Literature Synthesis on Curriculum-Based Measurement in Reading

Miya Miura Wayman, Teri Wallace, Hilda Ives Wiley, Renáta Tichá, and **Christine A. Espin,**
University of Minnesota

In this article, the authors review the research on curriculum-based measurement (CBM) in reading published since the time of Marston's 1989 review. They focus on the technical adequacy of CBM related to measures, materials, and representation of growth. The authors conclude by discussing issues to be addressed in future research, and they raise the possibility of the development of a seamless and flexible system of progress monitoring that can be used to monitor students' progress across students, settings, and purposes.

Curriculum-based measurement (CBM) is a method for monitoring student growth in an academic area and evaluating the effects of instructional programs on that growth (Deno, 1985). CBM was designed to be part of a problem-solving approach to special education whereby the academic difficulties of students would be viewed as problems to be solved rather than as immutable characteristics within a child (Deno, 1990). In the problem-solving approach, teachers were the "problem solvers" who constantly evaluated and modified students' instructional programs. For a problem-solving approach to be effective, it was necessary for teachers to have a tool that could be used to evaluate growth in response to instruction. CBM was developed to serve that purpose.

Two separate but related concerns drove the initial research into the development of CBM (Deno, 1985). The first was the concern for technical adequacy. If teachers were to use the measures to make instructional decisions, the measures would have to have demonstrated reliability and validity. The second was the concern for practicality. If teachers were to use the measures on an ongoing and frequent basis to evaluate instructional programs, the measures would have to be simple, efficient, easily understood, and inexpensive. These dual concerns led to the concept of "vital signs," or indicators of student performance (Deno, 1985). CBM measures were conceptualized to be short samples of work that would be indicators, or vital signs, of academic performance. The samples would need to be valid and reliable with respect to the broader academic domain they were representing, but would also need to be designed to be given on a frequent and repeated basis.

In 1989, Marston reviewed the existing research on CBM. At that time, CBM was viewed primarily as a progress-monitoring tool in basic skills for special education students at the elementary-school level (although there were discussions and instances of its uses more broadly, for example, see Shinn, 1989). Research in reading focused on two measures: word identification and reading aloud. The results of Marston's review provided support for the use of these two measures as indicators of general reading proficiency. In terms of reliability, results of five studies revealed test–retest reliability coefficients ranging from .82 to .97, with most coefficients above .90, and alternate-form reliability coefficients ranging from .84 to .96, with most coefficients above .90. Interrater agreement was .99. In terms of validity, 14 studies were reviewed. Criterion-related validity coefficients with published measures of reading ranged from .63 to .90, with most above .80. Criterion-related validity coefficients with basal reading series criterion mastery tests ranged from .57 to .86, with half above .80. Reading aloud correlated with teacher judgment and with various measures of reading comprehension, discriminated between lower and higher performing students, and was sensitive to growth.

Since the time of Marston's (1989) review, the research on CBM has expanded considerably—especially in the area of reading—making an updated review timely. Our purpose in writing this review is to gather, summarize, and reflect on the expansive body of literature published over the last 18 years on CBM in reading. We focus this review on issues of technical adequacy as they relate to measures, materials, and growth. Given the vast amount of material and the diversity of topics covered in this review, we insert summaries and discussion points throughout the article. In our final section, we draw conclusions, raise issues related to future research, and discuss the potential development of a seamless and flexible system of progress monitoring that can be used across students, settings, and purposes.

Method

The first step in our review process was to identify all articles addressing CBM in reading, writing, and math in kindergarten (K) to Grade 12. Electronic databases—including ERIC,

Address: Miya Miura Wayman, Department of Educational Psychology, 250D Burton Hall, 178 Pillsbury Dr. SE, University of Minnesota, Minneapolis, Minnesota 55455. E-mail: miura001@umn.edu

Science Citation Index Expanded, PsycInfo, Digital Dissertation, and the Expanded Academic Index—were searched using the following terms: *curriculum based measurement, curriculum-based measurement, curriculum based measure, curriculum-based measure, general outcome measure,* and *progress monitoring*. This initial search yielded 578 articles, dissertations, and reports related to CBM. Titles and abstracts of these documents were screened to confirm that they were related to CBM, and Method sections were screened to identify those that reported results of empirical studies of CBM, yielding 160 documents. These documents were then reviewed by a team of educational psychology graduate students and grouped by subject area (reading, mathematics, spelling, and writing). Ninety (56%) of the documents addressed reading measures. In addition to documents identified through the literature search, the complete set of technical reports produced by the Institute for Research on Learning Disabilities (IRLD) at the University of Minnesota was accessed.

Given the vast database on reading, and the limitations of space accorded a review article, we chose to narrow our review in several key ways. First, we focused on research published since the time of Marston's 1989 review. Second, we focused on studies related to the technical adequacy of reading measures. Studies in which teachers used CBM to monitor progress and to make instructional decisions have been reviewed recently (Stecker, Fuchs, & Fuchs, 2005). Third, we focused on research conducted with school-age students and thus did not include research on the development of early literacy (e.g., *Dynamic Indicators of Basic Early Literacy* [DIBELS]; Kaminski & Good, 1998) or preliteracy measures (e.g., *Individual Growth and Development Indicators;* McConnell, McEvoy, & Priest, 2002). Finally, we focused on three common CBM reading measures: reading aloud, maze selection, and word identification (see Note 1). Before reporting the results of our review, we present a brief discussion of the issue of validity, relating it specifically to the approach taken to investigate the validity of measures in CBM.

Issues of Validity

Messick (1989a, 1989b) described construct validity as a multifaceted, but unified, concept that takes into account the evidential and consequential factors related to test interpretation and test use. This conceptualization is helpful in understanding the research on CBM, which has examined the evidence supporting the interpretation and use of CBM scores, as well as the consequences associated with that interpretation and use. In this review, we focus on the evidential basis for validity of CBM measures, although, in keeping with the view of validity as a unified concept, we continuously consider the potential consequences associated with interpretation and use of the measures.

In a CBM approach, evidence for the validity of a measure is determined by examining the extent to which the measure serves as a vital sign or an indicator of a broader academic domain (Deno, 1985). To determine the evidential basis for the validity of a measure, the pattern of relations—also referred to as the nomological net (see Cronbach & Meehl, 1955; Messick, 1989b)—between the selected measure and many different criterion measures, each reflecting the construct of interest, is examined. Criterion measures might include other measures of the construct, such as standardized achievement tests, but might also include student age, group membership, and change in performance in response to an intervention. It is not just the pattern of relations but also the pattern of nonrelations, or the discriminative validity of the measure, that is considered. For example, a measure of reading would be expected to relate more closely to another reading measure than to a math measure. After a pattern of relations is established for particular students, settings, or purposes, the generalizability (Messick, 1989b) of the measure, or the extent to which the validity of the measure holds across different settings, students, and purposes, is examined.

Establishing the validity of a measure is an ongoing and recursive process (Messick, 1989b). Validity is determined not by one study or by one correlation but by the body of evidence amassed over time. The question arises, then, as to how one determines when the data are strong enough to support the use of the measure for the purpose for which it was intended. In other words, how good is good enough? There is no set standard for determining when a measure is "good enough" to be considered valid for the purpose for which it was designed. One approach is to consider the consequences of decision making with and without the measure (Messick, 1989b). For example, one might ask whether using a measure improves the ability to make decisions over using no measure at all. In an area in which few measures exist, this question might be warranted. Under such circumstances, correlations of .30 between the selected measure and the criterion measures might be considered strong enough to warrant use of an instrument for decision-making purposes. However, in the area of reading, where many measures exist, it seems appropriate to apply more stringent criteria. In addition, the early research on CBM, as illustrated in Marston's (1989) review, set a relatively high standard for validity and reliability, with many correlations reaching levels of .70 to .90. Thus, in this article, we adopt the following guidelines in interpreting the strength of the reliability and validity coefficients: Strong relations are those that are .70 and above; moderate relations are those that are .50 to .70; and weak relations are those that are below .50. We remind the reader, however, that these levels are arbitrarily chosen and used merely to help the reader interpret the strength of relations compared to previous research in reading in CBM. To evaluate the overall validity of the measures, it is necessary to consider the entire body of research.

Our review is organized into three sections: Technical Adequacy of CBM Measures, Effects of Text Materials, and Issues About Measuring Growth. The studies described in each section are also outlined in Table 1. We begin our review

(text continues on p. 105)

TABLE 1. Characteristics of Studies Examining Technical Features of CBM in Reading

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
					Reading Aloud				
Fuchs, Fuchs, & Maxwell (1988)	35	4–8	SE	Reading aloud	1	WRC	*Stanford Achievement Test* Reading Comprehension subtest: .91 Word Skills subtest: .80		
Shinn, Good, Knutson, & Tilly (1992)	238	3 & 5 3 5	ND	Reading aloud	1	WRC	Confirmatory factor analysis Unitary Model Reading Competence: .88–.90 Two-Factor Model Decoding: .89–.90 Comprehension: .74–.75		
Hosp & Fuchs (2005)	310	1–4 1 2 3 4	ND	Reading aloud	1	WRC	*Woodcock Reading Mastery Test–Revised* (WRMT-R) Word Attack: .71 Word Identification: .91 Passage Comprehension: .79 Basic Skills: .86 Total Reading–Short: .90 Word Attack: .82 Word Identification: .88 Passage Comprehension: .83 Basic Skills: .89 Total Reading–Short: .91 Word Attack: .82 Word Identification: .88 Passage Comprehension: .84 Basic Skills: .87 Total Reading–Short: .91 Word Attack: .72 Word Identification: .73 Passage Comprehension: .82 Basic Skills: .78 Total Reading–Short: .83	Test–retest .92–.97	
Kranzler, Brownell, & Miller (1998)	57	4	GE	Reading aloud	1	WRC	Elementary Cognitive Tasks: –.13 *Kaufman Brief Intelligence Test* Matrices: .24 *Kaufman Test of Educational Achievement* Reading Comprehension: .41 Reading aloud explained 11% of the variance when controlling for general cognitive ability and mental speed		

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
Markell & Deno (1997)	42	3	GE	Reading aloud	1	WRC	Large increases in reading aloud scores were associated with increases in performance on maze selection and comprehension questions. An inverse relationship was found between difficulty level and mean reading aloud scores.		
Hamilton & Shinn (2003)	66	3	ND	Reading aloud	1	WRC	Word Callers performed lower than did Similarly Fluent Peers on Reading Aloud, Maze Selection, Comprehension, Oral Question Answering, and *Woodcock Reading Mastery Test* Passage Comprehension.		
				Maze Selection					
Espin, Deno, Maruyama, & Cohen (1989)	2,604	1–6	ND	Maze selection	1	CS	Reading Aloud		Pattern of growth found within and between grades
		1							
		2							
		3					.77		
		4					.86		
		5					.86		
		6							
Fuchs & Fuchs (1992)									(18 weeks, 2 times per week)
Year 1	63	5.12 (average grade)	SE	Maze selection	2.5	CS			.31 CS per week
Year 2	63	5.12 (average grade)	SE	Maze selection	2.5	CS			.29 CS per week
	257	—	GE	Maze selection	2.5	CS			.39 CS per week
Jenkins & Jewell (1993)	335	2–6	ND	Maze selection	1	CS	*Gates-MacGinitie* Total Reading: .65–.76 *Gates-MacGinitie* Comprehension: .63–.75 *Metropolitan Achievement Test–6* (MAT-6) Total Reading: .66–.76 MAT-6 Comprehension: .60–.74 Teacher judgment: .56		Pattern of growth found within and between grades
				Reading aloud	1	WRC	*Gates-MacGinitie* Total Reading: .67–.88 *Gates-MacGinitie* Comprehension: .63–.86 MAT-6 Total Reading: .60–.87 MAT-6 Comprehension: .58–.84 Teacher judgment: .66		

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
Ardoin et al. (2004)	75	3	GE	Maze selection	3	CS	*Woodcock-Johnson–III* Broad Reading: .50 Reading Fluency: .51 Passage Comprehension: .31 Letter Word Identification: .43 *Iowa Test of Basic Skills* Total Reading: .46 Vocabulary: .35 Reading Comprehension: .49		
	77	3	GE	Reading aloud (single passage)	1	WRC	*Woodcock-Johnson–III* Broad Reading: .70 Reading Fluency: .74 Passage Comprehension: .42 Letter Word Identification: .62 *Iowa Test of Basic Skills* Total Reading: .64 Vocabulary: .35 Reading Comprehension: .58		
				Word Identification					
Daly, Wright, Kelly, & Martens (1997)	30	1	GE	Word identification	1	WRC	Concurrent: *Woodcock-Johnson–Revised* Broad Reading: .40 Predictive Word Identification: .71 Reading Aloud: .73	Test–retest .94	
Fuchs, Fuchs, & Compton (2004)	151	1	LA	Word identification	1	WRC	Concurrent *Woodcock Reading Mastery Test–Revised* (WRMT-R) Word Attack: .52–.59 WRMT-R Word Identification: .77–.82 *Comprehensive Reading Assessment Battery* (CRAB) Fluency: .93 CRAB Comprehension: .73 Predictive WRMT-R Word Attack: .45 Word Identification: .63 CRAB Fluency: .80 CRAB Comprehension: .66 Slope (1 year) WRMT-R Word Attack: .50 WRMT-R Word Identification: .79 CRAB Fluency: .85 CRAB Comprehension: .66		

