



ELSEVIER

Journal of School Psychology 47 (2009) 427–469

Journal of
**School
Psychology**

Curriculum-Based Measurement Oral Reading as an indicator of reading achievement: A meta-analysis of the correlational evidence

Amy L. Reschly^{a,*}, Todd W. Busch^b, Joseph Betts^c,
Stanley L. Deno^d, Jeffrey D. Long^d

^a *University of Georgia, United States*

^b *University of St. Thomas, United States*

^c *Renaissance Learning, United States*

^d *University of Minnesota, United States*

Received 17 February 2008; received in revised form 22 June 2009; accepted 11 July 2009

Abstract

This meta-analysis summarized the correlational evidence of the association between the CBM Oral Reading measure (R-CBM) and other standardized measures of reading achievement for students in grades 1–6. Potential moderating variables were also examined (source of criterion test, administration format, grade level, length of time, and type of reading subtest score). Results indicated a significant, strong overall correlation among R-CBM and other standardized tests of reading achievement and differences in correlations as a function of source of test, administration format, and reading subtest type. No differences in the magnitude of correlations were found across grade levels. In addition, there was minimal evidence of publication bias. Results are discussed in terms of existing literature and directions for future research.

© 2009 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.

Keywords: Curriculum Based Measurement; R-CBM; Reading achievement

In the 1970s, Deno and colleagues from the University of Minnesota set out to create a set of measurement procedures that could be used to efficiently monitor student progress in core educational skills and evaluate the effectiveness of instructional interventions, with the

* Corresponding author. Department of Educational Psychology & Instructional Technology, University of Georgia, Athens, GA, 30602, United States.

E-mail address: reschly@uga.edu (A.L. Reschly).

ACTION EDITOR: Randy Floyd.

0022-4405/\$ - see front matter © 2009 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.

doi:10.1016/j.jsp.2009.07.001

goal of accelerating student achievement. This work, referred to as Curriculum-Based Measurement (CBM), resulted in measures that were amenable to frequent administrations, sensitive to small changes in growth, inexpensive, and time efficient (Deno, 1985, 1992). Thirty years of research provides evidence of the reliability and validity of these measures as indicators of student achievement in reading, mathematics, and writing.

The data garnered from CBM measures have been used for a variety of purposes in general, remedial, and special education. Research indicates that when CBM data are used to monitor student performance and guide instructional modifications, student achievement is raised (Fuchs, Deno, & Mirkin, 1984; Fuchs & Fuchs, 1986; Fuchs, Fuchs, Hamlett, & Ferguson, 1992; Mirkin, Deno, Tindal, & Kuehnle, 1982; Stecker & Fuchs, 2000). In addition to individual progress monitoring, CBM data have been used in Individual Education Program (IEP) goals to evaluate the reintegration of special education students into general education classes (Powell-Smith & Stewart, 1998; Shinn, Powell-Smith, Good, & Baker, 1997), to create school- and district-level norms (Shinn, 1989), and for program evaluation (Tindal, 1989). Measures have also been modified for use with exceptional populations (Allinder & Eccarius, 1999; Morgan & Bradley-Johnson, 1995), translated into other languages and countries (Kaminitz-Berkooza & Shapiro, 2005; Yeh, 1992), and used with English Learners (e.g., Baker & Good, 1995; Wiley & Deno, 2005). Further, there have been extensions of the CBM measurement philosophy to social, pre-academic skill areas and content areas (e.g., Individual Growth and Development Indicators, Dynamic Indicators of Basic Early Literacy Skills [DIBELS]; Espin, Shin, & Busch, 2005; Greenwood, Dunn, Ward, & Luze, 2003; Kaminski & Good, 1996, 1998; Lembke, Foegen, Whittaker, & Hampton, 2008; McConnell, McEvoy, & Priest, 2002) and age groups (i.e., infancy, preschool, kindergarten, and secondary levels).

Of all CBM measures and skill areas, the most widely researched and utilized in schools across the U.S. is the oral reading or read-aloud measure (R-CBM).¹ For the R-CBM measure, students are given a reading passage, typically one at their grade or instructional level, and asked to read aloud for 1 min. At the end of 1 min, passages are scored for the number of words read correctly. R-CBM, and its less frequently used CBM-reading counterpart, MAZE, are general outcome measures (Fuchs & Deno, 1991). General outcome measures are standardized, longitudinal assessments that use the same methods and presumably equivalent content over time. Further, the measures may be used to designate the performance desired from a student at the end of the monitoring period (Fuchs & Fuchs, 1999). Therefore, an educator may set a long-range goal, typically the end of the school year, and monitor student progress toward that goal using different reading passages of presumably equivalent difficulty. These monitoring data are used to determine whether the student is on-track to reach the specified goal and inform instructional efforts to accelerate student performance. In the case of reading, progress toward this goal shows that a student is *overall* becoming a better reader.

Scores from CBM measures in general, and R-CBM in particular, have been evaluated according to traditional psychometric criteria for reliability and validity (Deno, 1992). R-CBM scores have been shown to be technically adequate in terms of reliability and are

¹ R-CBM is the term used throughout this article to describe data derived from the standard administration and scoring of the CBM read-aloud measure.

moderately to highly correlated with scores from other standardized measures of reading achievement (e.g., Marston, 1989). The psychometric properties of the scores of the R-CBM measure, in combination with its ability to function as a general outcome measure and the ease of administration, time efficiency, low cost, and frequency with which the measures may be given, has led to use by educators across U.S.

Until recently, questions raised about CBM were largely academic in nature and concerned philosophical issues and misunderstandings of its purpose (Shinn & Bamonto, 1998). In practice, the primary concerns expressed were the amount of time taken to assess rather than instruct and the face validity of R-CBM as a measure of general reading proficiency (i.e., whether data derived from scoring a student's passage reading for speed and accuracy could possibly represent their overall reading skill; Foegen, Espin, Allinder, & Markell, 2001; Wesson, Deno, & King, 1984; Yell, Deno, & Marston, 1992). However, the zeitgeist created by increased accountability and high-stakes assessments stimulated the use of R-CBM data for related purposes, such as screening to identify lower performing or "at-risk" students (Deno et al., 2009), benchmarking (Good, Simmons, & Kame'enui, 2001; Shinn, 1989), predicting performance on high stakes assessments (e.g., Buck & Torgesen, 2003; Crawford, Tindal, & Steiber, 2001; Hintze & Silbergitt, 2005; McGlinchey & Hixson, 2004; Shapiro, Keller, Lutz, Santoro, & Hintze, 2006), and as the basis for creating school- and district-wide improvement models (Deno et al., 2009). In addition, CBM in general, and R-CBM in particular, have emerged as a cornerstone of special education reform efforts (Grimes & Tilly, 1996; Marston, Muyskens, Lau, & Canter, 2003; Reschly & Bergstrom, 2009) and featured prominently in legislative initiatives such as *Reading First* and the reauthorization of the *Individuals with Disabilities Education Act* [IDEA]. The raised profile of CBM and increased use of the R-CBM measure and other CBM-like measures, such as the DIBELS, has spurred interest, criticism, and debate (e.g., Goodman, 2006; Kamii & Manning, 2005; Manzo, 2005a).

