address this issue, they compared WW to a variety of new scoring procedures, including number of nouns, verbs, and adjectives; long words; WSC; total and correct punctuation; capitalization; complete sentences; CWS; sentence fragments; simple sentences, and computer-scored variables. Interscorer reliability ranged from $r = .70$ (sentence fragments) to .96 (WW). Alternate-form reliability was weak to moderate, ranging from $r = .006$ (long words) to .62 (WW). Similar to Tindal and Parker's (1991) findings, criterion validity coefficients were weak, with none above .40. Correct punctuation and CWS accounted for 34% of the variance with teacher rankings, 33% to 45% of the variance with the *Iowa Test of Basic Skills* (ITBS; Hoover, Hieronymus, Frisbie, & Dunbar, 1996) language subscales, and 16% to 32% of the variance with writing subtests on the *Louisiana Educational Assessment Program* (LEAP; Mitzel & Borden, 2000).

Gansle et al. (2004) then used six "promising" variables—as determined by Gansle et al. (2002): WW, total and correct punctuation, words in complete sentences, CWS, and total simple sentences—to index students' writing improvement following a brief intervention. Participants responded to one of two counterbalanced story prompts, received 25 min of instruction on the writing process, and then responded to the second prompt. Interscorer agreement was above .90 for all scoring indices except simple sentences ($r = .78$). Validity coefficients with the *Woodcock Johnson–Revised* (WJ-R; Woodcock & Johnson, 1989) Written Samples subtest were weak ($r = -.05$ to $r = .42$). Only WW improved following instruction. In general, findings of these studies lend little support to the technical adequacy of either new or existing scoring procedures.

Measures for Beginning Writers. Lembke, Deno, and Hall (2003) examined the technical adequacy of measures for beginning writing. Lembke et al. examined some new types of writing tasks, including word and sentence copying and dictation tasks. Scores on these tasks were correlated with two types of criterion variables: "atomistic" (discrete, countable indices, including average WW, WSC, CWS, and correct minus incorrect word sequences [CIWS] obtained from a writing sample) and "holistic" (teachers' global ratings of the same writing samples).

Correlations between word copying scores and the atomistic variables were weak to moderate (*r*s = .10–.69). WSC and CLS obtained from word dictation appeared to be more strongly related to the atomistic variables (*r*s = .82–.92 for average WW and WSC; *r*s = .52–.92 for average CWS and CIWS). Moderate correlations were found between WSC and CWS on sentence copying and average WW and WSC (*r*s = .74–.79). Scores on sentence dictation had a wide range of correlations with atomistic variables (*r*s = .39–.92). Weaker correlations were associated with average CIWS, and stronger correlations were associated with average WW and WSC. Most correlations with holistic ratings were weak to moderate (*r*s = .06–.67) with the exception of WSC on word dictation ($r = .83$) and CIWS on sentence dictation ($r = .84$).

Summary and Discussion of Elementary Studies. Several researchers have extended IRLD research by examining technical features of both existing and new scoring procedures. Whereas most researchers reported interscorer reliability, few examined test–retest or alternate-form reliability. The exception was Gansle et al. (2002), who reported relatively weak alternate-form reliability coefficients for a variety of new scoring procedures. The lack of reliability data is problematic, as reliability is a necessary precondition for validity (Thorndike, 2005).

With respect to criterion validity, the results of these studies were substantially less positive than were the results of the IRLD studies. There are several possible reasons for this. First, weak criterion validity might be a function of weak reliability; as mentioned, this information was unavailable for most studies. Weaker validity coefficients may also reflect differences in study samples. The IRLD studies often included multigrade samples and often reported correlations across grades, whereas more recent studies reported correlations within grades. The use of multigrade samples may have served to increase the range of scores and thus increase reliability and validity coefficients (indeed, when IRLD studies included within-grade analyses, correlations were weaker). Continued work is needed to develop measures of written expression that have technical adequacy within, as well as across, elementary grades.