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
Compton, Fuchs, Fuchs, & Bryant (2006)	206	1	LA	Word identification	1	WRC	Predictive Grade 1 5-week word ID level: .80 5-week word ID slope: .42 Grade 2 Untimed Word-Level Reading: .39 Timed Word-Level Reading: .46 Reading Comprehension: .42 Composite Score: .47		
Generalizability of Research: Student Populations and Purposes									
Espin & Deno (1993a)	121	10	HA & LA	Reading aloud (English)	1	WRC			
			HA				*Tests of Achievement and Proficiency* (TAP) Complete Composite: .36 Grade point average: –.02 English study task: .34 Science study task: .43		
			LA				TAP–Complete Composite: .71 Grade point average: .39 English study task: .54 Science study task: .36		
Espin & Deno (1993b)	121	10	ND	Reading aloud (English)	1	WRC		Alternate-form .91 Test–retest .91	
Espin & Deno (1995)	120	10	ND	Reading aloud (English)	1	WRC	English study task: .41		
				Vocabulary matching (English)	10	CM	.44		
				Reading aloud (Science)	1	WRC	Science study task: .32		
				Vocabulary Matching (Science)	10	CM	.40		
Fewster & MacMillan (2002)	465	6–7	GE	Reading aloud	1	WRC	8th-grade English mark: .46 8th-grade Social Studies mark: .39		
				Reading aloud	1	WRC	9th-grade English mark: .38 9th-grade Social Studies mark: .36		
				Reading aloud	1	WRC	10th-grade English mark: .31 10th-grade Social Studies mark: .30		

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
Yovanoff, Duesbery, Alonzo, & Tindal (2005)	6,012	4–8	ND	Reading aloud	1	WRC			
		4					Vocabulary: .35–.63 Comprehension: .60–.65		
		5					Vocabulary: .61–.65 Comprehension: .60–.62		
		6					Vocabulary: .60–.61 Comprehension; .42–.52		
		7					Vocabulary: .49–.53 Comprehension; .42–.48		
		8					Vocabulary: .53–.57 Comprehension: .51–.52		
							Relative importance of fluency decreases as grade increases. Vocabulary is a significant predictor across grades.		
Espin & Foegen (1996)	176	6–8	ND	Reading aloud	1	WRC	Comprehension: .57 Acquisition: .54 Retention: .52		
				Maze selection	2	CS	Comprehension: .56 Acquisition: .59 Retention: .62		
				Vocabulary matching	5	CM	Comprehension: .65 Acquisition: .64 Retention: .62		
Hixson & McGlinchey (2004)	442	4	ND	Reading aloud	1	WRC	*Michigan Educational Assessment Program* (MEAP) Reading Caucasian: .54 African American: .54 Paid Lunch: .53 Free/Reduced Lunch: .50 *Metropolitan Achievement Test–7* (MAT-7) Caucasian: .78 African American: .64 Paid Lunch: .75 Free/Reduced Lunch: .66 Reading aloud, lunch status, and race each significantly contributed to predicting performance on MEAP and MAT-7		

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
Kranzler, Miller, & Jordan (1999)	326	2–5	GE	Reading aloud	1	WRC	At Grades 4–5, significant intercept biases were found based on race/ethnicity. At Grade 5, significant intercept and slope bias was found based on gender. No biases found at Grades 2–3.		
Hintze, Callahan, Matthews, Williams, & Tobin (2002)	136	2–5	GE	Reading aloud	1	WRC	Age and reading aloud each significantly contributed when predicting performance on *Woodcock Johnson–Revised* Reading Comprehension. Race and socioeconomic status did not contribute significantly when added to and reading aloud model.		
Baker & Good (1995)	76	2	GE/ELL	Reading aloud	1	WRC	Convergent Construct English-only *Stanford Diagnostic Reading Test* (SDRT)–Total Score: .51 *Stanford Reading* Comprehension subtest–pretest: .56 Teacher rating: .82 Bilingual SDRT–Total Score: .53 *Stanford Reading* Comprehension subtest–pretest: .73 *Stanford Reading* Comprehension subtest–posttest: .76 Teacher rating of reading: .80 Discriminant Construct English-only English Language Fluency: .54 Teacher rating of English: .62	English-only Point .87 Level .99 Slope .39 Bilingual Point .92 Level .99 Slope .49	(10 weeks, 2 times per week) Significantly different slopes were found for English-only and bilingual students with bilingual students showing greater growth.

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
							Bilingual English Language Fluency: .44 Teacher rating of English: .62 Language Assessment Scale–English: .47		
Wiley & Deno (2005)		3 & 5	GE & ELL				*Minnesota Comprehensive Assessment* (MCA)		
	36	3	GE	Reading aloud	1	WRC	.71		
				Maze selection	1	CS	.73		
			ELL	Reading aloud			.61		
				Maze selection			.52		
	33	5	GE	Reading aloud			.57		
				Maze selection			.73		
			ELL	Reading aloud			.69		
				Maze selection			.57		
Graves, Plasencia-Peinado, Deno, & Johnson (2005)	77	1	ELL, HA, AA, & LA	Reading aloud		WRC			(6 weeks, weekly)
			ELL/HA						1.40 WRC per week
			ELL/HA						2.30 WRC per week
			ELL/ LA						2.10 WRC per week
Klein & Jimerson (2005)	4,000	1–3	GE	Reading aloud	1	WRC	*Stanford Achievement Test–9* Total Reading, concurrent Hispanic: .71–.82 Caucasian: .63–.82 Female: .75–.82 Male: .73–.85 Free/Reduced Lunch: .73–.82 Regular Lunch: .64–.83 Home Language Spanish: .70–.82 Home Language English: .67–.82 SAT-9 Total Reading, predictive Hispanic: .66 Caucasian: .58 Female: .72 Male: .62 Free/Reduced Lunch: .62		

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
							Regular Lunch: .70 Home Language Spanish: .65 Home Language English: .63 Intercept bias evident when ethnicity and home language factors were combined		
Morgan & Bradley-Johnson (1995)	15	3–7	VI	Reading aloud		WRC		Alternate form	
					1		*Diagnostic Reading Scales* Decoding Level: .60	.88–.93	
					2		.66	.94–.98	
					3		.65	.94–.96	
								Test–retest	
					1		*Diagnostic Reading Scales* Comprehension Level: .78	.84–.94	
					2		.80	.93–.96	
					3		.82	.93–.97	
					1		*Diagnostic Reading Scales* Overall Reading Level: .76		
					2		.80		
					3		.80		
Allinder & Eccarius (1999)	36	—	D/HH	Reading aloud		WRC	*The Test of Early Reading Ability—Deaf and Hard of Hearing Version*	Alternate form	
					1		.30	.85	
					3		.21	.94	
Crawford, Tindal, & Stieber (2001)	51	2 & 3	GE		1	WRC	Statewide reading assessment (multiple-choice)		
		2					Predictive: .66		
		3					Concurrent: .60		
Good, Simmons, & Kame'enui (2001)	706	1 & 3	GE	Reading aloud	1	WRC	Benchmark goal of 40 WRC in 1 min in Grade 1 predicted continued reading success in Grade 2. Benchmark goal of 110 WRC in 1 min in Grade 3 predicted success on the *Oregon Statewide Assessment*.		
Hintze & Silberglitt (2005)	1,766	1–3	ND	Reading aloud	1	WRC	*Minnesota Comprehensive Assessment* (MCA)	Alternate form	
		1					Predictive: .49–.58	.89–.91	
		2					Predictive: .61–.68	.80–.85	
		3					Concurrent: .69	.83–.87	

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
							Results of discriminative analysis, logistic regression, and Receiver Operating Characteristic (ROC) curves indicated that reading aloud is an efficient method for predicting success on the MCA.		
McGlinchey & Hixson (2004)	1,362	4	ND	Reading aloud	1	WRC	*Michigan Educational Assessment Program* (MEAP): .49–.81 Sensitivity = 75% Specificity = 74% Correct classification = 74% MEAP base failure rate = 54% MEAP base pass rate = 46%		Test–retest .87–.95
Stage & Jacobson (2001)	173	4	ND	Reading aloud	1	WRC	*Washington Assessment of Student Learning* (WASL) Level: .51 Sensitivity = 66% Specificity = 76% Correct classification = 74% Base failure rate = 20% Base pass rate = 80%		
Silberglitt & Hintze (2005)	2,191	1–3 1 2 3	GE	Reading aloud	1	WRC	*Minnesota Comprehensive Assessment* (MCA) Predictive: .57 Predictive: .67 Concurrent: .71 Logistic regression and ROC curve analysis were the strongest methods for establishing cut scores. ROC curve analysis was the most flexible.		
						Curriculum Effects			
Tindal, Marston, Deno, & Germann (1982, IRLD #93)	660	1–6	GE	Reading aloud Word identification	1	WRC	Different curricula produced mean level differences in reading aloud and word identification scores. All curricula showed across grade growth.		
Tindal, Flick, & Cole (1992)	12	2–5	SE	Reading aloud	1	WRC			(31 weeks, 1–2 times per week); student performance on passages from instructional material was similar to performance on passages from mainstream material.

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
Fuchs & Deno (1992)	91	1–6	ND	Reading aloud	1	WRC	*Woodcock Reading Mastery Test* Passage Comprehension		
				Ginn 720 (preprimer–7)			.89–.93		
				Scott-Foresman Unlimited (preprimer–6)			.90–.93		
Hintze, Shapiro, Conte, & Basile (1997)	57	2–4	ND	Reading aloud	1	WRC	*The Degrees of Reading Power Test*		
				Authentic trade books (Levels 1–5)			.64–.69		
				Literature-based basal (Levels 1–5)			.62–.69		
Hartman & Fuller (1997)	26	1	GE				*Stanford Achievement Test*	Test–retest .91–.99	
	24	2		Reading aloud	1	WRC	Word Reading Skills: .72 Reading Comprehension: .89		
				Word identification	1	WRC	Word Reading Skills: .71 Reading Comprehension: .82		
	45	3		Reading aloud	1	WRC	Word Reading Skills: .71 Reading Comprehension: .80		
				Word identification	1	WRC	Word Reading Skills: .52 Reading Comprehension: .56		
Brown-Chidsey, Johnson, & Fernstrom (2005)	21	5	ND	Maze selection controlled passage	2	CS	Concurrent Literature-Based Passage: .74–.92		Controlled and literature-based passages showed significant growth across fall, winter, and spring. Mean scores were consistently higher for controlled passages.
Bradley-Klug, Shapiro, Lutz, & DuPaul (1998)		2 & 5	GE	Reading aloud	1	WRC			
	28	2							Literature = 1.36 WRC per week Basal = 1.32 WRC per week
	30	5							Literature = 0.84 WRC per week Basal = 0.32 WRC per week (10 weeks, 2 times per week)

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
Hintze, Shapiro, & Lutz (1994)	24	3	GE	Reading aloud	1	WRC		Parallel form	(9 weeks, 2 times per week)
				Literature-based passage				.56–.95	Literature Group = −1.04 WRC per per week Traditional Group = -0.34 WRC per week
				Traditional basal passage				.60–.96	Literature Group = 0.66 WRC per week Traditional Group = 1.72 WRC per week
Hintze & Shapiro (1997)	160	2–5	GE	Reading aloud	1	WRC		Parallel form .65–.97	(8 weeks, 2 times per week)
		2		Literature-based passage					Literature Group = 0.56 WRC per week Traditional Group = 0.68 WRC per week
				Traditional basal passage					Literature Group = −0.46 WRC per week Traditional Group = −0.26 WRC per week
		3		Literature-based passage					Literature Group = 1.81 WRC per week Traditional Group = 1.58 WRC per week
				Traditional basal passage					Literature Group = 0.67 WRC per week Traditional Group = 0.38 WRC per week
		4		Literature-based passage					Literature Group = 2.69 WRC per week

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
		5		Traditional basal passage Literature-based passage Traditional basal passage					Traditional Group = 2.29 WRC per week Literature Group = 1.72 WRC per week Traditional Group = 1.90 WRC per week Literature Group = 1.18 WRC per week Traditional Group = 2.05 WRC per week Literature Group = 1.72 WRC per week Traditional Group = 2.82 WRC per week
Powell-Smith & Bradley Klug (2001)	36	2	LA/ Ch. 1	Reading aloud per week	1	WRC			Basal = 2.69 WRC Generic = 2.41 WRC per week (5 weeks, 2 times per week)
Riley-Heller, Kelly-Vance, & Shriver (2005)	13	2	LA	Reading aloud	1	WRC			.65 WRC per day Generic = Curriculum-dependent (5 weeks, 2 times per week)
				Difficulty Level					
Shinn, Gleason, & Tindal (1989)	30	3–8	SE/Ch. 1	Reading aloud	1	WRC			(4 weeks, 4 times per week) 1 below grade level = 4.30 WRC per week 1 above grade level = 3.70 WRC per week

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
									2 above grade level = 4.55 WRC per week 4 above grade level = 2.35 WRC per week
Dunn & Eckert (2002)	20	2–3	GE	Reading aloud	1	WRC			Similar material .65 WRC per week Challenging material .92 WRC per week (8 weeks, 2 times per week)
Hintze, Daly, & Shapiro (1998)	80	1–4	ND	Reading aloud	1	WRC			(11 weeks, 2 times per week)
		1							Grade Level = 3.29 WRC per week Goal Level = 1.95 WRC per week
		2							Grade Level = .72 WRC per week Goal Level = .30 WRC per week
		3							Grade Level = .16 WRC per week Goal Level = .06 WRC per week
		4							Grade Level = 1.57 WRC per week Goal Level = 1.85 WRC per week
Fuchs, Tindal, & Deno (1981, IRLD #48, Study III)	20	2-4 (reading instructional level)	GE	Word identification	5 .	WRC			Slope (2 weeks, daily): Grade-Specific Domain ≠ Entire Grade Domain Entire Grade Domain ≠ Across Grade Domain Standard Error of Estimates:

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
									Grade-specific domain ≠ Entire grade domain Entire grade domain = across grade
Poncy, Skinner, & Axtell (2005)	37	3	ND	Reading aloud	1	WRC		Student skills accounted for 81% of the variance, and passage accounted for 10% of the variance.	
Hintze & Christ (2004)	99	2–5	ND	Reading aloud	1	WRC			(11 weeks, 2 times per week)
		2							Uncontrolled = −.06 WRC per week Controlled = .25 WRC per week
		3							Uncontrolled = −.05 WRC per week Controlled = .50 WRC per week
		4							Uncontrolled = .48 WRC per week Controlled = .24 WRC per week
		5							Uncontrolled = .42 WRC per week Controlled = 1.02 WRC per week
Compton, Appleton, & Hosp (2004)	248	2	AA &LA	Reading aloud	1	WRC	Percentage of high-frequency words and decodable words accounted for 41% of the variance in reading accuracy and 54% of the variance in reading fluency.		