Some of the recent debate about R-CBM has revolved around the *use* of the tests. In particular, whether or not these and similar tests (i.e., DIBELS) should have featured so prominently in *Reading First* assessment plans (e.g., Glenn, 2007; Manzo, 2005b, 2007) and the use of CBM and R-CBM in determining a students' response to intervention and potential identification for a Learning Disability under the 2004 revision to IDEA (e.g., Naglieri & Crockett, 2005). Underlying the debate over the use of these measures in *Reading First* is a larger question over whether R-CBM is a measure of *Fluency* (e.g., Samuels, 2007), one of the core elements of reading instruction outlined by the National Research Panel (National Institute of Child Health and Human Development [NICHD], 2000). In the original work on R-CBM, the measure was referred to as *read aloud* or *oral reading*; however, in recent years another term, *oral reading fluency*, has been used to describe this measure, drawing the R-CBM measure into a larger debate over the nature of reading development and fluency and assessment of this construct (Samuels, 2007). The R-CBM measure is sometimes described as a fluency measure in that it is a time-limited task on which performance is quantified in terms of both speed and accuracy. The utility of R-CBM for making instructional decisions, however, is based on the many and varied empirical relationships between R-CBM and other measures rather than the validity of R-CBM as a measure of the construct of reading fluency, as described in theories of reading (e.g., LaBerge & Samuels, 1974) or by the National Reading Panel (NICHD,

2000). The debate over the use of R-CBM and CBM-like measures in *Reading First* and IDEA and the nature of fluency aside, this article focuses on the empirical literature base of the R-CBM measure and its associations with other standardized measures of reading achievement.

There is considerable evidence supporting the use of R-CBM as a measure of general reading proficiency and comprehension (Fuchs, Fuchs, Hosp, & Jenkins, 2001). However, a number of questions remain. For instance, a concern often voiced by practitioners and others is that R-CBM taps students' decoding skills rather than general reading achievement or comprehension (Hamilton & Shinn, 2003), and there are conflicting results regarding the consistency of the strength of the relationship between R-CBM and other standardized measures of reading achievement across grades. For example, Jenkins and Jewell (1993) found that the correlation coefficients between scores from R-CBM and those from tests of comprehension and total reading decreased across grades two through six. Similar findings were reported by Kranzler, Miller, and Jordan (1999). In contrast, Hosp and Fuchs (2005) found stable, high correlation coefficients among R-CBM scores and the Comprehension subtest scores of the Woodcock Reading Mastery Tests (WRMT-R; Woodcock, 1987) across grades one through four; however, the same investigation reported a small but significant decrease in the magnitude of correlations between R-CBM scores and total reading scores on the WRMT-R in grade four when compared to grades one, two, and three. Relatively few studies have examined R-CBM with middle and high school students.

In addition to lingering questions related to the fitness of R-CBM as an indicator of general reading proficiency and comprehension and the suggestion of grade-level declines in the magnitude of the association between scores derived from R-CBM probes and other standardized measures of reading achievement across elementary school, a number of other questions warrant further attention. For example, R-CBM is increasingly used to predict performance, sometimes across years, on state-derived high-stakes assessments and to establish benchmarks for passing these tests. Further, it is recognized that there is a great deal of variability in the difficulty and relative proficiency levels of various state tests (Peterson & Hess, 2005) and very weak associations have been found between proficiency levels on state tests and the National Assessment of Educational Progress (NAEP; Linn, 2005).

In a related vein, over the last 30 years, external validity examinations of the relations between R-CBM probe scores with scores derived from other standardized tests of reading proficiency have shifted from individually administered achievement tests (e.g., Marston & Deno, 1982) to those from group-administered achievement tests that are frequently used for purposes of state-level, high-stakes assessment (e.g., Silberglitt & Hintze, 2005). Correlation coefficients derived from R-CBM probe scores and scores from other individually administered tests of reading proficiency may be higher due to similarity in administration format or potentially higher quality test construction of individually administered tests of reading proficiency. For example, different criteria have been suggested for evaluating the psychometric evidence of inferences drawn from test scores based on how the scores are used, with scores from screening and group tests requiring lower minimum reliability and validity coefficients than those used for individual decision-making (Salvia & Ysseldyke, 2007). Correlation coefficients between R-CBM probe scores

and scores derived from individual- and group-administered tests may also vary due to potentially greater self-direction required to complete group-administered tests, which would increase the likelihood that fatigue, inattention, and other variables may negatively affect student performance.

An important task in accumulating validity evidence for inferences drawn from scores on a particular test is determining the extent to which the scores function similarly for various individuals and subgroups (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, [AERA, APA, NCME], 1999). In the case of R-CBM, one must ask whether correlation coefficients from scores of R-CBM probes and other tests of reading achievement are similar across gender, socioeconomic, racial-ethnic, and general and special education groups. Given the use of the scores to establish benchmarks, predict future performance on high-stakes assessments, and identify students in need of additional intervention, systematic over- or under-prediction of performance for a particular subgroup of students would be problematic and may indicate a need to have different prediction equations for various subgroups (e.g., one for native English speakers, another for students who are English Learners). Studies of predictive bias of R-CBM with respect to measures of reading comprehension and general reading achievement have found disparate results with respect to bias in terms of racial-ethnic group membership and home language (e.g., Hintze, Callahan, Matthews, Williams, & Tobin, 2002; Klein & Jimerson, 2005; Kranzler et al., 1999).

To date, reviews of the R-CBM literature have been narrative, rather than quantitative in nature (Marston, 1989; Wayman, Wallace, Wiley, Ticha, & Espin, 2007) and a number of potential moderating variables have yet to be systematically examined. The purpose of this meta-analysis was to organize extant empirical results to obtain a quantification of the level of linear relation between R-CBM scores and commonly used reading tests across numerous published research endeavors. In addition, specific estimates were computed for facets of the relationship that were thought to have a moderating effect on the magnitude of the correlation coefficients. In this study, we examined the correlational evidence in the form of Pearson's correlation coefficients (r) for students in grades 1–6 with R-CBM administered according to standard CBM procedures in English prior to or concurrently with norm-referenced, standardized tests of reading achievement. Moderating variables were those related to students (grade-level, demographic characteristics) and criterion measures (source of test, administration format, and length of time between R-CBM and the administration of the reading criterion measures).

Method

Data collection

A systematic search of the literature was conducted to locate articles and technical reports for inclusion in the meta-analysis. Dissertations and conference proceedings were excluded because it was not possible to systematically access these sources. However, technical reports, accessible via the Internet, were included. The latter was an important consideration in that many correlations between R-CBM scores and state-specific assessment scores were only available in technical report format.

Key search terms were selected from those commonly used in the CBM reading literature and entered into the electronic search engines ERIC, Education Full Text, and PsychInfo. The number of hits for each term in ERIC, Education Full Text and PsychInfo, respectively, were 958, 1201, 737 for *read aloud*; 170, 142, 226 for *oral reading fluency*; 2468, 713, 2381 for *oral reading*; 205, 92, 314 for *CBM* (205, 92, 314); 4, 7, 15 for *R-CBM*; 0, 2, 0 for *CBM-OR*; 3, 3, 4 for *CBM-R*; and, 35, 23, 72 for *DIBELS*. Literature from the inception of CBM in the late 1970s (Deno & Mirkin, 1977) through June of 2008 was examined. As in other published meta-analyses (e.g., Swanson, Trainin, Denise, Necochea, & Hammill, 2003), we also conducted a hand search of journals that frequently publish this type of research (i.e., *Journal of School Psychology*, *Journal of Special Education*, *Remedial & Special Education*, *School Psychology Quarterly*, and *School Psychology Review*) to ensure the inclusion of articles that may have been missed by one of the search engines. In addition, the published work of authors who frequently publish CBM research was examined (D. Fuchs, L. Fuchs, Hintze, Marston, Shinn, Shapiro, Tindal, as well the authors of this paper) to ensure the comprehensiveness of the literature search. Finally, the reference lists of potential articles were scanned for additional citations. The use of these search terms also led to the examination of several articles that were intervention rather than criterion validity studies. These articles were included if correlations among R-CBM and other reading measures were reported and other criteria for inclusion were met. The first and second authors conducted the literature search and reviewed abstracts of articles identified in the keyword, author, hand, and reference searches. Using these methods, 105 studies were selected for additional review.