Another difference between earlier and later studies is the measures used as criterion variables. The strongest validity coefficients obtained in the IRLD studies were with the TOWL and DSS, which included analytic scoring of a variety of writing domains. In later studies, less direct measures of written expression were used as criterion measures, such as the language subtests of the ITBS (Gansle et al., 2002). Likewise, holistic ratings used in different studies varied in both the range of possible scores, criteria for different ratings, and who completed the ratings (e.g., teachers vs. researchers). A further complication with holistic ratings is that in some studies (Parker et al., 1991b; Tindal & Parker, 1991), they were applied to the same writing samples that were scored using CBM procedures. Holistic ratings may not be a valid approach for evaluating the quality of such brief samples. Applying different scoring procedures to the same writing samples could also inflate correlations, as compared to using completely separate criterion measures. Of course, these reasons for discrepant findings between recent research and IRLD studies are largely speculative; what is clear is that continued research is needed to determine the best ways to index elementary students' writing performance.

In addition to raising questions about the reliability and criterion validity of CBM writing measures, researchers extended the IRLD work in other ways. Tindal and Parker (1991) explored differences between quantitative and qualitative scoring procedures and suggested that using both approaches provides the most useful data for making educational decisions, thus addressing the *structural* aspect of validity (Messick, 1995). Tindal and Parker and Parker et al. (1991b) also found that while measures reliably distinguished among students at different grades or who were served in different educational programs, they were less effective in identifying students at risk for writing difficulties, which presents limitations for screening. Gansle et al. (2004) demonstrated that WW improved following a brief writing intervention. These studies provide some insight into *generalizability* (i.e., how well measures work across different populations) and *consequential* validity (i.e., utility of measures for educational decision making; Messick, 1995).

Secondary Studies

Researchers have also extended the study of CBM to address writing for middle and high school students. These researchers have examined (a) measures integrating reading and writing skills, (b) measures for students requiring remedial and special education, (c) measures for screening and monitoring progress, (d) more complex scoring procedures, (e) different writing tasks and durations, and (f) validity of measures for predicting performance on school-based indicators.

Measures Integrating Reading and Writing Skills. Tindal and Parker (1989b) proposed that measures requiring integration of basic skills with recall of content area information might provide more functional information for secondary-level teachers than do traditional CBM measures that treat "basic skills in isolation" (p. 329). Thus, they examined how well a written retell would relate to other reading and writing tasks. Here, we focus on findings related to the writing tasks.

Middle and high school students' creative writing samples were rated holistically based on communicative effectiveness (1 = *very poor;* 5 = *very effective*). Written retells of a grade-level reading passage were scored for WW, communicative effectiveness, and the number of passage-related idea units. On the written retell task, most students retold only 10% of the main ideas from the reading passages, and scores dropped steadily from one grade to the next, rather than improving as might be expected. Further, scores on the written retells and the writing samples were not significantly correlated. The authors concluded that written retell skills are different from creative writing skills and strongly encouraged further research exploring other approaches to monitoring secondary students' progress.

Measures for Students Requiring Remedial and Special Education. Tindal and Parker (1989a) found reliable differences between students requiring remedial and special education on teachers' holistic ratings, %WSC, %CWS, and mean length of correct word sequences (ML/CWS). Watkinson and Lee (1992) produced similar results: Middle school students with and without LD differed significantly on CWS, incorrect word sequences, %WSC, and %CWS.

Tindal and Parker (1989a) analyzed intercorrelations among scoring procedures and defined two clusters, which they identified as "production dependent" (fluency measures that relied on length; WW, WSC, and CWS), and "production independent" (percentage measures that relied on accuracy; %WSC, %CWS, and percent legible words [%LW]). Percentage measures had moderately strong correlation coefficients with holistic ratings (*r*s = .73–.75 for %CWS and %LW), whereas fluency variables produced weaker coefficients (*r*s = .10–.59). Tindal and Parker concluded that percentage measures were more predictive of holistic ratings of the writing of students with low performance than were fluency measures, but cautioned that percentage measures do not have equal interval scales and are thus difficult to interpret when trying to distinguish among students at different skill levels. Moreover, they are problematic for monitoring progress (e.g., 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 any growth, possibly masking important progress).