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
Reliability of Growth Rates									
Hintze, Owen, Shapiro, & Daly (2000) Study 1	160	2–5	GE	Reading aloud	1	WRC		Individual differences accounted for 48% of the variance and grade accounted for 19% of the variance.	(8 weeks, 2 times per week)
Study 2	80	1–4	ND	Reading aloud	1	WRC		Individual differences accounted for 42% of the variance, and grade accounted for 36% of the variance.	(10 weeks, 2 times per week)
Brown-Chidsey, Davis, & Maya (2003)	476	5–8	ND	Maze selection	10	CS		Grade level accounted for 68%–71% of the variance on two passages. Individual differences in scores accounted for 84% of the variance on one passage.	
Hintze & Pelle Petitte (2001)	12	3–4	GE & SE	Reading aloud	1	WRC		Individual differences accounted for 62% of the variance, and reading group accounted for 15% of the variance.	(8 weeks, 2 times per week)
Marston & Deno (1982, IRLD #106)	26	3	GE	Reading aloud	1	WRC			Paired *t* test at Weeks 1 and 16 showed significant growth on reading aloud; reading aloud showed greater growth when compared to basal series tests.

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
Marston, Deno, & Tindal (1983, IRLD #126)	83	3–6	LA	Word identification	1	WRC			Paired *t* test at Weeks 1 and 10 showed significant growth for all word lists ($p < .001$).
Skiba, Deno, Marston, & Wesson (1986)	67	1–7	SE	Reading aloud	1	WRC			(6 months, 3–5 times per week)
		1							1.78 WRC per week, *SD* = 1.21
		2							1.62 WRC per week, *SD* = .71
		3							1.42 WRC per week, *SD* = .92
		4							1.36 WRC per week, *SD* = .66
		5							
		6							
		7							
Shin, Deno, & Espin (2000)	43	2	GE/ Ch. 1	Maze selection	3	CS		Alternate form .75–.90	1.20 CS per month .91 CS per month (9 months, monthly)
Speece & Ritchey (2005)	276	1	HA/AA & LA	Reading aloud	1	WRC			(20 weeks, weekly & monthly)
			HA/AA						1.5 WRC per week
			LA						.77 WRC per week
Standard Rates of Growth									
Marston, Lowry, Deno, & Mirkin (1981, IRLD #49)	58	1–6	GE						Average % of growth
		1		Word identification	1	WRC			251
		2							84
		3							25

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
		4							9
		5							14
		6							21
		1		Reading aloud	1	WRC			150
		2							75
		3							26
		4							24
		5							28
		6							30
MacMillan (2000)	1,691	2–7	GE	Reading aloud	1	WRC			Significant growth within grades. Growth rates decreased as grade increased. Significant differences were found by gender (female > male). (1 school year, 3 times a year)
Fuchs, Fuchs, Hamlett, Walz, & Germann (1993)									(1 school year, weekly)
Study 1	117	1	ND	Reading aloud	1	WRC			2.10 WRC per week, *SD* = .80
		2							1.46 WRC per week, *SD* = .69
		3							1.08 WRC per week, SD = .52
		4							0.84 WRC per week, *SD* = .30
		5							0.49 WRC per week, *SD* = .28
		6							0.32 WRC per week, *SD* =.33
Study 2	257	1	ND	Maze selection	2.5	CS			0.34 CS per week, *SD* = .39
		2							0.39 CS per week, *SD* = .24
		3							0.47 CS per week, *SD* = .37

(table continues)

(Table 1 continued)

Study	Sample			Reading measure			Results		
	N	Grade(s)	Level	Type of measure	Time (min)	Scoring procedure	Validity	Reliability	Growth
		4							0.38 CS per week, *SD* = .32
		5							0.36 CS per week, *SD* = .23
		6							0.27 CS per week, *SD* = .25
Deno, Fuchs, Marston, & Shin (2001)	2,999	1–6	GE & SE	Reading aloud	1	WRC			(1 school year, weekly or 3 times a year)
		1	GE						1.80 WRC per week, *SE* = .15
			SE						0.83 WRC per week, *SE* = .15
		2	GE						1.66 WRC per week, *SE* = .09
			SE						0.57 WRC per week, *SE* = .09
		3	GE						1.18 WRC per week, *SE* = .10
			SE						0.58 WRC per week, *SE* = .11
		4	GE						1.01 WRC per week, *SE* = .05
			SE						0.58 WRC per week, *SE* = .05
		5	GE						0.58 WRC per week, *SE* = .05
			SE						0.58 WRC per week, *SE* = .08
		6	GE						0.66 WRC per week, *SE* = .04
			SE						0.62 WRC per week, *SE* = .14

Note. AA = average achieving; Ch. 1 = Chapter 1; CM = correct matches; CS = correct selections; D/HH = deaf or hard of hearing; ELL = English language learners; GE = general education; HA = high achieving; LA = low achieving; ND = sample included general education and special education, analysis not differentiated by group; *SD* = standard deviation; *SE* = standard error (of measurement); SE = special education; VI = visual impairment; WRC = words read correctly. If no specific numbers were reported for students receiving special education services, the participants were assumed to be in general education. **Tests:** *Stanford Achievement Test–7* (Gardner et al., 1982a, 1983); *Woodcock Reading Mastery Test Revised* (Woodcock, 1987); *Kaufman Brief Intelligence Test* (Kaufman & Kaufman, 1990); *Kaufman Test of Educational Achievement* (Kaufman & Kaufman, 1985); Gates-MacGinitie (MacGinitie et al., 1978); *Metropolitan Achievement Test–6* (Prescott et al., 1984); *Woodcock-Johnson–III* (McGrew & Woodcock, 2001); *Iowa Test of Basic Skills* (Hoover et al., 1996); *Woodcock-Johnson–Revised* (Woodcock & Johnson, 1989); *Comprehensive Reading Assessment Battery* (Fuchs, Fuchs, & Hamlett, 1989); *Tests of Achievement and Proficiency* (Scannell et al., 1986); *Michigan Educational Assessment Program* (Michigan State Board of Education, 1999); *Metropolitan Achievement Test–7* (Harcourt Educational Measurement, 1998); *Stanford Diagnostic Reading Test* (Karlsen & Gardner, 1985); *Minnesota Comprehensive Assessment* (Minnesota Department of Children Families, and Learning, 1998-2002); *Stanford Achievement Test–9* (Harcourt Brace & Company, 1997); *Diagnostic Reading Scales* (Spache, 1981); *The Test of Early Reading Ability—Deaf and Hard of Hearing Version* (Reid et al., 1991); *Oregon Statewide Assessment* (Oregon Department of Education, 2000); *Washington Assessment of Student Learning* (1998); *The Degrees of Reading Power Test* (Koslin et al., 1989).

by examining evidence related to the technical adequacy of the various measures used in CBM reading and consider generalizability of the research to other students and purposes.

Technical Adequacy of CBM Measures

CBM Reading Measures

We focus our review on the three measures most commonly used in CBM reading: reading aloud, maze selection, and word identification. In reading aloud, students read aloud from a passage, usually for 1 min, and the number of words read correctly is scored (Deno, 1985). Omissions, insertions, substitutions, hesitations, and mispronunciations are marked as errors. In maze selection, students read through a passage in which (typically) every seventh word has been deleted and replaced with three word choices—one correct choice and two distracters. (Rules for creating maze passages can be found in D. Fuchs and L. S. Fuchs, 1992.) Students read the passage silently, usually for 1 to 3 min, making selections as they read. The number of correct selections is scored. In word identification, students read aloud from a list of high-frequency words, usually for 1 min, and the number of words read correctly is scored (Deno, Mirkin, & Chiang, 1982). Omissions, insertions, substitutions, hesitations, and mispronunciations are marked as errors. Words are usually selected from word lists or from the reading curriculum. By far, the majority of research has focused on the reading-aloud measure. Recently, however, interest in maze selection and word identification has grown as CBM has been extended to younger and older students and as computerized progress monitoring has become a possibility.

Reading Aloud

Despite the early support for the technical adequacy of the reading-aloud measures, practitioners and researchers alike continued to express doubts about the relation between the simple measure of reading aloud from text for 1 minute and reading proficiency, especially proficiency in reading comprehension (e.g., see Mehrens & Clarizio, 1993; Yell, 1992). Thus, in the 1980s and 1990s, efforts were made to more closely examine the nature of the relationship between reading aloud and general reading proficiency, especially reading comprehension. This research took two different approaches. The first approach sought to clarify the relation between reading aloud and reading comprehension by considering alternative measures that might be more closely linked to reading comprehension and by examining the theoretical underpinnings of the relation between reading aloud and reading proficiency. The second approach sought to examine the concomitant relation between CBM reading aloud and reading comprehension, with a focus on the individual student.

Clarification of the Relation Between Reading Aloud and Reading Comprehension. Fuchs, Fuchs, and Maxwell (1988) compared the validity of CBM reading-aloud measures to that of other measures typically used to assess reading comprehension, including cloze (where every seventh word is deleted from a text and replaced with a blank), story retell, and question-answering measures. Participants were students with mild disabilities in Grades 4 to 8. Results revealed that reading-aloud scores correlated more strongly with scores on the comprehension and word skills subtests of a standardized achievement test (r = .91 and r = .80, respectively) than did scores from the other "typical" comprehension measures (rs = .76 to .82 for the reading comprehension and .66 to .76 for the word skills subtests, respectively). Results of the Fuchs et al. study suggested that reading aloud was more than just a measure of fluent decoding, a notion that was supported in subsequent research investigating the theoretical nature of the relationship between reading aloud and reading comprehension.

Shinn, Good, Knutson, Tilly, and Collins (1992) used confirmatory factor analysis to examine the role of reading aloud as it related to decoding, fluency, and comprehension skills for students in Grades 3 and 5. A single-factor model of "reading competence" was validated for third-graders, with all reading skills making significant contributions. In contrast, a two-factor model including decoding and comprehension as two separate but highly related factors was validated for fifth graders, with reading aloud loading on the decoding factor. Hosp and Fuchs (2005) also observed changes in the nature of the relationship between reading aloud and reading proficiency associated with age. Relationships between CBM reading aloud and the Decoding, Word Reading, and Comprehension subtests of the *Woodcock Reading Mastery Test–Revised* (WRMT; Woodcock, 1987) were similar in magnitude for students in Grades 2 and 3 (ranging from .82 to .88), but in Grade 4, lower correlations were observed for the Decoding and Word Reading subtests (rs = .72 and .73, respectively) than for the Reading Comprehension subtest (r = .82).

Kranzler, Brownell, and Miller (1998), in a somewhat different approach, posed the hypothesis that the number of words read aloud from text in 1 min might merely be a reflection of general speed of processing. Kranzler et al. examined the roles of general cognitive ability, speed and efficiency of elemental cognitive processing, and reading aloud in the prediction of reading comprehension for students in Grade 4. Multiple regression analyses revealed a significant relationship between reading aloud and reading comprehension measures that could not be explained by general cognitive ability or speed and efficiency of elemental cognitive processing. Results suggested that reading aloud was not merely an indicator of general cognitive processing speed. Kranzler et al. noted, however, that although reading aloud had the highest standardized regression coefficient when compared to cognitive ability and mental speed, it explained only 11% of the unique variance found in reading comprehension.

Concomitant Change in Reading Aloud and Reading Comprehension. The studies reviewed in the previous section focused on patterns of results across groups; however, the concerns raised by practitioners often focus on the nature of the relationship between CBM and reading comprehension for the individual student. One such concern is whether reading aloud and reading comprehension change concomitantly. Markell and Deno (1997) addressed this issue by experimentally manipulating the difficulty level of reading material. Students in Grade 3 read passages that were two levels below, at, and two levels above grade level. Students also completed two comprehension tasks for each passage—question answering and maze selection. Results revealed that, on average, students read significantly fewer words in 1 min on the more difficult passages, answered fewer questions correct, and selected fewer correct maze choices, supporting the general relation between words read aloud in 1 min and reading comprehension. However, at the individual level, it appeared that the amount of performance change was an important factor to be considered. Results revealed that for only 52% of the students did the rank ordering of reading-aloud scores match the rank ordering of comprehension scores on all three levels of materials. Taking a more liberal approach, and controlling for a ceiling effect on the comprehension measures, the percentage of students for whom rankings matched on only the highest and lowest levels were examined. This analysis resulted in an agreement of 100% and 96% for question answering and maze, respectively. These results suggested that a relatively large change in the number of words read in 1 min (perhaps as large as 15–20 words) was needed to predict with certainty a concomitant change in reading comprehension.