Three Internet sites that provided data used in this study were AIMSweb (<http://www.aimsweb.com>), DIBELS (<http://www.dibels.uoregon.edu>) and a searchable database of reports compiled by the federally funded Research Institute on Progress Monitoring (RIPM; <http://www.progressmonitoring.org>; Grant No. H324H030003) at the University of Minnesota. Technical reports from the Institute for Research on Learning Disabilities (1977–1983), the federally funded center in which much of the early validation work with CBM measures was conducted, were available from the RIPM site. From these three sites, 78 reports were selected for additional review.

Evaluation of articles and reports

From these searches, the first and second authors created a list of potential articles and reports and reviewed the full-text of each article or report for inclusion in the study. To be retained in the meta-analysis, R-CBM probes had to be administered and scored according to standard CBM criteria (e.g., instructions, scored for number of words read correctly in 1 min) in English prior to (across academic years) or within the same academic year as other norm-referenced, standardized tests of reading achievement for students in grades one through six. Studies were excluded if the grade level was above grade 6 (e.g., Fuchs, Fuchs, & Maxwell, 1988), standardized achievement scores were used to predict R-CBM performance across academic years (e.g., Alonso & Tindal, 2003), or R-CBM assessment tasks were modified for use with special populations (e.g., Allinder & Eccarius, 1999), and for technical reasons (i.e., the use of Grade Equivalent scores for the criterion measure as in Pressley, Hilden, & Shankland, 2005 and the collapse of grade-level data into one score for

analysis as in Wilson, Schendel, & Ulman, 1992). Following these procedures, a total of 41 studies were retained for analysis (Table 1).

Coding

The coding of data was guided by *a priori* questions and possible moderating variables. When more than one set of correlation coefficients was reported within a paper, those specific to grade levels of interest and those for whole groups—rather than for only educationally relevant subgroups—were entered into analyses. For example, in the Deno, Mirkin, Chiang, and Lowry (1980) paper, correlation coefficients were reported separately for general education students and those with learning disabilities as well as for the overall group. Those for the overall group were entered for further analysis.

Data were coded into eight categories. The coding of two categories, *correlation coefficient* and *sample size* associated with each respective coefficient, was straightforward. These were drawn directly from articles and reports. The remaining six categories are described below.

Source of test

Criterion measures were coded according to the dichotomy of state specific (e.g., Minnesota Comprehensive Assessment) or nationally normed, commercially available (national) tests of student reading achievement (e.g., WRMT-R; Woodcock, 1987).

Administration format

Criterion measures were also coded according to how the tests were administered: individually or to groups of students.

Type of criterion score

Correlation coefficients from criterion measures of reading achievement were coded according to the type of score reported: Comprehension, Vocabulary, Word Identification, Decoding, and Total Reading Score. These groupings came from the studies' authors descriptions of the tests rather than an analysis of stimulus materials and response formats. This rule was also true for the Total Reading Scores. Scores coded into this group were typically composites or overall scores as delineated in the authors' test descriptions. A list of tests and subtests, subtest groupings, and types of Total Reading Scores is presented in Table 2. At the outset, our goal was to examine the association between R-CBM and measures of achievement as a function of grade level, reading score type, and both grade level and score type. However, there were insufficient data to examine the correlations within grades for various score types.

Time

In general, the length of time between the administrations of two or more measures affects the magnitude of the correlations between these measures. The majority of studies completed to date have involved the administration of R-CBM and other standardized tests of reading achievement within the same academic year; however, a few studies have examined R-CBM and other tests across academic years. Correlations from studies retained

Table 1
Descriptive information for included studies.

Author	Region and school type	Sample characteristics ^a	Time ^b	Individual or group adm.	Source of test(s)	Included correlations (<i>r</i>) and score type
Ardoin et al. (2004)	Southeastern US	<i>N</i> = 77 3rd graders in regular education (35 females, 42 males; 58% White, 40% African American, 2% other)	W	Ind & Gp.	Natl.	.70, TR .74, C .42, C .62, WI .64, TR .35, V .58, C .73, TR .73, C .41, C .69, WI .66, TR .42, V .60, C
Bain and Garlock (1992)	2 elementary schools in rural, western FL	Grade 1 <i>N</i> = 66, Grade 2 <i>N</i> = 198, Grade 3 <i>N</i> = 215 Chapter 1 students	W	Gp.	Natl.	.69, TR .54, TR .79, TR .70, TR .74, TR
Baker et al. (2008)	16 school districts, 34 Reading First schools in OR	Each cohort approx. 2,400 students, 69% FRL	W, A	Gp.	State & Natl.	.72, TR .82, TR .63, TR .72, TR .72, TR .79, TR .80, TR .58, TR

Baker and Good (1995)	Rural district in WA State	N=76 2nd grade students	W	Ind.	Natl.	.63, TR .63, TR .65, TR .68, TR .67, TR .51, TR .56, C
Barger (2003)	NC	N=38 3rd graders	W	Gp.	State	.73, TR
Buck and Torgesen (2003)	13 schools from one FL district	N=1103 3rd grade students 49% female, 83% White, 7% African American, 6% Hispanic, 1% LEP, 19% Special Education, 46% FRL	W	Gp.	State	.70, TR
Colon and Kranzler (2006)	North Central FL	N=50 5th graders 44% male, 58% Caucasian, 22% African American, 6% Asian, 4% Hispanic, 2% Native American, 8% Other	W	Ind.	Natl.	.741, C .512, C .805, TR .813, C .465, C .832, TR .60, TR .66, TR
Crawford et al. (2001)	Rural school district in Western Oregon	N=51 students in both years of study (2nd and 3rd grade) 95% White, 57% female	W, A	Gp.	State	.87, WI .82, C .73, C .76, C .71, D .78, C .80, C
Deno et al. (1980, 1982)	Study I: Suburban school in St. Paul, MN	Study I: N=33 students in grades 1–5 N=18 regular education students (50% males) N=15 students with Learning Disabilities (73% males)	W	Ind. & Gp.	Natl.	
Fuchs et al. (1982)	Study III: Three urban schools in Minneapolis, MN Midwestern urban elementary school	Study III: N=66 43 regular education students, 23 students with Learning Disabilities in grades 1–6 N=30 English speaking students in grades 1–6 randomly selected from 90 participated in a larger study	W	Ind.	Natl.	.91, WI .89, WI

(continued on next page)

Table 1 (continued)