Measures for Screening and Monitoring Progress. Parker et al. (1991b, Study 2) further examined fluency and percentage indices to identify suitable screening measures for 6th- through 11th-graders. Criterion validity with a 7-point holistic rating scale (applied to the same writing sample) was weak to moderate within each grade for fluency measures (*r*s = .39–.56) and for percentage measures (*r*s = .34–.46). Students' scores increased from fall to spring on all measures. CWS discriminated better among 8th- and 11th-graders and students at middle- and high-score ranges than they did among students at lower grades and score ranges. Percentage CWS appeared most sensitive for discriminating among low scorers, but lacked precision. The authors cautioned that lack of precision is problematic as it increases the likelihood of identifying false negatives.

Parker, Tindal, and Hasbrouck (1991a) examined writing samples obtained from struggling middle school writers across 6 months. Interscorer reliabilities for all scoring procedures were strong (*r*s = .83–.98). Split-half reliabilities, calculated by correlating scores from the first and last 3 min of each sample, were moderate to strong (*r*s = .69–.81). Only WW increased regularly over the 6 months for students in each grade. There were no reliable differences across grades for any of the writing indices, so data were aggregated to examine criterion validity. Correlations between holistic ratings and WSC, %WSC, %LW, CWS, and ML/CWS were weak to moderate (*r*s = .43–.76). The remaining indices were not sufficient predictors of holistic ratings. Similar patterns were found between the writing indices and the TOWL, although coefficients were generally weaker (*r*s = .15–.56).

Parker et al. (1991a) also conducted both "static" and "dynamic" comparisons of different scoring procedures. Static comparisons involved correlating scores obtained at two time points to determine growth stability. The most consistently strong correlations were obtained for WW, LW, and WSC, although these varied (*r*s = .68–.83). Dynamic comparisons involved creating profiles for each measure by plotting standardized mean scores with confidence bands. The authors again expressed skepticism regarding the measures; whereas some appeared promising in terms of validity, stability, or sensitivity to growth, none was adequate in all of these areas. CWS, ML/CWS, and %LW appeared promising for screening, but not for monitoring progress. Parker et al. suggested that further research was needed to examine whether greater standardization of writing topics, greater structuring of writing tasks, or combining scores from more than one writing sample would provide more stable writing indices.

More Complex Scoring Procedures. Espin, Scierka, Skare, and Halverson (1999) explored the utility of combining measures and using computerized scoring. They examined the writing of four skill groups: students with LD, students in basic skills English classes, students in regular English classes, and students in enriched English classes. Writing samples were typed into a word-processing program, and the grammar-check function was used to obtain WW, WSC, characters written, characters per word, and sentences written. CWS and ML/CWS were counted manually. Reliability of the measures was not examined in this study. Validity coefficients were weak to moderate for each scoring procedure (see Table 1).

The four skill groups differed significantly on characters per word, sentences, and ML/CWS, with students who had LD performing the lowest, followed by students in the basic, regular, and enriched English classes. A multiple regression revealed that a combination of characters per word, sentences, and ML/CWS provided a better index of writing than did any single variable. Espin et al. (1999) concluded that using a combination of scores might be necessary to assess secondary students' writing proficiency. Like Parker et al. (1991a), Espin et al. suggested that using more than one sample might yield stronger criterion validity. At the same time, they emphasized the need for continued research to identify the most valid and reliable indicators of secondary students' writing that would preserve the simplicity and efficiency of CBM.

Type of Writing Task and Sample Duration. Espin, Shin, Deno, Skare, Robinson, and Benner (2000) addressed two issues that had not yet been explored in secondary-level writing research. First, because much of secondary-level writing is expository, they wondered whether expository samples (rather than narrative samples) would better reflect students' writing proficiency. Second, they examined whether duration mattered. Participants produced two narrative and two expository samples in 5 min each, marking their places at 3 min. Alternate-form reliability coefficients for incorrect words, ML/CWS, and characters per word were consistently weak (*r*s < .60). For the remaining measures, coefficients were moderate to strong, especially for WW, WSC, CWS, CIWS, and number of characters (all *r*s > .72). A multiple regression indicated that CIWS was the only reliable predictor of holistic ratings. There were no substantial differences in technical adequacy between type (narrative vs. expository) or duration (3 min vs. 5 min) of samples.