A second concern raised by practitioners is the existence of "word callers," that is, students who can read fluently but do not comprehend. Hamilton and Shinn (2003) examined teachers' ability to identify word callers. Third-grade teachers were asked to identify one to two students who were word callers (WC) and one to two similarly fluent peers (SFP). Similarly fluent peers were students whom teachers judged as having fluency rates similar to the WC students but with higher levels of comprehension. Results confirmed differences in comprehension levels between the students in the WC and SFP groups, with SFP students scoring higher on comprehension tasks. However, results also revealed differences in reading fluency, with scores for WC students lower than those for SFP students, calling into question teachers' ability to identify students as word callers. We note that the Hamilton and Shinn (2003) study did not address the actual existence of word callers, just teachers' judgment of word callers. To examine the existence of word callers, one would need to examine the relative standing of students on reading aloud and reading comprehension measures. Word callers would be those whose relative standing on the reading aloud measures were substantially higher than their standing on the reading comprehension measures.

Maze Selection

Although the number of words read aloud in 1 min demonstrated good technical adequacy in the early IRLD research, the measure was limited in the sense that it had to be administered individually, it lacked face validity, and it was unclear whether the measure would be appropriate for older students who might presumably reach an asymptote in reading aloud performance. These factors, combined with improvements in technology and changes in the field of special education leading to larger caseloads, led to consideration of the maze measure. The maze could be administered in groups, appeared to be more of a reading comprehension measure than a reading aloud measure, could be administered via the computer, and was considered to be more acceptable for older students. The maze was not a new measure. An untimed version of the maze had been studied in 1970s by Guthrie as a measure of reading comprehension and was shown to have good stability, to correlate with standardized measures of reading proficiency, and to separate readers with and without disabilities (Guthrie, 1973; Guthrie, Seifert, Burnham, & Caplan, 1974). The use of the maze as a timed measure within a CBM framework did not appear until the late 1980s and early 1990s.

In 1989, Espin, Deno, Maruyama, and Cohen reported on the technical adequacy of a maze measure that was part of a group-administered screening instrument called the *Basic Academic Skills Samples* (BASS; Deno, Maruyama, Espin, & Cohen, 1989). The reading portion of the BASS consisted of three 1-min maze selection tasks that were approximately at a first- to second-grade reading level. The BASS was administered to more than 2,000 students in Grades 1 through 6 across 31 schools. Correlations between the BASS maze and 1-min reading-aloud passages for a random sample of students from Grades 3, 4, and 5 were .77, .86, and .86, respectively. Data from the entire sample revealed a stable pattern of increase in maze scores from Grades 1 to 6, as well as from winter to spring within each grade.

Fuchs and Fuchs (1992) extended the research on maze selection in their search for a CBM reading measure that would be suitable for data collection via the computer and that might have greater acceptance for teachers than would reading aloud. Technical adequacy and level of teacher acceptance were compared for several alternative CBM measures, including question answering, story recall, cloze, and maze selection. Maze selection in this study was a 2.5-min measure administered twice weekly for 18 weeks via computer. Earlier research (Fuchs & Fuchs, 1990) had revealed correlations of .83 between scores on maze and reading aloud and of .77 between scores on maze and the Reading Comprehension subtest of the *Stanford Achievement Test* (SAT; Gardner, Rudman, Karlsen, & Merwin, 1982). Results of the Fuchs and Fuchs (1992) study revealed that the maze task was sensitive to change in performance over time and, unlike other measures, had a relatively small ratio of slope to standard error of estimate (SEE),

making it easier to detect growth on a graph. In addition, teachers rated their satisfaction with maze highly, reporting that they believed the maze reflected multiple dimensions of reading, including decoding, comprehension, and fluency. Finally, students reported that they liked taking the maze.

In a direct comparison of the technical adequacy of reading aloud and maze selection measures, Jenkins and Jewell (1993) examined the validity of the two measures across Grades 2 to 6. All students in the study completed three 1-min maze tasks and three 1-min reading-aloud passages. Passages were at a first- to second-grade level. Criterion variables were scores on the *Gates-MacGinitie Reading Tests* (Gates; MacGinitie, Kamons, Kowalski, MacGinitie, & McKay, 1978) and the *Metropolitan Achievement Tests* (MAT; Prescott, Balow, Hogan, & Farr, 1984). Within-grade correlations were moderate-strong to strong for both measures, ranging from .63 to .88 with the Gates and from .58 to .87 with the MAT. In Grades 2 through 4, correlations tended to be stronger for reading aloud than for maze, but in Grades 5 and 6, this pattern of differences disappeared. Looking across grades, correlations between the reading aloud and the criterion measures dropped from the .80s in Grades 2 through 4 to .60s to .70s in Grades 5 and 6. In contrast, correlations for the maze remained consistent across the grade levels, with most between .65 and .75. Finally, both measures revealed increases across Grades 2 to 6, and from fall to spring within grade. For the reading aloud measures, change was greatest from Grades 2 to 3, after which it leveled off. Maze, in contrast, reflected more even rates of change across the grades. Results suggested that reading aloud might be a better measure than maze for primary-grade students, a conclusion supported by Ardoin et al. (2004), who found that adding a maze task did not add significantly to the prediction of performance on a standardized achievement test in reading for students in third grade.

Word Identification

Although both word identification (word ID) and reading aloud were included as potential indicators of reading proficiency in early CBM research (e.g., Marston, Deno, & Tindal, 1983; Marston, Lowry, Deno, & Mirkin, 1981; Shinn, Ysseldyke, Deno, & Tindal, 1986; Tindal, Marston, Deno, & Germann, 1982), reading aloud emerged as the more commonly used measure, perhaps because it appeared to be more closely related to the construct of reading than did word ID. Reading aloud, however, proved difficult to use with beginning first-graders because many of these students could not read any words from text in the fall, creating a floor effect in the measure (e.g., see Bain & Garlock, 1992; Fuchs, Fuchs, & Compton, 2004). As interest in early identification and prevention grew, interest in CBM word identification reemerged: The words presented in a word ID task could be controlled for difficulty and were not constrained by the requirement of fitting the words into a coherent story.

The technical adequacy of several potential CBM reading measures for students in first grade was compared in a study by Daly, Wright, Kelly, and Martens (1997). Reading measures included word ID, letter reading, letter copying, letter-sound production, and letter-sound selection. Word ID probes were created using words selected from the Harris-Jacobson pre-primer word list (Harris & Jacobson, 1972). Results revealed that letter reading and word ID produced the best technical adequacy data. Test–retest reliabilities for the word ID and letter-reading measures were .94 and .87, respectively, compared to .42 to .65 for the other measures. Concurrent validity coefficients with the broad reading subtest of the *Woodcock-Johnson–Revised* (Woodcock & Johnson, 1989) were .40 and .35, respectively. Predictive validity coefficients between the two measures and a passage-reading and word-ID task administered 4 months later were .73 (word ID) and .71 (letter reading) with passage reading and .71 (word ID) and .69 (letter reading) with word ID. Predictive validity coefficients for the other measures ranged from –.09 to .53.

Fuchs, Fuchs, and Compton (2004) compared the validity of a word-ID task and nonsense word fluency (NWF) task for first graders at risk in reading. The word-ID task was created using words from the Dolch preprimer, primer, and first-grade-level lists. The NWF measure was taken from DIBELS (Good, Simmons, & Kame'enui, 2001). Criterion measures included the word attack and word identification subtests of the *Woodcock Reading Mastery Test–Revised* (WRMT-R; Woodcock, 1987) and the *Comprehensive Reading Assessment Battery* (CRAB; Fuchs, Fuchs, & Hamlett, 1989). Students were tested in the fall and spring of the year on the criterion measures and were tested at least once weekly on both CBM measures. Alternate-form reliability for the word-ID and NWF tasks was .88 and .87, respectively. Validity was examined for both level and slope of performance for the two CBM measures. In general, correlations with the criterion measures were consistently and reliably stronger for word ID than for NWF. Concurrent validity coefficients for word-ID level ranged from .52 to .93, compared to .50 to .80 for NWF. Predictive validity for word-ID level ranged from .45 to .80, compared to .46 to .64 for NWF. Finally, the slopes produced by word ID were more strongly correlated to the criterion variables (with most coefficients at .45 or above) than they were for NWF (with only 2 of 12 coefficients at .45 or above). Results not only lent support for the concurrent and predictive validity of word ID as an indicator of performance but also provided information supporting the technical adequacy of the growth rates produced by the measure.

In a recent study, Compton, Fuchs, Fuchs, and Bryant (2006) examined the use of a word ID measure within a response-to-intervention (RTI) approach for first-grade students. Students were administered a prediction battery in the fall of first grade consisting of word ID, rapid naming, phonemic awareness, and oral vocabulary measures. Students were also progress-monitored for a 5-week period using the word-

ID task, and the slope of improvement was calculated. Students were followed until the end of second grade. Compton et al. examined the use of word-ID level, slope, and a combination of level and slope as a part of a process for predicting performance of participants at the end of second grade. (They also examined different classification and data analytic approaches, which are not reviewed here.) Results revealed that adding word-ID level and slope significantly improved classification accuracy for the identification of at-risk students over and above the use of phonemic awareness, rapid naming, and oral vocabulary measures.

Generalizability of Research: Student Populations and Purposes

Recent work has examined the validity of CBM reading measures—primarily reading aloud—to diverse groups of students and for new uses. In terms of generalizability to new student populations, CBM research has been extended to students at the secondary-school level and to students with diverse backgrounds and characteristics (racial/ethnic and language backgrounds, gender, socioeconomic status, and sensory disabilities). In terms of generalizability of uses, CBM research has been extended to examine the uses of reading measures for predicting performance on state standard tests. Given space limitations, and the fairly limited amount of research within each category, we briefly review these studies in this section.

Student Populations. Extensions of CBM reading research to the secondary-school level initially focused on the use of reading measures to predict content-area performance, rather than to predict general reading performance (e.g., Espin & Deno, 1993a, 1993b; Espin & Deno, 1995; Fewster & MacMillan, 2002; Yovanoff, Duesbery, Alonzo, & Tindal, 2005). More recent research has focused on predicting general reading performance (Espin & Foegen, 1996; Espin, Wallace, Lembke, Campbell, & Long, 2007; Muyskens & Marston, 2006; Tichá, Espin, & Wayman, 2007). These studies have all been conducted at the middle school level. Results have generally shown that both reading aloud and maze exhibit strong alternate-form reliability and moderate to strong criterion-related and predictive validity. However, Espin et al. (2007) and Tichá et al. (2007) also found that whereas reading aloud did not reflect change in performance over time, maze selection did. Growth on maze was related to performance on a state reading test and to changes on a standardized achievement test in reading.

Extensions of CBM reading research to students from diverse backgrounds have produced mixed results. With regard to racial/ethnic backgrounds, Hixson and McGlinchey (2004) and Kranzler, Miller, and Jordan (1999) found that CBM reading aloud resulted in overestimation of reading performance for African American students and underestimation of performance for Caucasian students. (Overestimation of performance might result in underidentification for services.) However, Hintze, Callahan, Matthews, Williams, and Tobin (2002) found that CBM reading aloud resulted in neither overnor underestimation of performance for African American or Caucasian students. In terms of language, results generally have revealed moderate to strong reliability and criterion-related validity coefficients for reading-aloud measures with English learners (ELs; Baker & Good, 1995; Wiley & Deno, 2005). In addition, gains on reading-aloud measures for EL students have been found to be similar to gains seen for non–EL students (Graves, Plasencia-Peinado, Deno, & Johnson, 2005).

In a comprehensive study of the effects of home language, gender, ethnicity, and socioeconomic status on the technical adequacy of CBM reading aloud measures, Klein and Jimerson (2005) followed three cohorts of students ($N = 398$) longitudinally from Grades 1 through 3. The first cohort consisted of Caucasian students who spoke English as their home language, the second group was Hispanic students who spoke English as their home language, and the third cohort was Hispanic students who spoke Spanish as their home language. Results revealed a strong relationship between the CBM reading-aloud and SAT reading scores for all three groups of students at each grade level, with most correlations between .63 and .82. Linear regression analyses revealed that only a combination of ethnicity and home language resulted in bias in the measures. Specifically, the reading proficiency of Hispanic students whose home language was Spanish was systematically overpredicted (which could lead to systematic underidentification for services), whereas the reading proficiency for Caucasian students whose home language was English was systematically underpredicted (which could lead to systematic overidentification).

Only two studies have examined the validity of CBM reading measures with students with sensory disabilities. The first, Morgan and Bradley-Johnson (1995), provided support for the validity of the measures for students in Grades 3 to 7 with visual impairments. The second, Allinder and Eccarius (1999), did not provide support for the validity of either reading aloud or maze selection for students who were deaf and hard of hearing.

Purposes. Recent work has extended the work on CBM reading measures to examine their use for predicting performance on state standards tests. Early studies focused on establishing benchmark scores that would predict passing or failing a state reading test (Crawford, Tindal, & Stieber, 2001; Good, Simmons, & Kame'enui, 2001). Subsequent studies that examined correlations between CBM reading-aloud measures and performance on state standards tests reported diagnostic efficiency statistics, including sensitivity, specificity, positive predictive power, and negative predictive power (Hintze & Silberglitt, 2005; McGlinchey & Hixson, 2004; Silberglitt & Hintze, 2005; Stage & Jacobsen, 2001; see Note 2). Sensitiv-

ity is the percentage of students below a cut score who fail a test. Specificity is the percentage of students above a cut score who pass a test. Positive predictive power is the probability that a student with a score below the cut score will truly fail the test, whereas negative predictive power is the probability that a student with a score above the cut score will truly pass the test.