Author	Region and school type	Sample characteristics ^a	Time ^b	Individual or group adm.	Source of test(s)	Included correlations (r) and score type
Good et al. (2001)	Urban district in the Pacific Northwest (OR)	N=364 3rd graders 6 elementary schools in the district: 37% to 63% FRL 10% minority, 18% at or below poverty	W	Gp.	State	.67, TR
Hintze et al. (2002)	Small urban school in the Northeastern U.S.	N=136 (Grade 2 N=33, Grade 3 N=31, Grade 4 N=34, Grade 5 N=38) 49% male, 48% African American, 52% White School demographics: 47% low income, 33% middle income, 20% high income N=57 (32 males, 25 females) Grade 2 N=19, Grade 3 N=20, Grade 4 N=18	W	Ind.	Natl.	.65, C
Hintze et al. (1997)	Northeastern United States One suburban elementary school	86% received reading instruction in regular education with no additional assistance 82% Caucasian, 9% African-American, 4% Latino, 5% Asian N=1815	W	Gp.	Natl.	.67, C .66, C
Hintze and Silbergitt (2005)	North central United States 7 elementary schools, 5 cohorts of 1st grade students	51% male, 3% Native American, 1% Asian Pacific Islander, 1% Hispanic, 1% Black not Hispanic, 94% White not Hispanic, 5% Special Education, 30% FRL N=310	W, A	Gp.	State	.49, TR .58, TR .61, TR .68, TR .68, TR .66, TR .68, TR .69, TR
Hosp and Fuchs (2005)	Southeastern U.S. Four elementary schools	Grade 1 (N=74), Grade 2 (N=81), Grade 3 (N=79), Grade 4 (N=76) 57%, 56%, 53%, 53% male, respectively	W	Ind.	Natl.	.71, D .91, WI .79, C .86, TR

Table 1 (continued)

Author	Region and school type	Sample characteristics ^a	Time ^b	Individual or group adm.	Source of test(s)	Included correlations (<i>r</i>) and score type
Ketterlin-Geller and Tindal (2004)	Urban school district in the Pacific Northwest	<i>N</i> = 1153 3rd graders 68% White, 5% Asian, 3% African American, 2% Native American, 9% Other, 51% female, 16% special education	W	Gp.	State	.58, C .41, TR
Klein and Jimerson (2005)	School district in Southern CA	<i>Cohort 2</i> : Grade 1 (<i>N</i> = 473), Grade 2 (<i>N</i> = 642), Grade 3 (<i>N</i> = 731) 73%, 64%, 65% Hispanic (Grades 1–3, respectively) 48%, 57%, 56% FRL 56%, 51%, 53% Spanish Home Language	W, A	Gp.	Natl.	.84, TR .74, TR .81, TR .80, TR .77, TR .77, TR .68, TR
Kranzler et al. (1998)	North Central FL	<i>Cohort 3</i> : Grade 1 (<i>N</i> = 759), Grade 2 (<i>N</i> = 600), Grade 3 (<i>N</i> = 709) 74%, 72%, 74% Hispanic (Grades 1–3, respectively) 49%, 48%, 45% FRL 61%, 56%, 59% Spanish Home Language Longitudinal: 1st grade R-CBM to 3rd grade SAT-9: <i>N</i> = 401 District: 24% Caucasian, 71% Hispanic, 55% FRL, 56% Spanish Home Language <i>N</i> = 57 4th graders (28 males, 29 females) 77% White, 19% African American, 4% Hispanic, 23% low income	W	Ind.	Natl.	.41, C
Kranzler et al. (1999)	Public elementary school in north central FL	<i>N</i> = 326 general education students Grades 2–5 (<i>N</i> s: 84, 76, 94, 72, respectively) 69% Caucasian, 24% African American, 49% female, English primary language of all students	W	Gp.	Natl.	.63, C .52, C .54, C .51, C
Marston and Deno (1982)	Minneapolis, MN	<i>N</i> = 26 3rd grade students	W	Gp.	Natl.	.59, V .84, W1 .88, C

Author(s)	Study Description	W	Gp.	State	Effect Size	
McClintchey and Hixson (2004)	7 years of data from 1 elementary school in urban MI. One year all students in the district included.	Total 4th grade N across years=1,362 Yr 1 N=139, Yr 2 N=68, Yr 3 N=64, Yr 4 N=843, Yr 5 N=61, Yr 6 N=73, Yr 7 N=55, Yr 8 N=59	W	Gp.	State	.90, TR
						.84, D
						.77, TR
						.69, TR
						.74, TR
						.63, TR
						.49, TR
						.65, TR
						.81, TR
						.76, TR
McIntosh, Graves, and Gersten (2007)	Large southern CA school district	District: 52% non-Caucasian, 60% FRL N=59 for the longitudinal data	W, A	Ind.	Natl.	.51, C
						.73, C
Riedel (2007)	26 schools in Memphis, TN that received a Reading Excellence Act grant	N=1395 92% African American, 85% FRL, 55% female, 4% EL	W, A	Gp.	Natl.	.59, TR
						.49, TR
Roehrig, Petscher, Nettles, Hudson, and Torgesen (2008)	Students in Reading First schools in FL	N=35,207 49% female, 36% White, 36% African American, 23% Latino, 3% multiracial, 1.5% Asian, <1% Native American, 75% FRL, 17% SpEd,	W	Gp.	State & Natl.	.66, TR
						.68, TR
						.68, TR
						.68, TR

(continued on next page)

Table 1 (continued)

Author	Region and school type	Sample characteristics ^a	Time ^b	Individual or group adm.	Source of test(s)	Included correlations (<i>r</i>) and score type
		12% EL, 3% Gifted				.71, TR .71, TR .67, TR .69, TR .68, TR .68, TR .70, TR .70, TR .58, V .58, D .66, C .66, TR .63, V .66, D 71, C .71, TR .62, V .59, D .70, C 71, TR .63, V .63, D .74, C .74, TR .63, V .64, D .75, C .74, TR .61, V
Schilling, Carlisle, Scott, and Zeng, 2007; Carlisle et al., 2004	9 School districts, 49 schools, in MI Reading First	<i>N</i> =2970 1st graders, <i>N</i> =2,884 2nd graders, <i>N</i> =3,130 3rd graders	W, A	Gp.	Natl.	

Author	Study Description	W	Gp.	State & Natl.	Effect Size
Shapiro et al. (2006)	2 school districts in Eastern PA		Gp.	State & Natl.	.62, D
					.67, C
					.70, TR
					.58, V
					.61, D
					.65, C
					.67, TR
					.56, V
					.60, D
					.63, C
					.65, TR
					.68, TR
					.69, TR
					.67, TR
Shaw and Shaw (2002)	District 1, moderate size, urban and suburban		Gp.	State	.65, TR
					.66, TR
					.67, TR
					.25, TR
					.64, TR
					.62, TR
					.72, TR
					.54, D
					.67, V
					.67, C
Shaw and Shaw (2002)	District 2, Suburban		Gp.	State	.71, TR
					.53, D
					.63, V
					.67, C
					.70, TR
					.52, D
					.64, V
					.65, C
					.73, TR
					.73, TR

(continued on next page)

Table 1 (continued)

Author	Region and school type	Sample characteristics ^a	Time ^b	Individual or group adm.	Source of test(s)	Included correlations (r) and score type
Shinn et al. (1992)	Mostly White public school district in mid-size northwestern city	N=238 Grade 3 N=114, Grade 5 N=124 49% female; 96% received instruction in general education	W	Gp.	Natl.	.80, TR .57, C .58, C .69, D .58, C .60, C .59, D .60, C .55, C .49, D .62, C .54, C .48, D .63, TR .75, TR .61, TR .62, TR .65, TR .76, TR .76, TR .71, TR .68, TR .65, TR
Sibley, Bivier, and Hesch (2001)	IL	Group 1: N=112 5th graders, Group 2: N=114 6th graders District: 15% special education, Mostly white, middle to upper middle class, 7% FRL, 4% minority	W, A	Group	State & Natl.	
Silberglitt, Burns, Madyun, and Lail (2006)	Upper Midwest	N=5, 472 51.5% male, 2.3% Native American, 1.4% Asian, 1.0% Hispanic, 1.0% Black, 94.3% White Grade 3 N=3165, Grade 5 N=3283 FRL range across districts: 5.7% to 18.63% 7th and 8th grade data not included in table	W	Gp.	State	
Silberglitt and Hintze (2005)	Five rural and suburban districts in the Upper Midwest	N's range from 1441 to 2126	W, A	Gp.	State	.47, TR .57, TR