Espin, De La Paz, Scierka, and Roelofs (2005) further extended research at the secondary level by examining the relation of CWS and CIWS to the number of functional elements and quality ratings of text. They also examined whether the length of text affected reliability and validity and whether CWS and CIWS would be sensitive to growth. The researchers randomly selected pre- and posttest essays that students had written for a writing intervention (De La Paz, 1999). In the original study, students had 35 min to write their essays; thus, scores were not based on brief writing samples, as they had been in previous research. Correlations between WW, CWS, and CIWS and criterion measures were moderate to strong (*r*s = .58–.90). Espin et al. also calculated validity coefficients for the first 50 words of each writing sample, which resulted in a notable decrease in the size of correlations (*r*s = .33–.59). It should be noted that because the CBM scoring procedures and the criterion measures were applied to the same writing samples, these correlations may be somewhat inflated.

Espin et al. (2005) reported that WW, CWS, and CIWS written in 35 min increased reliably over time. When only the first 50 words of each sample were examined, increases in CWS and CIWS from pre- to posttest were observed for lower performers; however, longer samples were needed to detect growth in higher performers. Espin et al. concluded that CWS and CIWS obtained from expository essays were promising indicators of secondary students' writing proficiency. Further support for the measures was provided in that they detected growth when a systematic writing intervention was in place (De La Paz, 1999). However, strong correlations and sensitivity to growth were associated with samples written in 35 min, which is much longer than a typical CBM writing sample. Espin et al. again emphasized the need for research to identify sufficient durations while maintaining ease and efficiency of measurement.

Predicting Performance on School-Based Indicators. Fewster and MacMillan (2002) investigated whether middle school students' written expression performance was predictive of high school performance. District writing CBM data were collected from participants' sixth- or seventh-grade records. Participants were then divided into four groups based on their high school educational placement: special education, remedial, general education, or honors classes. English and social studies grade point averages (GPAs) were collected from students' 8th-, 9th-, or 10th-grade records. Correlations between middle school writing and high school GPAs were relatively weak (*r*s = .16–.34). This finding might have been due to differences in educational programming across grades, differences in grading standards among teachers, or differences in length of time between CBM administration and GPAs awarded for different students. A discriminant analysis indicated that CBM scores reliably distinguished among students in special education, remedial classes, general education, and honors classes, suggesting that the measures had utility for screening.

Espin, Wallace, Campbell, Lembke, Long, and Tichá (2006) examined the use of writing measures for gauging individual performance and progress toward state standards. Alternate-form reliability coefficients for WW, WSC, CWS, and CIWS produced in 3, 5, 7, and 10 min were moderate to strong (*r*s = .64–.85) and appeared to strengthen with duration, with coefficients above *r* = .70 for 5-min samples and above *r* = .80 for 7- and 10-min samples. Criterion validity with a holistically scored state writing test was weak to moderate (*r*s = .23–.60). Overall, CIWS obtained from 7-min samples appeared to have the strongest reliability and validity, although coefficients for 5-min samples were also deemed acceptable. The researchers used the 7-min samples to demonstrate how to construct "Tables of Probable Success," which used logistic regression to estimate the chances of passing the state standards test.

Summary and Discussion of Secondary Studies. In terms of reliability of measures of secondary-level writing, internal consistency (Parker et al., 1991a) and alternate-form reliability (Espin et al., 2000, 2005) of various scoring procedures generally appeared sufficient, with most *r*s > .70. With respect to criterion validity, however, the following seems clear: Simple, countable indices such as WW and WSC are not sufficient for assessing secondary students' writing. Tindal and Parker (1989a) and Parker et al. (1991a) found that for middle school remedial and special education students, percentage measures were better predictors of holistic ratings of writing (*r*s = .73 to .75) than were fluency measures (*r*s = .10 to .59). Moreover, percentage measures more reliably distinguished between students with and without LD than did fluency measures (Tindal & Parker, 1989a; Watkinson & Lee, 1992). However, while several measures appeared promising for screening, fine-grained analyses indicated that none was sufficiently sensitive to differences among low scorers, raising the concern of identifying false negatives (Parker et al., 1991b), and none was sufficient for monitoring progress (Parker et al., 1991a).