Correlations between scores on the reading-aloud measures and the various state reading tests generally ranged from .60 to .80 across studies. The exception to this pattern was the study by Stage and Jacobsen (2001), in which correlations between reading aloud and the Washington state test were .43 to .44. The differences for the Stage and Jacobsen study might have been related to the nature of the Washington state reading test, which required both short-answer and extended written responses, and thus involved not only reading but also writing. The other state tests did not require written responses.

Diagnostic efficiency statistics across the four studies were fairly consistent. Sensitivity values ranged from 65% to 76%; specificity values ranged from 74% to 82%. Positive predictive power values ranged from 55% to 77% for all but the Stage and Jacobsen (2001) study, in which the value was 41%. Negative predictive power values ranged from 83% to 90% for all but the McGlinchey and Hixson (2004) study, in which the value was 46%. Across studies, the use of CBM added significantly to positive and negative predictive power above base rates of prediction.

Summary and Discussion

Research on CBM reading measures has provided support for and clarified the nature of the relationship between reading aloud and general reading proficiency; has examined alternates to reading aloud, including maze selection and word ID; and has examined the generalizability of the results to different student populations and for different uses.

With regard to the reading-aloud measure, results generally replicated earlier research demonstrating a strong relationship between CBM reading aloud and reading proficiency, even when correlations were calculated within grade, addressing a concern raised about the early CBM research (see Mehrens & Clarizio, 1993). Reading aloud was found to be a better indicator of reading comprehension than were other "typical" comprehension measures, and results revealed that reading aloud was not just a speed-of-processing measure. In addition, research provided insight into the theoretical nature of the relationship between reading aloud and reading proficiency for elementary-school students.

However, reading aloud has limitations. First, reading aloud may not be the best choice for very young and older students. For readers at the very beginning stages of reading, reading-aloud measures produced a floor effect. Examination of an alternative measure, word ID, proved a promising alternative for very beginning readers. Reliability and validity coefficients for word ID were consistently strong for beginning readers, and research supported the use of word ID as a part of an RTI approach to early identification and prevention. Second, although correlations between reading-aloud and criterion measures remained moderate to strong across elementary school grades, they were strongest at the primary grades and decreased at the intermediate grades. No such decrease was seen for maze, which remained fairly stable across the grades. Thus, although reading aloud might be the best measure for primary-grade students, reading aloud and maze selection both seem to be appropriate for intermediate-grade students. For secondary-school students, maze may be the best choice. Although research is limited, initial results suggest that reading aloud does not reflect growth for middle school students, whereas maze selection does. Finally, if progress is to be monitored across school years, maze might prove to be the best choice. It has been shown to have reasonable validity and reliability for students across Grades 2 through 8, and the growth rates across grades have shown greater consistency than those for reading aloud. The reasons for the age-related differences seen between the measures are not clear and should be examined more closely. Perhaps the teachers from the Fuchs and Fuchs (1992) study were correct: Perhaps maze reflects multiple aspects of reading proficiency to a greater extent than reading aloud does. If true, the separation of reading proficiency into decoding and comprehension factors at the intermediate elementary grades, as seen in Shinn et al. (1992), would not affect the maze correlations but would affect reading-aloud correlations. Regardless of the reasons for the differences, given the moderate to strong reliability and criterion-related validity coefficients for maze, and given the advantages offered by maze in terms of group administration, appropriateness for computerized administration, potential for cross-grade measurement, and acceptance by teachers, we believe that in the future more attention should be devoted to maze selection as a potential CBM reading measure.

The extension of the research to various populations and for different uses is still in the early stages of development. Research at the secondary-school level is promising but has focused primarily on middle school students. Little has been done at the high school level. With regard to students of diverse backgrounds, evidence suggests that although the CBM reading measures may have reasonable reliability and criterion-related validity for various groups, the measures may overestimate the performance of African American students and of Hispanic students whose home language is Spanish and underestimate the performance of Caucasian students. However, results are mixed, and there is a need to further examine these patterns of relations. We suggest that performance-level differences between groups be factored out in future research.

In terms of extensions for new purposes, the use of CBM reading-aloud measures for predicting performance on state standards tests has produced positive results, with the measures producing generally strong correlations with state read-

ing tests and fairly good sensitivity, specificity, positive predictive, and negative predictive power. These conclusions must be tempered by the fact that the technical adequacy of CBM reading measures for predicting performance on state tests may be affected by the nature of the state test, as seen in Stage and Jacobsen (2001).

Aside from a need for further research on the generalizability of the measures, two other issues have yet to be sufficiently addressed in the area of measure development. First, there is a need to further examine the relationship between reading aloud and reading comprehension at the individual level. Few studies focused on the individual level. One that did, Markell and Deno (1997), implied that large gains in reading-aloud scores might be necessary before gains in reading comprehension could be assumed. There is a need to replicate this research and to examine the implications of the results for individual progress monitoring. Relatedly, we think research on the existence of word callers is of interest. It is conceivable that a small group of students exists whose performance on the reading aloud measures is, relatively speaking, much higher than their performance on comprehension measures.

A second issue relates to methods for linking measures. If word ID is to be used for beginning readers, reading aloud for primary-grade readers, and maze for intermediate and secondary-school readers, how might the measures be linked to create a picture of growth across school years? The ability to link different measures across years would contribute to the development of a seamless and flexible system of progress monitoring.

Effects of Text Materials

Thus far, we have focused on various measures that can be used in CBM in reading. In this section, we turn our attention to the materials used to develop those measures. In 1994, Fuchs and Deno raised the question of whether instructionally useful performance assessment had to be based in the curriculum. The authors noted that although measures selected from the curriculum might have face validity, curriculum measures also introduced more error into progress measurement because of variations in passage difficulty and student familiarity with passages. In addition, passages selected from within an instructional curriculum might limit generalizability to other reading curricula.

A large number of studies were conducted beginning in the 1990s to address questions related to the materials used to develop CBM measures. The research fell into two broad categories. The first addressed the question of curriculum source, that is, whether technical adequacy of the measures differed with the curriculum used to generate measures. Included in this category were studies comparing reading passages generated from different curricula and studies comparing passages generated from an instructional versus a "generic" or noninstructional curriculum. The second category of research addressed the difficulty level of the materials chosen, specifically whether measurement had to be done within the student's instructional level.

We note two points before describing the literature in this section. First, the research on materials has been conducted almost exclusively with reading-aloud measures. Few studies have examined maze or word ID measures. Second, there are a surprisingly large number of studies addressing the question of materials. Given the space limitations accredited to a review article, and the similarities in pattern of results, we report the results generally, referring to details of specific studies only when needed.

Curriculum Effects

Three general themes emerged from the research on the effects of curriculum on CBM reading measures. First, level of performance differs significantly with curriculum source. Second, although technical adequacy does not vary with curriculum source, rates of growth may. Third, it is not necessary to match instructional and progress-monitoring material.

With respect to the first theme—differences in levels of performance—results consistently reveal mean level differences in scores on passages drawn from different curricula, beginning with Tindal, Marston, Deno, and Germann (1982). Studies have shown higher levels of performance on instructional materials than on mainstream materials (Tindal, Flick, & Cole, 1992), on literature-based materials than on authentic materials (Hintze, Shapiro, Conte, & Basile, 1997), on basal materials than on literature-based materials (Bradley-Klug, Shapiro, Lutz, & DuPaul, 1998), and on generic materials than on basal materials (Powell-Smith & Bradley-Klug, 2001). In addition, higher levels of performance have been found on maze measures drawn from materials controlled for difficulty than on literature-based materials (Brown-Chidsey, Johnson, & Fernstrom, 2005). Mean level differences are important for progress monitoring only insofar as they are used to compare students across classes, schools, or districts or when comparing student performance at one point in time to performance at a later point in time. For such uses, it is important to keep the source of material constant. However, performance-level differences do not address the issue of technical adequacy of the measures as indicators of reading performance or growth. Measures may produce differences in levels of performance but be equally good indicators of general reading proficiency.

The second theme to emerge from the curriculum materials research does relate to technical adequacy. Results across several studies reveal that there are few differences in the technical adequacy of reading-aloud measures selected from different curricula. For example, Fuchs and Deno (1992) compared reading-aloud passages drawn from two published basal curriculum series and found no differences in the magnitude of correlations with the *Woodcock Reading Mastery Test* (WRMT; Woodcock, 1973) in Grades 1 through 6. Similarly, Hintze, Shapiro, Conte, and Basile (1997) found no differ-

ences in alternate-form reliability or criterion-related validity for passages selected from authentic and literature-based curricula and no differences in the passages for classifying students into groups. Hartman and Fuller (1997) found that test–retest reliability and criterion-related validity coefficients for passages selected from literature-based materials for students in Grades 1 through 3 were similar to those reported in the other research for passages selected from basal-series texts. Finally, Brown-Chidsey et al. (2005) found high correlations between performances on maze passages selected from controlled and literature-based material.

Similar to the research on performance levels, research on the developmental growth rates (i.e., changes in scores across grade levels) produced by CBM reading measures reveal few differences related to curriculum source (Fuchs & Deno, 1992; Hintze et al., 1997). However, results from studies of intraindividual growth rates have been mixed. Bradley-Klug, Shapiro, Lutz, and DuPaul (1998) found no differences in growth rates produced by literature-based and basal-series curricula for second-graders. Differences in growth for fifth-graders were not statistically significant, but growth rates were .84 words per week for literature-based and .32 for basal-series probes. Brown-Chidsey et al. (2005) found that maze probes created from controlled and literature-based passages both produced positive growth rates from fall to winter to spring measurements. The significance of the difference in growth rates was not tested.

Other studies have found significant differences in growth rates. Hintze, Shapiro, and Lutz (1994) examined the growth rates produced by literature-based and basal-series curricula for two groups of third-grade students who were taught with either a literature-based or basal-series curriculum. Growth rates for the literature-based and basal-series students were −.35 and −1 when measured with literature-based probes, but were .66 and 1.70 when measured with basal-series probes. In the Hintze et al. (1994) study, passages were not equated for difficulty. Passage difficulty was equated in a subsequent study (Hintze & Shapiro, 1997) in which students in Grades 2 to 5 who were instructed with literature-based or basal-series material were monitored with literature-based and basal-series passages. Again, results revealed differences in growth rates related to curriculum source, although the results were inconsistent across grades. Students in Grades 2 through 4 achieved greater growth when monitored with the literature-based, rather than with the basal-series, probes (a pattern opposite that of Hintze et al. [1994], who found greater growth with literature-based probes than with basal-series probes for third-graders), whereas students in Grade 5 achieved greater growth with the basal-series probes (a pattern opposite that found by Bradley-Klug et al. [1998], who found greater growth rates on literature-based probes for fifth-grade students).

It is difficult to determine why there is such inconsistency in results regarding growth rates. However, given the general pattern of inconsistency in the results (i.e., literature-based sometimes producing higher and sometimes lower growth rates), we hypothesize that factors other than curriculum source may be contributing to differences in growth rates. For example, as will be discussed in a later section, it is quite difficult to determine the equivalency of "parallel probes" used to monitor progress, even when the probes are drawn from the same curriculum and matched on readability level. Moreover, slope values can easily be affected by a bunching of particularly difficult or easy probes near the beginning or end of a progress-monitoring session. One way to address these issues is to remove potential effects of nonequivalence of passages by counterbalancing the order of passages across students so that students do not read passages in the same order. Another method is to use techniques other than readability to establish equivalence of the passages (a point we will discuss later). In either case, based on current research, we cautiously suggest that growth rates are not affected by curriculum source; however, as with level of performance differences, we would still recommend consistency in progress monitoring with respect to curriculum source.

The third and last theme emerging from the research on materials is related to the need to match monitoring material to the materials used in instruction. The aforementioned Hintze et al. (1994) and Hintze and Shapiro (1997) studies resulted in no interaction between the growth rates generated by materials selected from particular curricula and the curricula used for instruction. In other words, students taught using a literature-based series did not grow differently on literature-based probes than on basal-series probes. Other studies have produced similar results. Tindal et al. (1992) found no differences in slope of improvements for 12 special education students in Grades 2 through 5 when monitored on highly controlled instructional materials or the general education basal curriculum. Powell-Smith and Bradley-Klug (2001) and Riley-Heller, Kelly-Vance, and Shriver (2005) found no differences in growth rates between probes derived from the students' instructional material and generic probes for second-grade students.

Difficulty Level

In terms of difficulty level of material, there are two general issues to consider. The first is whether students must be measured in instructional-level material or whether they can be measured in material outside their instructional level. The second is the importance of establishing equivalence of passage difficulty for repeated progress monitoring.

Need to Measure at Instructional Level. Addressing the issue of whether students must be measured with instructional-level material, Fuchs and Deno (1992) compared criterion-related validity and developmental growth rates produced by materials of various difficulty levels. Results revealed no differences related to material difficulty in the magnitude of the correlations. Average correlations across difficulty levels with the WRMT (Woodcock, 1973) were .91, ranging from

.89 to .93. Results also supported the use of a generic (in this case, third-grade) passage for tracking growth across grade levels. Growth rates produced with a common third-grade passage were relatively stable and linear. However, Fuchs and Deno also found that developmental growth rates decreased as the difficulty of the material increased; thus, sixth-grade material produced growth rates that were less steep than those produced with third-grade material.