						.60, TR
						.66, TR
						.67, TR
						.68, TR
						.70, TR
						.71, TR
						.65, WI
						.75, C
						.64, D
						.50, TR
Sofie and Riccio (2002)	West Central AL	N=40 20 students referred (ages 6–8) for reading disability	W	Ind.	Natl.	
		65% male, 90% White, 10% African American				
		20 students not referred with average achievement				
		40% male, 95% White, 5% African-American				
		N=276				
Speece and Ritchey (2005)	3 schools in a suburban district in the mid-Atlantic states	2 cohorts of 1st grade students in 2 consecutive academic years	W	Ind.	Natl.	
		57% male, 13% African American, 11% Asian American, 55% European American, 16% Hispanic, 5% Multiracial, 1% Other				
		N=173 4th graders				
		54% male, 6% SpEd				
Stage and Jacobsen (2001)	One elementary in Puget Sound	District: 90% European American, 5% Hispanic, 2% Native American, 1% African American, 15% FRL	W	Gp.	State	.43, TR
		N=772				.43, TR
		Grade 1 N=198, Grade 2 N=191, Grade 3 N=183, Grade 4 N=66, Grade 5 N=68, Grade 6 N=66				.44, TR
Tindal and Manston (1996)		43.5% White, 35.4% African American, 11% Asian, 7.4% Native American, 2.8% Hispanic	W	Gp.	Natl.	.81, V
						.73, V
						.75, V
						.30, C
						.77, V
						.20, C
						.87, V
						.63, C
						.87, V
						.68, C
						.73, V

(continued on next page)

Table 1 (continued)

Author	Region and school type	Sample characteristics ^a	Time ^b	Individual or group adm.	Source of test(s)	Included correlations (<i>r</i>) and score type
Uribe-Zarain (2007)	11 schools in DE that participated in Reading First for 2 or more yrs.	<i>N</i> = 852 3rd grade students 61% FRL, 52% female, 18% SpEd, >6% LEP	W	Gp.	State	.64, C .75, V .64, C .52, TR
VanDerHeyden, Witt, and Naquin (2003)	Rural community in southern LA	<i>N</i> = 182 in 1st and 2nd grade, 42% male School: 46% FRL, 85% Caucasian	W	Gp.	Natl.	.70, TR
Vander Meer, Lentz, and Stollar (2005)	3 elementary schools in suburban school district in southwest OH	<i>N</i> = 364 students 3 of 5 schools from a suburban district of 8800 students	W, A	Gp	State	.65, TR .63, TR .65, TR .65, TR .612, TR .61, TR .71, TR .57, TR
Wiley and Deno (2005)	Urban elementary school in St. Paul, MN	Grade 3 <i>N</i> = 36, Grade 5 <i>N</i> = 33 80%, 58% EL, 3rd and 5th grade, respectively <i>N</i> = 241 3rd grade students	W	Gp.	State	.741, TR
Wilson (2005)	AZ		W	Gp.	State	

Note. Table only includes data specific to grade levels included in the study.

^a EL = English Learner, FRL = Free or Reduced Lunch, IEP = Individualized Education Program, SpEd = Special Education.

^b W = measures administered within academic year, A = measures administered across academic years.

Table 2
Tests, subtests, and groupings for analyses.

Comprehension tests and subtests	Decoding subtests
CAT Reading Comprehension	ITBS Word Analysis
DRP	MAT-8 Sounds & Print
Gates–MacGintie Comprehension	SAT Word Study
ITBS Reading Comprehension	SAT Phonetic Analysis
MAT-6 Comprehension	SAT-7 Word Study Skills
MAT-8 Comprehension	SDRT Word Attack
PIAT Comprehension	WJ-R Word Attack
SDRT Literal Comprehension	WRMT Word Attack
SDRT Inferential Comprehension	
SDRT Reading Comprehension	Total scores
WJ-III Passage Comprehension	Comprehensive Test of Basic Skills
WJ-R Passage Comprehension	GRA+DE
WJ III Reading Fluency ^a	Gates–MacGintie Total Reading
WRMT Comprehension	ITBS Total Score
WRMT-R Passage Comprehension	MAT-6 Total Reading
SAT Comprehension	MAT-8 Total Reading
SAT-7 Reading Comprehension	NWEA Levels Tests (RIT Scores)
TerraNova — 2nd ed. reading subtest ^b	SAT Total Reading
	SAT-9
Vocabulary subtests	SAT-10
ITBS Vocabulary	SDRT
CAT Vocabulary	WJ-R BRS
MAT-8 Vocabulary	K-TEA Reading Subtest
SAT Vocabulary	WJ-III Broad Reading Cluster
	WRMT-Revised Total Score
Word identification subtests	WRMT-Revised Basic Skills
SAT Reading Words	State tests: Arizona, Delaware,
WJ-R Word Identification	Colorado, Illinois, Ohio, Oregon,
WJ III Letter Word Identification	
WRMT Word Identification	
WRMT-R Word Identification	

Note. Tests were grouped according to what the type of skill score the test purported to produce and how authors described the use of the particular test/subtest. Information regarding state tests of reading achievement may be found at the respective state's Department of Education.

CAT=California Achievement Test (CTB/McGraw-Hill, 1985; www.ctb.com); Comprehensive Test of Basic Skills (Comprehensive Tests of Basic Skills, CTB/McGraw-Hill, 1983; www.ctb.com) DRP=Degrees of Reading Power (Koslin, Koslin, Zeno & Ivens, 1989); Gates–MacGintie Total Reading (MacGintie, Kamons, Kowalski, MacGintie, & McKay, 1978); GRA+DE (Williams, 2001); K-TEA=Kaufman Test of Educational Achievement (Kaufman & Kaufman, 1985); ITBS=Iowa Tests of Basic Skills (www.education.uiowa.edu/itp/itbs/); MAT=Metropolitan Achievement Test (Prescott, Balow, Hogan, & Farr, 1984; pearsonassess.com); PIAT: Peabody Individual Achievement Test (Dunn & Markwardt, 1970); NWEA=Northwest Evaluation Association (www.nwea.org); SAT=Stanford Achievement Test (Madden, Gardner, Rudman, Karlsen, & Merwin, 1973; pearsonassess.com); SDRT=Stanford Diagnostic Reading Test (Karlsen, Madden, & Gardner, 1976); TerraNova (CTB/McGraw-Hill, 2003; www.ctb.com); WJ-R=Woodcock Johnson–Revised (Woodcock & Mather, 1990); WJ III=Woodcock Johnson III (Woodcock, McGrew, & Mathew, 2001); WRMT=Woodcock Reading Mastery Test (Woodcock, 1973), WRMT-R=Woodcock Reading Mastery Test–Revised (Woodcock, 1987).

^a In this case, the WJIII Reading Fluency test was placed in the comprehension test/subtest category due to the nature of the task: students are required to make a semantic judgment on the accuracy of each test item in the subtest.

^b Used by the school district as a comprehension measure (Riedel, 2007).

for analysis were coded according to length of time between the administration R-CBM and other measures as follows: within an academic year, across two academic years, and across more than two academic years.

Grade level

The grade in which R-CBM was administered required several codes. In addition to codes for grades one through six, it was necessary to code combinations of grades (e.g., grades 1–3 and grades 4–6) as well as an “other” category when children from a broad combination of grades (e.g. grades 1–5 or grades 1–6) were administered a single-level passage.