Espin and colleagues demonstrated that more complex measures, such as a combination of measures (Espin et al., 1999) or CIWS (Espin et al., 2000) might be needed. The latter is a promising finding because unlike percentage indices, CIWS is more viable for detecting growth, and unlike combinations of variables, it is likely to be less time consuming to score and interpret. Espin et al. (2000) also demonstrated that validity of CIWS does not appear to depend on type of writing prompt (narrative vs. expository) or sample duration (3 min vs. 5 min) for middle school students. However, Espin, De La Paz, et al. (2005) found that longer samples (35 min) yielded stronger validity than did 50-word samples. Further, Espin et al. (2006) demonstrated that for 10th-graders, validity of CIWS strengthened when the duration was increased to 7 min.

Espin et al. (2006) developed a procedure using Tables of Probable Success that could be used to predict students' probability of passing state standards tests based on CBM performance. Such a tool might be useful for identifying students at risk for failing state tests, setting reasonable goals, providing instructional accommodations, and monitoring student progress toward higher probabilities of passing. Research is needed to examine the validity of data generated by Tables of Probable Success and to determine whether schools' use of such a tool would lead to improved student outcomes. Such research, along with research examining the utility of measures for monitoring writing progress on a frequent basis to make instructional decisions, would shed further light on the consequential validity of secondary writing measures.

Studies Across Grade Levels

Results of studies reviewed thus far suggest that different scoring indices might be needed at different grade levels, raising questions about the "seamlessness" of writing measures across students of different ages. Secondary-level studies revealed that more complex scoring procedures had stronger technical characteristics than did simple scoring procedures. Elementary-level research is less conclusive. It is unclear whether simple indices such as WW and WSC are sufficient (as suggested by IRLD studies) or not sufficient (as suggested by more recent studies). To understand better whether different measures are needed at different levels, several researchers have compared the technical adequacy of measures across grade levels.

Malecki and Jewell (2003) observed that many districts were collecting normative writing data only on simple fluency measures (WW, WSC, CWS) and noted that whereas there is some evidence of the utility of these measures for elementary students (primarily from IRLD work), researchers have shown stronger technical adequacy of percentage measures (%WSC or %CWS; Tindal & Parker, 1989a) or CIWS (e.g., Espin et al., 2000) for older students. Malecki and Jewell investigated which indices were most appropriate at different grade levels. They also examined potential gender differences, which might be important for developing district norms.

Malecki and Jewell (2003) found reliable differences between grade levels on fall fluency and percentage measures, including CIWS, with sixth- through eighth-graders scoring higher than third- through fifth-graders did, who scored higher than first- through second-graders did. Girls reliably outperformed boys, and this gap widened over time on CWS and CIWS. For %WSC, girls outperformed boys in first through second grade, but the gap closed at later grades. At all grades, WW, WSC, CWS, and CIWS improved significantly from fall to spring, with no grade by time interactions. For the percentage measures, only first- and second-graders' scores showed reliable improvement.

Jewell and Malecki (2005) examined criterion validity of the fluency and percentage indices described above. Students at higher grades reliably outperformed students at lower grades with the exception of fourth- and sixth-graders on percentage measures. For second- and fourth-graders, weak to moderate correlations were found between most CBM scores and the SAT language subtests (*r*s = .34–.67) and between percentage scores, CIWS, and the SAT spelling subtest (*r*s = .43–.56). For sixth-graders, positive but relatively weak correlations were found between SAT subtests, percentage measures, and CIWS (*r*s = .41–.52). Language arts grades were weakly to moderately correlated with all scoring indices for fourth-graders (*r*s = .45–.61), but only with %WSC and CIWS for sixth-graders (.45 and .36, respectively). Finally, the analytic scoring system was weakly correlated with all scoring indices for second- and fourth-graders (*r*s = .34–.58) and with CWS, %WSC, %CWS, and CIWS for sixth-graders (*r*s = .33–.52).