Several studies have examined the influence of material difficulty on intraindividual growth rates. Shinn, Gleason, and Tindal (1989) monitored the reading-aloud progress of mildly handicapped students in Grades 3 to 8 who were randomly assigned to one of two groups: reading material one grade level below and above instructional placement, or two and four grade levels above instructional placement. Results revealed comparable slopes for students monitored one level below (4.3 words per week) and above (3.7 words per week) instructional placement level. Although not statistically significant, slopes for two and four levels above instructional placement did differ in magnitude (4.55 vs. 2.35 words per week), supporting the earlier finding that an increase in difficulty leads to flatter slopes. Standard error of estimates (SEEs) did not differ by difficulty level.

Dunn and Eckert (2002) found no differences in slopes or SEEs for grade-level versus challenging (approximately one grade level above instructional level) material for second- or third-grade students. Hintze, Daly, and Shapiro (1998), however, did find differences in the pattern of results for grade levels. Students in Grades 1 through 4 were monitored with material at and 1 year above grade level. No differences were seen in alternate-form reliability between materials, nor in the slopes obtained for students in Grades 3 and 4. However, for students in Grades 1 and 2, growth rates were higher on grade-level material than they were on material above grade level. Differences were especially notable for first-grade students. The authors surmised that students in the beginning stages of reading development experienced more fluency problems when the text was difficult than did students who were beyond the beginning stages of reading.

One study examined the effects of difficulty level on the slopes produced by word identification probes. Fuchs, Tindal, and Deno (1981) compared three types of word lists: those generated from grade-level material covered throughout the year (grade-level, comprehensive), from grade-level material covered during the time of the study (grade-level, limited), and across-grade-level material (preprimer to Grade 4). Differences were found in the magnitudes of the slopes produced by the measures. The steepest slopes were produced by the grade-level, limited word lists (.49), followed by the grade-level, comprehensive lists (.20), and the across-grade lists (−.07).

Equivalence of Passages. A second issue related to difficulty level of CBM materials, and one that has received far less attention than other materials-related issues have, is the importance of establishing the equivalence of the "parallel" passages used for repeated progress monitoring. As one might expect, CBM reading scores are sensitive to the difficulty level of the passages. For example, each of the studies described in the section above reported mean score differences on the reading-aloud measure for passages of varying difficulty levels (e.g., Dunn & Eckert, 2002; Fuchs & Deno, 1992; Hintze, et al., 1998; Shinn et al., 1989). In one respect, this is a positive finding and demonstrates the validity and sensitivity of the CBM reading-aloud measures. On the other hand, with respect to ongoing progress monitoring, the finding is problematic. It implies that scores on repeated progress-monitoring measures will be affected by variation in passage difficulty (a finding confirmed by studies reviewed in a later section on growth).

The importance of establishing equivalence in passage difficulty has been illustrated in two studies. In the first, Poncy, Skinner, and Axtell (2005) examined the effects of passage variability on reading-aloud scores. Participants were third-graders who read 20 grade-level passages with readabilities ranging from 2.8 to 3.1 grade level. Passages were presented to students in random order over a period of 4 days. Analyses revealed that 81% of the variance in students' scores was due to student skill, 10% was due to passage variability, and 9% was unaccounted for. By controlling the difficulty of the passages on the basis of students' average scores, variance due to student skill increased and variance to passage difficulty decreased.

In the second study, Hintze and Christ (2004) compared grade-level material controlled for difficulty level with randomly selected materials from graded readers for students in Grades 2 to 5. Students were administered both controlled and uncontrolled reading-aloud passages over the course of 11 weeks. The results indicated that estimates of both SEE and the standard error of the slope (SEb) were smaller when passages were controlled for difficulty than when they were not.

The results of Poncy et al. (2005) and of Hintze and Christ (2004) emphasize the need to establish passage equivalence. Yet, creating parallel passages is not as easy as it may seem. The most common method for establishing equivalence of CBM passages has been to examine the readability levels of the passages via the use of common readability formulas. However, Ardoin, Suldo, Witt, Aldrich, and McDonald (2005) found only modest relationships between the reading levels assigned to passages via readability formulas and the number of words read correctly (WRC) in 1 min from those passages for third-graders. The readability formula that produced the highest and most reliable mean associations with WRC was Forecast (Sticht, 1973), a formula that has not been used in the CBM research. In a component analysis, Ardoin et al. (2005) found that the two components significantly related to WRC were syllables per 100 words and words not on the Dale-Chall list of 3,000 words. Results also revealed inconsistencies in the levels assigned to passages among various readability formulas.

Others studies have demonstrated low correlations (ranging from rs = −.08 to .43) between scores on CBM reading-aloud

and readability formula levels (Bradley-Klug et al., 1998; Compton, Appleton, & Hosp, 2004; Hintze et al., 1994; Powell-Smith & Bradley-Klug, 2001) and low correlations among various readability formulas (r = .28; Compton et al., 2004). In addition, Compton et al. (2004) found that certain passage components were related to WRC in 1 min, including the number of high-frequency and decodable words, the number of multisyllabic words (negatively related), and sentence length.

Given the problems with readability formulas for determining passage difficulty, how can one determine the equivalence of passages used for CBM progress monitoring? Ardoin et al. (2005) proposed a process whereby a set of passages is selected based on specific components found to relate to reading fluency, such as those identified in Ardoin et al. (2005) and Compton et al. (2004). Selected passages are then field-tested with a large number of students. In this system, only passages that fall within 1 standard deviation of the mean WRC would be selected for use in CBM progress monitoring. The effects of such an approach on the stability of the growth rates produced by reading-aloud passages have yet to be examined.

Summary and Discussion

The research on curriculum source supports a robustness in the CBM measures: Even when developed from different curriculum sources, the measures seem to function consistently. Moreover, it does not appear necessary to develop CBM probes from material in which the student is being instructed. These conclusions are good news in terms of the development of a seamless and flexible system of progress monitoring. They imply that CBM progress monitoring using reading-aloud measures can be used across various curricula and instructional approaches. However, the research related to difficulty level poses some restrictions to the general robustness of CBM materials.

With respect to the issue of whether students need to be measured with instructional-level material, the literature indicates that CBM reading-aloud measures are fairly flexible with regard to difficulty level: Students can be measured with material that is easier or more difficult than their instructional level, and the technical adequacy of the measures is not affected. However, there are limits to this flexibility. Rates of growth may be affected if material is too difficult (e.g., two to three levels above instructional level). Further, there is some indication that beginning readers may be more affected by difficulty level than more advanced readers are. In terms of word identification, results reveal that the more closely the measure is tied to the instruction, the more sensitive it is to growth. Of course, this sensitivity must be balanced with generalizability of the results. If words are selected that cover only 1 month of instruction, students are likely to hit a ceiling in their performance in a relatively short amount of time.

The issue of difficulty as it relates to intraindividual growth monitoring is of greater concern. Generally, results emphasize the need to establish passage equivalence for CBM progress monitoring, especially if the measures are to be used as a part of a decision-making process that carries important social consequences. The issue of passage equivalence is perhaps less of a concern if CBM is used by a classroom teacher to monitor student progress and evaluate instructional programs. Given the time-consuming nature of the type of approach suggested by Ardoin et al. (2005), and the positive treatment validity results reported in Stecker et al. (2005), use of passages selected from controlled sources, such as basal-reading series, is probably sufficient for such classroom use. However, if CBM is to be used as part of a school- or districtwide decision-making process, or if it is to be used as part of an eligibility decision-making process, we feel that it is necessary to establish equivalence of passage difficulty for progress monitoring by adapting a process similar to that suggested by Ardoin et al. (2005).

Issues About Measuring Growth

We already have discussed in part some factors related to growth. For example, the amount of growth produced by CBM measures may vary with the curriculum source or difficulty level of the passages, and error in the production of growth rates is reduced when the difficulty of the passages used for progress monitoring is controlled. Two additional issues specific to growth have not yet been addressed: (a) How reliable or trustworthy are the rates of growth produced by CBM measures? (b) Can standards for growth be determined, and do they differ for students by performance or grade level? As we did in the section on the effects of text materials, we caution the reader that the majority of research conducted in this area has been with the reading-aloud measure, with little attention devoted to word ID or to maze selection.

Reliability of Growth Rates

The research on determining the trustworthiness of the growth rates produced by CBM measures includes examination of the dependability of single CBM scores and the reliability or accuracy of the slopes produced by multiple, repeated scores. With regard to single scores, studies have examined the amount and sources of error surrounding single CBM scores. Christ and Silberglitt (in press) calculated typical standard error of measurement (*SEM*) values across students in Grades 1 to 5, taking into consideration various levels of measurement reliability and sample variability. Values were calculated for data collected in fall, winter, and spring (three measurements per data-collection period) across a period of 8 years for 8,200 students. Results reveal that the median *SEM* across grades and conditions was 10 words read correctly, with a range of 5 to 15. Reliability, grade, and sample diversity affected *SEM*s, with smaller *SEM*s associated with higher levels of reliability, lower grade levels, and less sample variability. The au-

thors suggested that the use of standard *SEM*s could aid in the interpretation of CBM reading aloud data.

Sources of error associated with CBM scores have been examined recently via the application of G theory, or generalizability theory, to CBM reading aloud. G theory is designed to assess the dependability of behavioral measurements by specifying portions of error that can be accounted for by various situational variables under which measurements are taken (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Through G theory, "portions of variance that in classic test score theory are simply attributed to random error" (Hintze, Owen, Shapiro, & Daly, 2000, p. 53) are identified and explained. Results of a series of studies applying G theory have revealed consistent support for the dependability of the CBM reading measures for inter- and intraindividual decision making at the elementary school level, with the majority of variance in CBM scores accounted for by individual variation and grade level (Hintze et al., 2000) or individual variation and group membership (Brown-Chidsey, Davis, & Maya, 2003; Hintze & Pelle Petitte, 2001).

Although dependability of individual data points contributes to the dependability of the slope, it does not guarantee it. Other factors must be considered with regard to slope. Several approaches have been taken to evaluate the trustworthiness or the reliability of the slopes produced by CBM measures. In early CBM studies, the focus was on whether CBM reading measures were sensitive to change in performance. For example, Marston and Deno (1982) and Marston, Deno, and Tindal (1983) compared change in scores on reading-aloud and word ID measures to changes in scores for other reading measures, such as standardized achievement tests or basal-series tests. Results revealed that the CBM measures were more sensitive to growth over a short period of time than were these other measures.

Later studies focused more closely on the statistical properties of the growth rates produced by repeated CBM data. Skiba, Deno, Marston, and Wesson (1986) examined slopes for CBM reading-aloud data collected three to five times per week over a 6-month period for 67 students with disabilities (Grades 1–7). The mean increase in number of words read per week was 1.55 words per minute and tended to have an inverse relationship to grade level (a pattern that has been replicated in other research that will be reviewed later). The mean SEE across grade levels was 10.17, with a range of 8.45 in Grade 1 to 1.56 in Grade 4.

In 1992, Fuchs and Fuchs examined the ratio of the slope value to the standard error of estimate. SEE indexes the degree of intraindividual instability in the CBM data. A large SEE in proportion to the slope makes the graphs difficult to interpret. Fuchs and Fuchs (1992) found that maze had a relatively small SEE in proportion to the slope when compared to the other measures that might serve as alternatives for reading aloud.

Shinn et al. (1989) described a desirable slope as one that would be (a) easy to compute and interpret, (b) accurate in the sense of not producing systematic over- or underpredictions of performance, and (c) precise in the sense that individual errors of prediction are minimized. In a series of studies, these characteristics were compared for different methods for slope calculation (Good & Shinn, 1990; Parker & Tindal, 1992; Shinn, Good, & Stein, 1989). Results of these studies supported the use of ordinary least squares (OLS) for calculating slope compared to other methods (see Note 3). OLS was not the easiest method for calculating slope, but it did not systematically over- or underpredict future performance given a relatively small number (e.g., 10) of data points, and it minimized individual errors of prediction. Results of the slope studies revealed negatively accelerating growth rates within the academic year.

Dunn and Eckert (2002) discussed limitations of studies of slope, stating that the studies compared line-fitting methods to each other rather than to an absolute standard of technical adequacy and focused primarily on predicting later data from earlier data in the same data set. Dunn and Eckert proposed examination of the correlations between words read correctly and time (school day) as an indicator of accuracy of the slope. The square of the correlation coefficient would reflect the amount of variability in the words read correctly due to time. A higher percentage would indicate a more accurate slope line. Dunn and Eckert compared slopes for grade-level and above-grade-level material (see description of the study in the Materials section). Results revealed median correlation coefficients of .15 and .14, respectively, indicating that very little variability in the number of words read correctly could be attributable to time.

Both Hintze and Christ (2004) and Christ (2006) considered SEb in their investigations of slope reliability. Hintze and Christ (2004; reviewed earlier) found that controlling the difficulty of progress-monitoring material reduced both SEE and SEb. Christ (2006) examined the likely magnitudes of SEb for different values of SEE and for different durations of progress monitoring. Results revealed that the longer the progress-monitoring duration, the smaller the SEE and the smaller the SEbs (9.19 for 2 weeks compared to .42 for 15 weeks). If one assumed a moderate amount of SEE, 9 to 10 data points were needed to reduce SEbs to levels below 1. Ten to 12 data points reduced SEbs to between .59 and .78. Christ (2006) discussed the importance of controlling testing conditions and passage difficulty to reduce SEE.