Sample characteristics

Initially, we had hoped to examine the association between R-CBM and other standardized tests of reading achievement as a function of racial–ethnic background, students receiving special education services, and those eligible and not eligible for Free or Reduced Lunch subsidies. However, some studies reported demographic data for the school or district in which the data were collected but not demographic data specific to the sample included in the study, whereas others reported it for the sample but not specific to grade-level, which was central to the purpose of the meta-analysis. In both of these cases, sample characteristics specific to the students yielding the correlation coefficients would have been inferred from these other data. It was decided that this analysis would not be appropriate given the data that were available; however, descriptive information for the students included in each study can be found in Table 1.

Interrater agreement

Initial coding was completed by the first author for all studies and correlations included in analyses. Because more than one correlation coefficient was often reported within each article or report, information associated with 73 correlations (25% of all correlation coefficients in the study) was randomly selected by the second author to examine inter-rater agreement. The second author coded this information independently of the first author and calculated the inter-rater agreement for each coding category. Inter-rater agreement was calculated by dividing the number of agreements by the number of agreements plus disagreements, and multiplying the result by 100 to obtain a percentage. Inter-rater agreement across all coded categories met or exceeded 90%. Inter-rater agreement values across the seven categories coded were 100% for source of test, administration format, and grade level categories; 96% for length of time; 95% for R-CBM correlation coefficients; and 90% for sample characteristics and type of criterion score. For categories where 100% agreement was not reached, the two authors discussed the discrepant cases, and both independently recoded the data. After the second round of recoding, all categories had inter-rater agreement of 100%.

Statistical analyses

Modeling and study artifacts

A random-effects meta-analysis model (Hedges & Olkin, 1985; Hedges & Vevea, 1998; Hunter & Schmidt, 1990) was used to evaluate the study results. The random-effects model

has been recommended as a more appropriate approach to combining information across studies in meta-analytic research than the traditional fixed effects model (Hunter & Schmidt, 2000; National Research Council, 1992; Schmidt & Hunter, 2003). One of the main reasons that the random-effects model was chosen over a fixed-effects model for this research was that the latter assumes that correlations between the R-CBM scores and criterion reading test scores are fixed in the population and thus constant. This assumption seems implausible for the current analysis as numerous measures of R-CBM and criterion reading measures were considered and one would expect the population effect size to randomly vary from study to study.

The most obvious reason to question the assumption of a constant correlation in the population for the present research was that a diversity of reading measures were examined, which, while assumed to be measuring a similar construct of general reading skill, were not all developed or intended to measure the same aspects of reading. For instance, different state tests measure different standards in different ways and nationally developed tests have different approaches to measuring reading. Therefore, it was difficult to assume that there would be a fixed and constant correlation between all the different measures. With numerous and diverse measures of reading, it seems very difficult to assume that all the tests measure reading in the same way and with similar content to the same degree of precision and with the same degree of validity. Therefore, a fixed correlation in the population was regarded as an overly strict and tenuous assumption.

The use of a random-effects model allowed for heterogeneous variation underlying the correlations between R-CBM scores and those from criterion tests of reading across studies rather than assuming a single underlying validity as in the fixed effects model. Using this model allowed for the correlations between R-CBM and criterion measures to differ across studies, and should therefore capture the underlying variation in correlations that would result from different R-CBM measures and different criterion measures used across the studies in the sample. In addition, the random-effects model allows for generalizations beyond the studies included in this analysis (Field, 2001), which was deemed important because of the numerous R-CBM measures and the numerous tests of reading skills considered.

An important consideration for this study was effectively handling the study artifacts that can affect the reported correlation estimates within studies (Hunter & Schmidt, 1990). It was suspected that three main artifacts would need to be addressed: sampling error, errors in measurement of both R-CBM and criterion reading measures, and restriction of range. However, during the compilation of studies and review of the reported materials, we were unable to identify much of the necessary information to correct for unreliability and restrictions of range. For instance, only 7 of 41 studies (17%) reported any information on the reliability of the R-CBM measures used in the study (either on the study sample or from a publisher). For the criterion measures, 61% of the studies ($n=25$) did not report any reliability information for the criterion measures, and 32% ($n=13$) provided general information from a technical manual or secondary source, such as *Buros Mental Measurements Yearbooks* (e.g., Geisinger, Spies, Carlson, & Plake, 2007) or textbooks (e.g., Salvia & Ysseldyke, 2007). Only 17% ($n=7$) of studies provided specific information regarding any form of reliability for the particular subtests, grade levels, or both. One study provided general information for one criterion measure and specific information for a second criterion measure in the study.

Unfortunately, a number of studies did not provide sufficient descriptive data to account for potential restrictions in range on R-CBM or other standardized measures of reading achievement. Twenty-nine percent ($n=12$) of the studies did not report means or standard deviations for R-CBM, whereas 5% ($n=2$) reported partial data (e.g., a mean but not a standard deviation), and 15% ($n=6$) of the studies reported data for means and standard deviations for a particular subgroup of the study (e.g., at-risk students or general education and special education students) rather than the overall sample. A similar pattern was found for the other standardized measures of reading achievement. Only 7% ($n=3$) of studies reported means and standard deviations for the population and study sample. Forty-two percent ($n=17$) did not report a mean or standard deviation for either the population or sample, whereas another 42% ($n=17$) reported means and standard deviations for the sample but gave no information about the population. In some cases, the population values were known or readily accessible (e.g., Woodcock–Johnson III; Woodcock, McGrew, & Mathew, 2001), while others were more difficult to locate (e.g., state-specific tests). Seven percent ($n=3$) of studies gave partial information for the criterion population and sample and 2% ($n=1$) provided population information but did not report sample mean and standard deviation.

Therefore, not enough information was reported in the studies to correct for these two important artifacts, either at the individual study level or with respect to general artifact distributions (Hunter & Schmidt, 1990). Attempts to make corrections on the reported correlations would have had to have been based on somewhat capricious “estimates” by the authors or based on the potentially biased subsample of studies with the pertinent and necessary information. It was decided not to correct for these artifacts, and report the lack of consistency in reporting important statistics necessary for the evaluation of the accumulated evidence of studies such as these, which was thought to be an important finding in itself.

The random effect model can be conceptualized as a hierarchical linear model (HLM) and estimated using maximum likelihood methods (Raudenbush & Byrk, 2002; Raju and Drasgow, 2003; Konstantopoulos & Hedges, 2004). Estimates of sampling error were dealt with in the following explication of the random-effects model within the context of the general HLM. The HLM model for meta-analysis is a two-level model with level 1 (L1) modeling within-study variation and level 2 (L2) modeling between-study variation (Raudenbush & Bryk, 2002). The two-level model is,

$$d_j = \delta_j + e_j \quad (1)$$

$$\delta_j = \gamma_0 + \sum_s \gamma_s W_{sj} + u_j, \quad (2)$$

where d_j is the observed effect size measure in the j th study, δ_j is the population correlation, and e_j is the random sampling error. We assume e_j is normally distributed with zero mean and variance, V_j (which is independent across studies). The L2 model of Eq. (2) indicates that the study correlations might vary based on the moderator variables, W_{sj} . Thus, the L2 model is a regression model in which the study correlation is the criterion and the moderator variables are the predictors with regression coefficients, γ_s , indicating the strength of association. γ_0 is the intercept term and u_j is the random L2 error, assumed to have zero

mean and non-zero variance, τ^2 . The HLM allows for differences in sampling error across studies when assessing the between-study variability and moderator variable effects.