Based on these results, Jewell and Malecki (2005) concluded that simple measures, such as WW and WSC, become less valid as grade level increases. They suggested using percentage measures or CIWS with middle school students. They also noted that percentage measures and CIWS were more strongly related to criterion measures at all grades, a finding consistent with previous research (Tindal & Parker, 1989a). They further cautioned that none of the validity coefficients was overwhelmingly strong, which was also consistent with other findings at elementary and secondary levels (e.g., Espin and colleagues; Gansle et al., 2002; Tindal & Parker). Jewell and Malecki's overall conclusion was that it is critical to consider students' gender and grade, as well as the purpose of assessment, when deciding which measures to use.

Weissenburger and Espin (2005) examined reliability and validity of narrative writing prompts across 4th, 8th, and 10th grades. Alternate-form reliability was moderate to strong at each grade for WW (*r*s = .55–.84), CWS (*r*s = .59–.84), and CIWS (*r*s = .61–.82). Correlations were stronger for longer writing samples and weaker at higher grades. Criterion validity of the measures with the Language Arts Normal Curve Equivalent (NCE) scores from a statewide test was weak at each grade level for WW (*r*s = .04–.45) and slightly stronger at Grades 4 and 8 for CWS (*r*s = .47–.62) and CIWS (*r*s = .60–.68). Validity of CWS and CIWS was weak at Grade 10 (*r*s = .18–.36). Similarly, coefficients with holistic writing scores from the statewide test were weak for WW (*r*s = .33–.48) and moderate for CWS and CIWS (*r*s = .50–.65) for 4th- and 8th-graders (data were not available for 10th-graders). Although increased duration led to increased alternate-form reliability, it generally did not strengthen the validity of the measures.

Findings from the above studies suggest that criterion validity of CBM in written expression decreases as students get older. However, in these studies, only narrative samples were used. Because older students are often required to produce more expository than narrative writing (e.g., Deshler, Ellis, & Lenz, 1996), the validity of expository prompts warrants further investigation. Also, although the validity of measures administered in the Weissenburger and Espin (2005) study did not increase substantially with time, previous research has indicated that longer samples do increase validity of writing scores (e.g., Espin et al., 2005; Espin et al., 2006). The effect of duration on the criterion validity of writing samples should also be further explored.

Implications for Future Research

An extensive amount of work has been done to identify technically sound approaches to assessing written expression within a CBM framework, that is, using measures that are simple and efficient to obtain reliable and valid indicators of overall writing proficiency. This work has provided a foundation for further research on monitoring writing progress. It is clear that much work is needed to develop seamless and flexible writing measures to be used within a system of accountability, whereby students at risk for failing to meet standards are identified, intervention effectiveness is evaluated, and student progress within and across grade levels is monitored.

Reliability

Most of the research on CBM in written expression has reported reliability of static (one point in time) measures or stability across a wide timeframe (e.g., fall to spring). However, an important goal is to develop measures that can be used to monitor student progress on a frequent basis to facilitate instructional decisions. This is an area wide open for further investigation. For example, possible variation in scores obtained from different writing prompts has particular implications for future research addressing the use of measures for monitoring writing progress. There might be considerable variability in the interest and background knowledge that students bring to different writing prompts, which could affect the quality and quantity of their responses and could, in turn, lead to substantial bounce from one measurement point to the next. None of the studies in this review addressed reliability of slopes, yet this is a critical component of research needed in the development of progress-monitoring tools.

Validity

Procedures developed thus far have yielded more modest criterion validity coefficients than have those obtained in other areas of CBM research. In fact, recent research at both elementary and secondary levels has yielded only moderate criterion validity at best. This may be a result of criterion measures that do not directly assess written expression, such as language subtests of standardized measures, poorly constructed criterion measures (other writing measures that have questionable technical adequacy themselves), or varying criteria for holistic ratings that tend to have only moderate interscorer reliability.