Most studies of slope have focused on the stability of the slopes, but little attention has been paid to the validity of the slopes generated by CBM data. We use validity in this sense to refer to the degree to which slope values predict performance on external criterion measures. With the development and ease of use of new statistical techniques, investigations of the validity of slopes have become easier to accomplish. For example, Shin, Deno, and Espin (2000) used Hierarchical Linear Modeling (HLM) to examine growth rates on maze selection measures for second-grade students over the course of a year. Results revealed that maze selection sensitively re-

flected improvement in student performance across the year and reflected interindividual differences in growth rates. In addition, growth on maze was positively and significantly related to later performance on a standardized reading achievement test. Tichá et al. (2007) and Espin et al. (2007) used HLM to examine growth rates produced by maze selection for middle-school students. They found that growth on a maze selection task was significantly related to performance on a state standards test and to growth on a standardized achievement test. Finally, Speece and Ritchey (2005) used growth-curve analysis to examine the growth rates produced by reading aloud for a sample of at-risk first-graders. Results revealed that rate of growth on the reading-aloud measures in first grade predicted rate of growth in second grade, as well as end-of-year performance in second grade.

Standard Rates of Growth

A question often posed by practitioners is how much growth one can expect on CBM measures and whether growth expectations differ by age or performance levels. Several studies have examined these standard rates of growth. Marston, Lowry, Deno, and Mirkin (1981) examined growth rates from fall to winter to spring for students in Grades 1 through 6 for both reading aloud and word ID. Fifty-eight randomly selected general education students participated. Both word ID and reading-aloud measures reflected growth over time and exhibited steep continual, linear growth. Dramatic changes were observed in the earlier grades, with less dramatic changes in the upper grades, a result seen in other studies (e.g., Deno et al., 1982; MacMillan, 2000). Hasbrouck and Tindal (1992) published CBM norms for reading aloud on the basis of data collected over a period of 9 years from 7,000 to 9,000 students in Grades 2 through 5. They reported normative performance levels by time of year (fall, winter, spring), grade, and percentile level within grade.

The limitation of grade-level standards is that they do not take into account an individual's beginning level of performance. Students who are low performing may never reach grade-level standards, even when they are improving (Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993). Fuchs et al. (1993) addressed intraindividual norms for weekly growth rates on CBM reading-aloud and maze measures. Participants were two samples of students in Grades 1 through 6. During the first year of the study, students ($N = 117$) were measured in grade-level material once each week using reading aloud. During the second year, students ($N = 257$) were measured in grade-level material at least monthly using a computerized maze selection program. Results revealed that for the majority of students, a linear model fit the growth data for both reading aloud and maze, but for a proportion of the students, a quadratic model fit the data. For most of these cases, a slightly negative pattern of growth was found. Weekly growth rates were calculated by grade level. Results revealed differences in growth rates as a function of grade for reading aloud but not for maze. As with earlier research, the magnitude of slopes was found to decrease with an increase in grade. However, similar to the findings of Jenkins and Jewell (1993), no such relationship was found for maze. Fuchs et al. (1993) presented what they termed realistic and ambitious standards for weekly growth. For reading aloud, these standards were, by grade level, 2 and 3 words per week (Grade 1), 1.5 and 2.0 (Grade 2), 1.0 and 1.5 (Grade 3), .85 and 1.1 (Grade 4), .5 and .8 (Grade 5), and .3 and .65 (Grade 6). For maze, realistic and ambitious standards for growth remained the same across grades and were .39 and .84 word choices per week, respectively.

Deno, Fuchs, Marston, and Shin (2001) established academic growth standards for students in general and special education under typical and effective instructional practices. Growth rates under typical conditions were generated using extant databases from four educational agencies across the country. Growth rates under effective instructional conditions were generated by combining data across studies in which instructional practices had been implemented and shown to be effective. As with previous research, results revealed the greatest growth in the early grades, with a decrease in growth rates with age. Typical growth rates for general and special education by grade level were 1.8 and .83 (Grade 1), 1.66 and .57 (Grade 2), 1.18 and .58 (Grade 3), 1.01 and .58 (Grade 4), .58 and .58 (Grade 5), and .66 and .62 (Grade 6). Large differences existed between the growth rates of general and special education students up until Grade 4. In the second part of the study, Deno et al. (2001) examined growth rates for students with learning disabilities who had received effective instructional treatments across five different studies. Growth rates for both reading aloud and maze were reported. Growth rates for reading aloud ranged from .83 to 2.10 words per week. Growth rates for maze ranged from .56 to .70 words per week. With regard to the reading-aloud data, the results were striking in the sense that under effective instructional conditions, students with LD exhibited growth rates close to the typical growth rates for general education students seen in the first part of the study.

Summary and Discussion

Research on the technical adequacy of the growth rates produced by CBM measures supports the dependability of single measures; however, results are more mixed with regard to slope. Slopes have been found to predict future CBM performance and have also been found to predict performance and progress on external measures. However, concern exists about the variability of the data points around the slope and the variability of the slope values themselves. Concern also exists about establishing equivalence of passages used for progress monitoring. More research is needed on the technical properties of the slope values produced by various CBM reading measures, especially as they relate to the use of CBM within the framework of high-stakes decisions, such as those involved in eligibility determination. Future research should

also examine the relative effects of exceptionally high or low data points at various time points on slope values, and practitioners' abilities to interpret slope values and apply them to instructional decision making.

Although studies of standard growth rates make an important contribution to the CBM literature, the results must be viewed with caution. First, as acknowledged by the researchers, none of the studies employed a nationally representative data set. Second, as discussed earlier in this review, the growth rates obtained in these studies may have been affected by the materials in the studies, specifically the equivalence of the passages used to establish the growth rates. To illustrate this point, perusal of the various studies included in this review reveal growth rates for general education students in Grade 4 ranging from .24 to 2.69 words per week. Similar diversity in growth rates can be seen at the other grade levels. It may be important to consider establishing nationally standard growth rates by creating a standard set of passages that are controlled for difficulty and measuring a nationally representative sample of students on a weekly basis. In addition, following the lead set by Fuchs et al. (1993), it seems important to consider both typical and ambitious growth standards.

Conclusion

We begin our Conclusion section with a return to the question posed at the beginning of the article: Are CBM measures "valid enough" for the purposes for which they are used? We believe that the data have supported the validity of the CBM reading-aloud measure for use by classroom teachers as an indicator of the performance and progress in reading for elementary school students, Grades 2 to 5. The measures have been shown to relate to a variety of criterion measures across a multitude of studies conducted over many years with different participants, methods, materials, and researchers. In addition, the theoretical basis for the strength of reading aloud as an indicator of general reading proficiency has been supported, and the measures have been shown to be dependable.

Despite the positive support for the reading-aloud measures, we offer some words of caution about the CBM reading measures in general. First, with respect to reading aloud, the measures have not always been found to be strongly related to criterion measures. Although correlations are consistently in the .70s or above, there are studies that have resulted in correlations in the .40s between reading aloud and criterion variables. A closer look at the reasons for these exceptions is in order. Research has suggested that the relationship between reading aloud and criterion variables may decrease as students get older; however, other factors may also influence the relationship, such as the influence of various passage characteristics on the relation between reading aloud and criterion variables.

Second, the overwhelming majority of research in CBM reading has been done with reading aloud, and with samples of students in Grades 2 through 5. The question of the validity of the measures for younger (K–Grade 1) and older (Grades 6–12) students, and for students of diverse backgrounds, is still open for examination. With regard to student age, research suggests that a word-ID measure may be more appropriate than reading aloud would be for younger, beginning readers, and that maze may be more appropriate for older, middle school students. With regard to students with diverse needs, correlations between the CBM measures and criterion variables have generally been positive across various groups, but there is evidence that CBM reading aloud overestimates the performance of African American and EL students, which could result in underidentification for services. For students with sensory disabilities, too few studies have been conducted to draw conclusions.

Finally, we raise the issue of the uses of CBM measures. CBM measures are no longer simply a means for special education teachers to evaluate the effects of instructional programs. The use of CBM measures has shifted from that of monitoring the progress of students in special education to use in high-stakes decisions that carry important social consequences. This shift creates a new set of standards for the measures. The validity of the measures for these new purposes has yet to be established. There is little known, relatively speaking, about the technical characteristics of the slopes produced by CBM measures. For example, what is the best method for determining validity and reliability of the slope? How many data points are needed to obtain a reliable and valid slope? Does this number differ with the age of the student, the material used, or the equivalency of "parallel" passages? How do we establish parallel passages? There is also little known about normative or ambitious rates of performance and growth. For example, how much growth should be expected from students at different age and performance levels? Should national norms be developed? If so, must standard sets of materials be developed for this purpose? Finally, little is known about teachers' understanding of CBM progress data and the thinking processes teachers use as they interpret and use data for decision making. For example, how accurate are teachers at interpreting CBM data? How long does it take them to learn how to interpret and use CBM data? How easy is it for teachers to connect progress-monitoring data to instructional decisions, and are there methods to enhance their ability to tie data and instruction together? We believe that areas must be explored if CBM is to be used as a part of a decision-making process related to determining the need for special education services.

Despite these words of caution, we believe that the flexibility and durability of CBM in reading across different measures, materials, settings, students, and situations is notable. This flexibility and durability provide the basis for considering the development of a seamless and flexible system of progress monitoring that could be used across students of various ages and performance levels. Such a system might allow one to follow the progress of a student from kindergarten to Grade 12, using the same measures and materials or linking

measures and materials. Many of the issues raised above (e.g., establishment of normative levels and growth rates using standard material) would need to be addressed before such as system could be realized, but the research on the development of CBM measures in reading is at a point where development of such a seamless and flexible system is conceivable.

AUTHORS' NOTES

1. The Research Institute on Progress Monitoring at the University of Minnesota is funded by the U.S. Department of Education, Office of Special Education Programs (Award H324H030003) and partially supported the completion of this work.
2. We wish to thank the Netherlands Institute for Advanced Study in the Humanities and Social Sciences for its support in the preparation of this manuscript.
3. We wish to thank Stan Deno, Ed Shapiro, Ted Christ, Jongho Shin, and John Hintze for input on parts of this article.

NOTES

1. These measures can be found under various names in the literature. Reading aloud is often referred to as *oral reading fluency* and *oral reading*. *Maze selection* is referred to as *maze* and *multiple-choice cloze*. *Word identification* is also referred to as *isolated word reading, word list reading,* and *word identification fluency.* We have chosen to use *reading aloud, maze selection,* and *word identification* consistently throughout the article because these terms represent the behavior that the student performs on each CBM measurement task; that is, students read aloud, select maze choices, or identify words.
2. Hintze and Silberglitt (2005) and Silberglitt and Hintze (2005) examined various methods for determining cut scores using the CBM measures. Results supported the use of logistic regression either alone or in combination with Receiver Operating Characteristics (ROC) curve analysis. Given the fact that using logistic regression alone is easier than using it with ROC and given the similarities in the pattern of results between the two approaches, we report results here for logistic regression analyses only.
3. Although the research supported the use of OLS for calculating slope, this is not a method that can be easily calculated by hand by teachers. Parker and Tindal (1992) proposed an alternative, Tukey I, that had good statistical properties and could be calculated by hand. To our knowledge, this method has never caught on in either CBM research or practice. In research, the overwhelming majority of studies calculate slope using OLS.

REFERENCES

Allinder, R. M., & Eccarius, M. A. (1999). Exploring the technical adequacy of curriculum- based measurement in reading for children who use manually coded English. *Exceptional Children, 65,* 271–283.

Ardoin, S. P., Suldo, S. M., Witt, J., Aldrich, S., & McDonald, E. (2005). Accuracy of readability estimates' predictions of CBM performance. *School Psychology Review, 20,* 1–22.

Ardoin, S. P., Witt, J. C., Suldo, S. M., Connell, J. E., Koenig, J. L., Resetar, J. L., et al. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psychology Review, 33,* 218–233.

Bain, S. K., & Garlock, J. W. (1992). Cross-validation of criterion-related validity for CBM reading passages. *Diagnostique, 17,* 202–208.

Baker, S. K., & Good, R. H. (1995). Curriculum-based measurement of English reading with bilingual Hispanic students: A validation study with second-grade students. *School Psychology Review, 24,* 561–578.

Bradley-Klug, K. L., Shapiro, E. S., Lutz, J. G., & DuPaul, G. J. (1998). Evaluation of oral reading rate as a curriculum-based measure within literature-based curriculum. *Journal of School Psychology, 36,* 183–197.

Brown-Chidsey, R., Davis, L., & Maya, C. (2003). Sources of variation in curriculum-based measures of silent reading. *Psychology in the Schools, 40,* 363–377.

Brown-Chidsey, R., Johnson, P., Jr., & Fernstrom, R. (2005). Comparison of grade-level controlled and literature-based maze CBM reading passages. *School Psychology Review, 34,* 387–394.

Christ, T. J. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimates of standard error of slope to construct confidence intervals. *School Psychology Review, 35,* 128–133.

Christ, T. J., & Silberglitt, B. (in press). Curriculum-based measurement of oral reading fluency: The standard error of measurement. *School Psychology Review.*

Compton, D. L., Appleton, A. C., & Hosp, M. K. (2004). Exploring the relationship between text-leveling systems and reading accuracy and fluency in second-grade students who are average and poor decoders. *Learning Disabilities Research & Practice, 19,* 176–184.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98,* 394–409.

Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment, 7,* 303–323.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Daly, E. J., III, Wright, J. A., Kelly, S. Q., & Martens, B. K. (1997). Measures of early academic skills: Reliability and validity with a first grade sample. *School Psychology Quarterly, 12,* 268–280.

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

Deno, S. L. (1990). Individual differences and individual difference: The essential difference of special education. *The Journal of Special Education, 24,* 160–173.

Deno, S. L., Fuchs, L. S., Marston, D., & Shin, J. H. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30,* 507–524.

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

Deno, S. L., Maruyama, G., Espin, C. A., & Cohen, C. (1989). *The basic academic skills samples* (BASS). Minneapolis, MN: University of Minnesota.

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49,* 36–45.