Transformation of correlation coefficients

Further specification of the model was needed, as the effect size of interest, the correlation coefficient, can be treated in various ways (see Hunter & Schmidt, 1990). We chose to transform the reported correlations from the Pearson product-moment metric to Fischer's z metric using the following transformation:

$$z = \frac{1}{2} \ln \left[\frac{1+r}{1-r} \right],$$

where r indicates the observed correlation, and \ln represents the natural logarithm.

The z -transformation approach was chosen for three reasons. First, given that it was impossible to correct the reported correlations for study artifacts, any slight bias from the use of the z -transformation was thought to be minimal. This decision was not meant to imply that this transformation was more accurate than applying the corrections for study artifacts, but it was only a practical consideration given the lack of evidence provided in the studies to actually correct the correlations for the effects of two important artifacts. Second, the transformation has the effect of making the distribution of correlations more normal, which has advantages when pooling across studies and assessing publication bias (Duval & Tweedie, 2000). Finally, the original intention of the z -transformation was to provide an estimate of the standard error based solely on the sample size. Using this estimate of the standard error allowed for estimating the regression parameters by fixing the L1 error term for each study independently (Raudenbush & Bryk, 2002). Each study sampling error was estimated using the following formula for the standard error:

$$V_j = 1 / (n_j - 3),$$

where V_j is the L1 error variance, and n_j is the sample size of the j th study.

Given the above rationale, $d_j = z_j$ in the L1 model outlined above, the observed effect size for the j th study was the z -transformed Pearson product-moment correlation, and the L1 variance was assumed known from the sample size of the study. All HLM statistical analyses used the z -transformed correlations, but nearly all results (except publication bias) are reported using Pearson r for ease of interpretation.

Testing for moderating variables

The examination of moderator variable effects was performed in two stages. The first stage used only the Total Reading Scores as the criterion reading measures. For this analysis, the effects of the following moderating variables were investigated individually: grade level (grades 1 to 5), administration type (individual vs. group) for only the national tests, test type (state-specific test vs. national) for only the group-administered tests, and the length of time between tests (within vs. across academic years). The analysis of grade level was done by using the 3rd grade as the reference group and dummy coding 4 dichotomous variables to reflect grades 1, 2, 4, and 5.

The investigation of administration type was restricted to national tests, as there were no state-specific tests that were individually administered. Therefore, the investigation of the

moderating effect of state-specific and national tests was restricted to only the group administered tests. It was thought to be inappropriate to investigate the effects of national versus state tests without partitioning out the potentially more reliable individually administered tests from the national tests. For the analysis of length of time between assessments, all across-year coefficients were grouped. An *a priori* family-wise error rate was set at $\alpha = .05$. With the seven multiple tests outlined above, a Bonferroni corrected *p*-value of $\alpha^* = .007$ was used for the individual test values.

The second stage used only the scores on the subtests measuring specific areas of reading skill (e.g., Vocabulary and Comprehension). It was necessary to exclude Total Reading Scores, which are typically composites of subtest scores, to ensure the correlation coefficients were independent. For this analysis, the Comprehension subtests were used as the reference group, and dummy coding was used for the other three domains: Vocabulary, Word Identification, and Decoding. An *a priori* family-wise error rate was set at $\alpha = .05$. With the three multiple tests outlined above, a Bonferroni corrected *p*-value of $\alpha^* = .02$ was used for the individual test values.

Publication bias

Two methods were used to assess publication bias, a trim and fill analysis based on funnel plots (Duval, 2005; Duval & Tweedie, 2000), and a file drawer analysis (Orwin, 1983; Rosenthal, 1979). An example of a funnel plot based on Fisher's *z* transformation is shown in Fig. 1. In the figure, precision in the form of the inverse of the standard error ($1/SE(z)$) is plotted against Fisher's *z* for the total score of the score type correlations (see

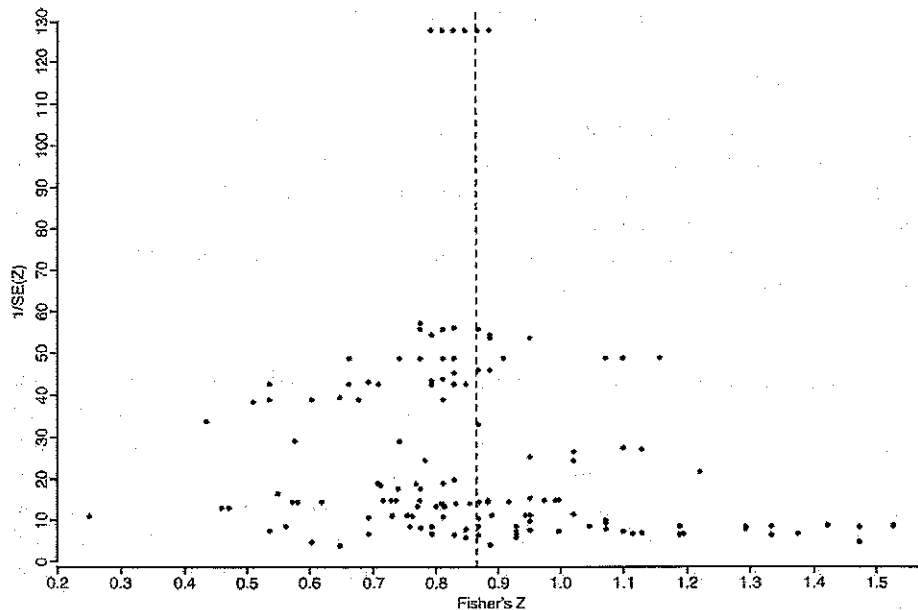


Fig. 1. Funnel plot of precision ($1/SE(z)$) as a function of Fisher's *z* for the score type of total score. Note: The vertical dashed line is the mean *z* used as a basis for evaluating symmetry (see text).

Table 3). The vertical dashed line is the mean value based on the 154 Fisher's z values. Sampling variability is a function of precision, which gives rise to the funnel-like shape of the points in Fig. 1 (wider at the bottom and narrower at the top). Publication bias is indicated by asymmetry about the mean value and typically is a result of missing values in the lower left of the funnel plot as low-valued non-significant correlations tend not to be published (Begg & Berlin, 1988).

Trim and fill analysis uses an iterative algorithm that assesses the asymmetry in the funnel plot (the trim part) and imputes the "missing" Fisher's z values to create a symmetric funnel plot (the fill part). If no missing values are imputed, this is an indication of statistically sufficient funnel plot symmetry and minimal publication bias. When missing values are imputed, the mean z value and confidence interval are re-estimated based on the observed and filled-in values. An important index is the difference in the mean z value for the initial funnel plot and the mean z value for the filled-in funnel plot. Smaller differences

Table 3
Study correlations and publication bias results.