Modest validity coefficients might also be a reflection of the difficulty associated with measuring the complex, multifaceted construct of writing proficiency. Although validity coefficients for CBM writing measures have generally been lower than those seen for CBM measures in other academic areas, coefficients in many of the studies in this review are similar to or better than those seen for other commonly used measures of written expression. For example, the criterion-related validity reported for the TOWL-3 (Hammill & Larsen, 1996) ranges from .34 to .68 for the various subtests (with only the spelling subtest above .60). Given the general difficulties associated with measuring writing proficiency in evaluating the technical adequacy of CBM, it is important to search for consistent findings across multiple criterion measures that tap various aspects of this construct (Messick, 1995). Further, it is important to examine divergent validity, that is, whether CBM writing measures correlate more strongly with other writing measures than they do with measures in other domains, such as reading or math.

Two critical aspects of validity, discussed earlier in this review, are the *generalizability* and *consequential validity* of measures, which are necessary if writing measures are to be used within a seamless and flexible system of progress monitoring. In terms of generalizability, much work in written expression has focused on students in general education or those with high-incidence disabilities; whether the materials and procedures are appropriate for other populations of students, such as English learners and those with significant disabilities who access the general education curriculum, is not yet well understood. With respect to consequential validity, to facilitate educational decision making within and across grades, research is needed to determine which measures are most appropriate at which grade levels and to establish methods to connect student progress both within and across grades. Finally, if measures are used by teachers to monitor progress and make instructional decisions, it is necessary to demonstrate that student performance improves as a result.

Implications for Practice

Clearly, much work remains to identify the most useful measures for monitoring students' writing progress. Yet, it is evident that teachers, schools, and districts are already using such measures (Fewster & MacMillan, 2002; Gansle et al., 2002; Malecki & Jewell, 2003) and that WW or WSC is often used as the primary index of writing proficiency. Educators should use caution in interpreting results of these measures. There is some evidence that simple, countable indices of written expression are useful for screening (Parker et al., 1991a, 1991b; Watkinson & Lee, 1992), and percentage measures appear to be more technically sound for this purpose than do fluency measures. There is also evidence that, at least at the higher grades, more complex scoring procedures, such as CIWS, are more technically sound (e.g., Espin et al., 2000; Jewell & Malecki, 2005). Further, for instructional decision making, educators might wish to consider qualitative, as well as quantitative, aspects of students' writing (Tindal & Hasbrouck, 1991). Finally, educators should keep an eye toward the research for further development of progress-monitoring approaches in written expression. It is our hope that upcoming research will lead to great improvements in the technical soundness and instructional utility of CBM in written expression, eventually leading to a seamless and flexible system for monitoring student progress in writing.

REFERENCES

Berninger, V. W. (1994). Developmental skills related to writing and reading acquisition in the intermediate grades. *Reading and Writing: An Interdisciplinary Journal, 6,* 161-196.

CTB/McGraw-Hill. (1985). *California achievement test.* Monterey, CA: Author.

CTB/McGraw-Hill. (1996). CTB/McGraw-Hill, TerraNova, Monterey, CA: Author.

De La Paz, S. (1999). Self-regulated strategy instruction in regular education settings: Improving outcomes for students with and without learning disabilities. *Learning Disabilities Research & Practice, 14*, 92–106.

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

Deno, S. L., Marston, D., & Mirkin, P. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children Special Education and Pediatrics: A New Relationship, 48*, 368–371.

Deno, S., Marston, D., Mirkin, P., Lowry, L., Sindelar, P., & Jenkins, J. (1982). *The use of standard tasks to measure achievement in reading, spelling, and written expression: A normative and developmental study* (Vol. IRLD-RR-87). University of Minnesota, Institute for Research on Learning Disabilities.

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.

Deshler, 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. *The 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. *The Journal of Special Education, 34*, 140–153.

Espin, C., Wallace, T., Campbell, H., Lembke, E. S., Long, J. D., & Ticha, R. (2006). *Predicting the success of secondary-school students on state standards tests: Validity and reliability of curriculum-based measures in written expression.* Manuscript submitted for publication.

Fewster, S., & MacMillan, P. D. (2002). School-based evidence for the validity of curriculum-based measurement of reading and writing. *Remedial and Special Education, 23,* 149–156.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to responsiveness-to-intervention: What, why, and how valid is it? Reading Research Quarterly. *Reading Research Quarterly, 41,* 92–99.