Dunn, E. K., & Eckert, T. L. (2002). Curriculum-based measurement in reading: A comparison of similar versus challenging material. *School Psychology Quarterly, 17,* 24–46.

Espin, C. A., & Deno, S. L. (1993a). Content-specific and general reading disabilities of secondary students: Identification and educational relevance. *The Journal of Special Education, 27,* 321–337.

Espin, C. A., & Deno, S. L. (1993b). Performance in reading from content area text as an indicator of achievement. *Remedial and Special Education, 14,* 47–59.

Espin, C. A., & Deno, S. L. (1994-95). Curriculum-based measures for secondary students: Utility and task specificity of text-based reading and vocabulary measures for predicting performance on content-area tasks. *Diagnostique, 20,* 121–142.

Espin, C. A., Deno, S. L., Maruyama, G., & Cohen, C. (1989, March). *The Basic Academic Skills Samples (BASS): An instrument for the screening and identification of children at risk for failure in regular education classrooms.* Paper presented at the National Convention of the American Educational Research Association, San Francisco, CA.

Espin, C. A., & Foegen, A. (1996). Validity of general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children, 62,* 497–514.

Espin, C., Wallace, T., Lembke, E., Campbell, H., & Long, J. (2007). *Creating a progress measurement system in reading for secondary students: Monitoring progress towards meeting high stakes standards.* 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, L. S., & Deno, S. L. (1992). Effects of curriculum within curriculum-based measurement. *Exceptional Children, 58,* 232–243.

Fuchs, L. S., & Deno, S. L. (1994). Must instructionally useful performance assessment be based in curriculum? *Exceptional Children, 61,* 15–24.

Fuchs, L. S., & Fuchs, D. (1990). [Relations among the Reading Comprehension subtest of the Stanford Achievement Test and maze, recall, question answering, and fluency measures]. Unpublished data.

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21,* 45–58.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. Exceptional Children, 71, 7–21.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989). Effects of instrumental use of curriculum-based measurement to enhance instructional programs. *Remedial and Special Education, 10,* 43–52.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22,* 27–48.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal comprehension measures. *Remedial and Special Education, 9,* 20–28.

Fuchs, L. S., Tindal, G., & Deno, S. L. (1981). Effects of varying item domain and sample duration on technical characteristics of daily measures in reading. *Institute for Research on Learning Disabilities, 48.*

Gardner, R. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1982). *Stanford achievement test.* Iowa City: Harcourt Brace Jovanovich.

Good, R. H., & Shinn, M. R. (1990). Forecasting accuracy of slope estimates for reading curriculum-based measurements: Empirical evidence. *Behavioral Assessment, 12,* 179–193.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5,* 257–288.

Graves, A. W., Plasencia-Peinado, J., Deno, S. L., & Johnson, J. R. (2005). Formatively evaluating the reading progress of first-grade English Learners in multiple-language classrooms. *Remedial and Special Education, 26,* 215–225.

Guthrie, J. T. (1973). Reading comprehension and syntactic responses in good and poor readers. *Journal of Educational Psychology, 65,* 294–299.

Guthrie, J. T., Seifert, M., Burnham, N. A., & Caplan, R. I. (1974). The maze technique to assess, monitor reading comprehension. *The Reading Teacher, 28,* 161–168.

Hamilton, C., & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills. *School Psychology Review, 32,* 228–240.

Harcourt Brace & Company. (1997). *Stanford Achievement Test Series–Ninth edition: Technical data report.* San Antonio, TX: Author.

Harcourt Educational Measurement. (1998). *Metropolitan achievement test* (7th ed.). San Antonio, TX: Harcourt Assessment.

Harris, A. P., & Jacobson, M. D. (1972). *Basic elementary reading vocabularies.* New York: MacMillan.

Hartman, J. M., & Fuller, M. L. (1997). The development of curriculum-based measurement norms in literature-based classrooms. *Journal of School Psychology, 35,* 377–389.

Hasbrouck, J. E., & Tindal, G. (1992). Curriculum-based oral reading fluency norms for students in grades 2 through 5. *Teaching Exceptional Children, 24*(3), 41–44.

Hintze, J. M., Callahan, J. E., Matthews, W. J., Williams, S. A. S., & Tobin, K. G. (2002). Oral reading fluency and prediction of reading comprehension in African American and Caucasian elementary school children. *School Psychology Review, 31,* 540–553.

Hintze, J. M., & Christ, T. J. (2004). An examination of variability as a function of passage variance in CBM progress monitoring. *School Psychology Review, 33,* 204–217.

Hintze, J. M., Daly, E. J., & Shapiro, E. S. (1998). An investigation of the effects of passage difficulty level on outcomes of oral reading fluency progress monitoring. *School Psychology Review, 27,* 433–445.

Hintze, J. M., Owen, S. V., Shapiro, E. S., & Daly, E. J. (2000). Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly, 15,* 52–68.

Hintze, J. M., & Pelle Petitte, H. A. (2001). The generalizability of CBM oral reading fluency measures across general and special education. *Journal of Psychoeducational Assessment, 19,* 158–170.

Hintze, J. M., & Shapiro, E. S. (1997). Curriculum-based measurement and literature-based reading: Is curriculum-based measurement meeting the needs of changing reading curricula? *Journal of School Psychology, 35,* 351–375.

Hintze, J. M., Shapiro, E. S., Conte, K. L., & Basile, I. M. (1997). Oral reading fluency and authentic reading material: Criterion validity of the technical features of CBM survey-level assessment. *School Psychology Review, 26,* 535–553.

Hintze, J. M., Shapiro, E. S., & Lutz, J. (1994). The effects of curriculum on the sensitivity of curriculum-based measurement in reading. *The Journal of Special Education, 28,* 188–202.

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review, 34,* 372–386.

Hixson, M. D., & McGlinchey, M. T. (2004). The relationship between race, income, and oral reading fluency and performance on two reading comprehension measures. *Journal of Psychoeducational Assessment, 22,* 351–364.

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

Hosp, M. K., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review, 34,* 9–26.

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children, 59,* 421–432.

Kaminski, R. A., & Good, R. H., III. (1998). Assessing early literacy skills in a problem-solving model: Dynamic Indicators of Basic Early Literacy Skills. In M. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 113–142). New York: Guilford Press.

Karlsen, B., & Gardner, E. (1985). *Stanford diagnostic reading test* (3rd ed.). San Antonio, TX: Psychological Corp.

Kaufman, A. S., & Kaufman, N. L. (1985). *Kaufman test of educational achievement* (brief form). Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1990). *Kaufman brief intelligence test.* Circle Pines, MN: American Guidance Service.

Klein, J. R., & Jimerson, S. R. (2005). Examining ethnic, gender, language, and socioeconomic bias in oral reading fluency scores among Caucasian and Hispanic students. *School Psychology Quarterly, 20,* 23–50.

Koslin, B. L., Koslin, S., Zeno, S. M., & Ovens, S. H. (1989). *The Degrees of Reading Power Test: Primary and standard forms.* Brewster, NY: Touchstone Applied Science Associates.

Kranzler, J. H., Brownell, M. T., & Miller, M. D. (1998). The construct validity of curriculum-based measurement of reading: An empirical test of a plausible rival hypothesis. *Journal of School Psychology, 36,* 399–415.

Kranzler, J. H., Miller, M. D., & Jordan, L. (1999). An examination of racial/ethnic and gender bias on curriculum-based measurement of reading. *School Psychology Quarterly, 14,* 327–342.

MacGinitie, W. H., Kamons, J., Kowalski, R. L., MacGinitie, R. K., & McKay, T. (1978). *Gates-MacGinitie Reading Tests* (2nd ed.). Chicago: Riverside.

MacMillan, P. (2000). Simultaneous measurement of reading growth, gender, and relative-age effects: Many-faceted Rasch applied to CBM reading scores. *Journal of Applied Measurement, 1,* 393–408.

Markell, M. A., & Deno, S. L. (1997). Effects of increasing oral reading: Generalization across reading tasks. *The Journal of Special Education, 31,* 233–250.

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. L. (1982). Implementation of direct and repeated measurement in the school setting. *Institute for Research on Learning Disabilities,* 106.

Marston, D., Deno, S. L., & Tindal, G. (1983). A comparison of standardized achievement tests and direct measurement techniques in measuring pupil progress. *Institute for Research on Learning Disabilities,* 126.

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

McConnell, S. R., McEvoy, M. A., & Priest, J. S. (2002). "Growing" measures for monitoring progress in early childhood education: A research and development process for individual growth and development indicators. *Assessment for Effective Intervention, 27,* 3–14.

McGlinchey, M. T., & Hixson, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33,* 193–203.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III technical manual*. Itasca, IL: Riverside.

Mehrens, W. A., & Clarizio, H. F. (1993). Curriculum-based measurement: Conceptual and psychometric considerations. *Psychology in the Schools, 30,* 241–254.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5–11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Michigan State Board of Education. (1999). 1999 *Update of the Essential Skills Reading Test Blueprint Michigan Educational Assessment Program*. Lansing: Michigan State Board of Education.

Morgan, S. K., & Bradley-Johnson, S. (1995). Technical adequacy of curriculum-based measurement for Braille readers. *School Psychology Review, 24,* 94–103.

Minnesota Department of Children, Families, and Learning. (1998–2002). *Test specifications: Minnesota Comprehensive Assessment–Reading.* St. Paul, MN: Author.

Muyskens, P., & Marston, D. B. (2006). *The relationship between curriculum-based measurement and outcomes on high-stakes tests with secondary students.* Minneapolis Public Schools. Unpublished manuscript.

Oregon Department of Education. (2000). *Statewide assessment results 2000*. Retrieved from the World Wide Web: http://www.ode.state.or.us/

Parker, R., & Tindal, G. (1992). Estimating trend in progress monitoring data: A comparison of simple line-fitting methods. *School Psychology Review, 21,* 300–312.

Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment, 23,* 326–338.

Powell-Smith, K. A., & Bradley-Klug, K. L. (2001). Another look at the "C" in CBM: Does it really matter if curriculum-based measurement reading probes are curriculum-based? *Psychology in the Schools, 38,* 299–312.

Prescott, G. A., Balow, I. H., Hogan, T. P., & Farr, R. C. (1984). *Metropolitan achievement tests* (MAT-6). San Antonio, TX: Psychological Corp.

Reid, D. K., Hresko, W. P., Hammill, D. D., & Wiltshire, S. M. (1991). *Test of early reading ability: Special edition for students who are deaf or hard of hearing.* Austin, TX: PRO-ED.

Riley-Heller, N., Kelly-Vance, L., & Shriver, M. (2005). Curriculum-based measurement: Generic vs. curriculum-dependent probes. *Journal of Applied School Psychology, 21*(1), 141–162.

Scannell, D. P., Haugh, O. M., Schild, A. H., & Ulmer, G. (1986). *Tests of achievement and proficiency.* Chicago, IL: Riverside.

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education, 34,* 164–172.

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

Shinn, M. R., Gleason, M. M., & Tindal, G. (1989). Varying the difficulty of testing materials: Implications for curriculum-based measurement. *The Journal of Special Education, 23,* 223–233.

Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21,* 459– 479.

Shinn, M. R., Good, R. H., & Stein, S. (1989). Summarizing trend in student achievement: A comparison of methods. *School Psychology Review, 18,* 356–370.

Shinn, M. R., Ysseldyke, J. E., Deno, S. L., & Tindal, G. A. (1986). A comparison of differences between students labeled Learning Disabled and Low Achieving on measures of classroom performance. *Journal of Learning Disabilities, 19,* 545–551.

Silberglitt, B., & Hintze, J. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23,* 304–325.

Skiba, R. J., Deno, S. L., Marston, D., & Wesson, C. (1986). Characteristics of time-series data collected through curriculum-based reading measurement. *Diagnostique, 12,* 3–15.

Spache, G. (1981). *Diagnostic reading scales*. Monterey, CA: CTB/McGraw-Hill.

Speece, D. L., & Ritchey, K. D. (2005). A longitudinal study of the development of oral reading fluency in young children at risk for reading failure. *Journal of Learning Disabilities, 38,* 387–399.

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30,* 407–419.

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.

Sticht, T. G. (1973). Research toward the design, development and evaluation of a job-functional literacy program for the U.S. Army. *Literacy Discussions, 4,* 339–369.

Tichá, R., Espin, C. A., & Wayman, M. M. (2007). *Reading progress monitoring for secondary-school students: Reliability, validity, and sensitivity to growth of reading aloud and maze selection measures*. Manuscript submitted for publication.

Tindal, G., Flick, D., & Cole, C. (1992). The effect of curriculum on inferences of reading performance and improvement. *Diagnostique, 18,* 69–84.

Tindal, G., Marston, D., Deno, S. L., & Germann, G. (1982). Curriculum differences in direct repeated measures of reading. *Institute for Research on Learning Disabilities, 93*.

Washington Assessment of Student Learning: Test sample. (1998). Olympia, WA: Office of the Superintendent of Public Instruction.

Woodcock, R. W. (1973). *Woodcock reading mastery tests manual.* Circle Pines, MN: American Guidance Service.

Woodcock, R. W. (1987). *Woodcock reading mastery test–Revised.* Circle Pines, MN: American Guidance Service.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson test of achievement: Standard and supplemental batteries.* Allen, TX: DLM Resources.

Wiley, H. I., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education, 26,* 207–214.

Yell, M. L. (1992). Barriers to implementing curriculum-based measurement. *Diagnostique, 18,* 99–112.

Yovanoff, P., Duesbery, L., Alonzo, J., & Tindal, G. (2005). Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. *Educational Measurement: Issues and Practice, 24*(3), 4–12.