Fuchs, L. S., Deno, S. L., & Marston, D. (1982). *Use of aggregation to improve the reliability of simple direct measures of academic performance* (Vol. IRLD-RR-94). University of Minnesota, Institute for Research on Learning Disabilities.

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. (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.

Hammill, D. D., & Larsen, S. C. (1978). *Test of written language.* Austin, TX: PRO-ED.

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

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. Gregg & E. Steinberg (Eds.), *Cognitive processes in writing* (pp. 31–50). Hillsdale, NJ: Erlbaum.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1996). *Iowa tests of basic skills.* Itasca, IL: Riverside.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (Research Report No. 3). Champaign, IL: National Council of Teachers of English.

Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. § 1400 *et seq.* (2004)(reauthorization of the Individuals with Disabilities Education Act of 1990)

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.

Lee, L., & Canter, S. M. (1971). Developmental sentence scoring. *Journal of Speech and Hearing Disorders, 36,* 335–340.

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.

Madden, R., Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1978). *Stanford achievement test.* New York: Harcourt Brace Jovanovich.

Malecki, C. K., & Jewell, J. (2003). Developmental, gender, and practical considerations in scoring curriculum-based measurement writing probes. *Psychology in the Schools, 40*, 379–390.

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 Press.

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.

Marston, D., Lowry, L., Deno, S. L., & Mirkin, P. (1981). *An analysis of learning trends in simple measure of reading, spelling, and written expression: A longitudinal study* (Vol. IRLD-RR-49). University of Minnesota, Institute for Research on Learning Disabilities.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review, 8*, 299-325.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *The American Psychologist, 50*, 741–749.

Minnesota Department of Children, Families, and Learning and NCS Pearson. (2002). *Minnesota Basic Skills Test (BST) technical manual for the academic year 2001–2002.*

Mitzel, H. C., & Borden, C. F. (2000). *LEAP for the 21st century: 1999 operational final technical report.* Monterey, CA: CTB/McGraw-Hill.

National Center for Education Statistics. (2003). *The nation's report card: Writing.* Retrieved April 23, 2006, from http://nces.ed.gov/nationsreportcard/writing/

No Child Left Behind Act of 2001, 20 U.S.C. 70 § 6301 *et seq.* (2002)

Nolet, V., & McLaughlin, M. J. (1997). Using CBM to explore a consequential basis for the validity of a state-wide performance assessment. *Diagnostique, 22*, 147–163.

Nolet, V., & McLaughlin, M. J. (2000). *Accessing the general curriculum: Including students with disabilities in standards-based reform.* Thousand Oaks, CA: Corwin Press.

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

Parker, R. I., Tindal, G., & Hasbrouck, J. (1991b). Progress monitoring with objective measures of writing performance for students with mild disabilities. *Exceptional Children, 58*(1), 61–73.

Shinn, M. R., Ysseldyke, J., Deno, S. L., & Tindal, J. (1982). *A comparison of psychometric and functional differences between students labeled learning disabled and low achieving* (Vol. IRLD-RR-71). University of Minnesota, Institute for Research on Learning Disabilities.

Speece, D. L., Case, L. P., & Molloy, D. E. (2003). Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities Research & Practice, 18,* 147–156.

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

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson Education.

Tindal, G., Germann, G., & Deno, S. L. (1983). *Descriptive research on the Pine County norms: A compilation of findings* (Vol. IRLD-RR-132). University of Minnesota, Institute for Research on Learning Disabilities.

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. (1989a). Assessment of written expression for students in compensatory and special education programs. *The Journal of Special Education, 23*, 169–183.

Tindal, G., & Parker, R. (1989b). Development of written retell as a curriculum-based measure in secondary programs. *School Psychology Review, 13*, 328–343.

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

Videen, J., Deno, S. L., & Marston, D. (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.

Watkinson, J. T., & Lee, S. W. (1992). Curriculum-based measures of written expression for learning-disabled and nondisabled students. *Psychology in the Schools, 29*, 184–192.

Weissenburger, J. W., & Espin, C. A. (2005). Curriculum-based measures of writing across grade levels. *Journal of School Psychology, 43*, 153–169.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson psychoeducational battery* (Rev. ed.). Allen, TX: DLM.