Main effect	Initial analysis						Trim and fill analysis						File drawer	
	N	P25	P50	P75	Mean r	LCI	UCI	Mean r	LCI	UCI	Missing studies	Diff. r	N_f	$5N+10$
All types	289	0.61	0.68	0.74	0.68	0.67	0.69	0.65	0.64	0.66	49	0.03	1649	1455
Grade														
1	32	0.57	0.66	0.72	0.68	0.64	0.71	0.68	0.64	0.71	0	0	174	170
2	61	0.63	0.7	0.76	0.71	0.69	0.72	0.69	0.67	0.71	4	0.01	353	315
3	97	0.62	0.68	0.73	0.67	0.66	0.68	0.65	0.64	0.66	18	0.02	559	495
4	48	0.59	0.65	0.76	0.67	0.64	0.7	0.67	0.64	0.7	0	0	258	250
5	32	0.53	0.64	0.69	0.62	0.59	0.66	0.59	0.55	0.63	5	0.03	168	170
6	4	0.59	0.62	0.66	0.62	0.56	0.67	0.62	0.56	0.67	0	0	21	30
1–3 Comb	3	0.64	0.65	0.75	0.68	0.57	0.77	0.68	0.57	0.77	0	0	17	25
Other	12	0.68	0.77	0.86	0.77	0.72	0.82	0.76	0.7	0.81	1	0.01	77	70
Score type														
Comprehension	72	0.58	0.66	0.74	0.67	0.64	0.69	0.64	0.62	0.66	8	0.02	414	370
Vocabulary	27	0.59	0.63	0.73	0.66	0.63	0.68	0.63	0.6	0.65	5	0.03	138	145
Word identification	11	0.69	0.87	0.89	0.83	0.76	0.88	0.83	0.76	0.88	0	0	76	65
Decoding	25	0.52	0.61	0.7	0.61	0.59	0.63	0.61	0.58	0.63	1	0	128	135
Total score	154	0.64	0.68	0.74	0.69	0.68	0.69	0.65	0.64	0.66	37	0.04	893	780
Time between assessments														
Within year	260	0.62	0.68	0.74	0.68	0.68	0.69	0.66	0.65	0.67	37	0.02	1499	1310
Across years	20	0.61	0.63	0.68	0.64	0.61	0.66	0.63	0.61	0.66	1	0.01	106	110
Across years (2 or more yrs)	9	0.5	0.58	0.66	0.6	0.54	0.65	0.6	0.54	0.65	0	0	40	55
Type of test administration														
Individual	48	0.69	0.77	0.86	0.77	0.73	0.81	0.76	0.72	0.8	3	0.01	302	250
Group	241	0.6	0.67	0.72	0.66	0.66	0.67	0.64	0.63	0.65	29	0.02	1374	1215
Test type														
National	218	0.61	0.69	0.76	0.69	0.68	0.7	0.67	0.66	0.68	26	0.02	1252	1100
State-specific	71	0.61	0.66	0.69	0.64	0.63	0.66	0.63	0.62	0.65	12	0.01	401	365

Note. P25=25th Percentile, LCI=lower confidence interval bound, UCI=upper confidence interval bound, Diff. r =difference in mean r , N_f =failsafe N .

indicate minimal publication bias whereas larger differences indicate greater bias. The trim and fill method is based on symmetry assumptions, so that the funnel plot should be based on Fisher's z rather than Pearson's r (Duval & Tweedie, 2000). In addition, one must condition on the grouping variables, so that a trim and fill analysis will be performed for each group defined in Table 3.

In a file drawer analysis, the focus is on the number of non-significant studies it takes to nullify the overall effect (Rosenthal, 1979). More specifically, one estimates the number of studies in which the null hypothesis, $H_0: \rho=0$, was not rejected that are required to render the mean result in the meta-analysis to be statistically non-significant. This number is referred to as the failsafe N , denoted as N_f (Orwin, 1983). A large value of N_f represents minimal bias as this implies very many non-significant studies would have to be in the file drawer and thus, missing from the meta-analysis. Conversely, a small N_f represents greater bias as the meta-analysis results are vulnerable to being nullified based on a few file drawer studies.

Results

Preliminary results and overall correlational results

A forest plot of all the correlations used in the meta-analysis may be found in Fig. 2. The filled symbols are the values of Pearson's r , the vertical lines indicate the 95% confidence intervals, and the dotted horizontal line is the overall mean. Additional information is shown in Table 3 listed by study characteristics (main effects), including the total number of correlations (N), the quartiles, and the mean with its associated confidence interval limits. Fig. 2 shows the variability among the correlations, and Table 3 provide a general picture of the results based on the study characteristics.

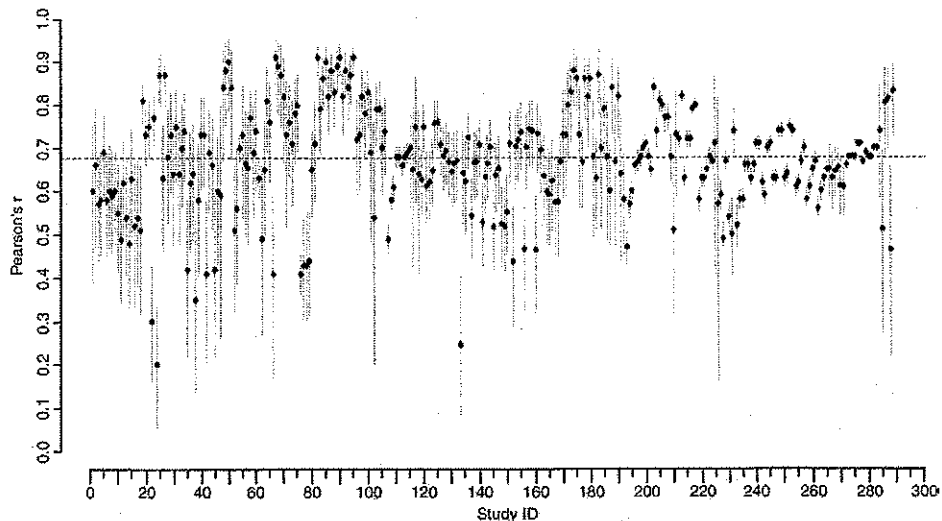


Fig. 2. Forest plot of all correlations. Note: Filled symbol is Pearson's r , vertical line is the 95% confidence interval, and the dashed horizontal line is the mean.

The weighted average and standard deviation (SD) were computed using the formula found in Hunter and Schmidt (1990). The median correlation coefficient across all 289 coefficients reported in the studies was .68 with an interquartile range of .61 to .74 (see Table 3). These results were very similar to the weighted average, .67, and SD, .06. Variability was found across the levels of most of the main effects with the means and medians quite similar in magnitude and the interquartile range and standard deviation similarly wide. Most correlation coefficients tended to be in the .60 to .70 range with a few outliers, such as .55 for across years validity with the criterion test being given 2 or more years after the R-CBM measures, or a .75 on the individually administered tests.

Results reported in Table 3 are the Pearson product-moment correlation estimates from the studies themselves and might differ from the results reported in the following evaluation of main effects of possible moderating variables for two main reasons. First of all, the Fisher z transformation was used for the analysis of main effects, thus the transformation back to the Pearson correlation metric might be slightly biased upwards (Hunter & Schmidt, 1990). Second, some main effects analyses were run on a subset of the total number of factors. For example the comparison of nationally derived or state specific tests was only done at the level of Total Reading Score and, thus, coefficients from subtests on the nationally available tests would be found within the computations in Table 3, but *not* in the evaluation of main effects.

Comparison of state-specific and national group-administered tests

A statistically significant estimate of the z -transformed correlation on the state-specific tests was found ($N=70$), $\gamma_0=0.77$, $t(139)=46.92$, $p<.001$. This confirmed evidence that R-CBM was a significant predictor of state-specific tests of reading standards. There was also a significant positive increase related to the national tests ($N=71$), $\gamma_1=0.18$, $t(139)=4.56$, $p<.001$. The reliability of estimating the population z -transformed correlation coefficient was found to be .81. Significant variation was also found between studies over and above sampling error and the test type moderator variable, $\chi^2(139)=2668.84$, $p<.001$. Thus, the expected correlation coefficient with national tests of .74 was found to be significantly higher than the correlation coefficient of .65 with state-specific tests. It should be noted that the estimate for the national tests was based on about one-third of the total number of coefficients noted in Table 3, which suggested that the ratio of subtests for the national tests to Total Reading Scores was about 2:1 in the dataset. In addition, significant variability between studies remained, suggesting that other potential moderating variables might be found. A forest plot of the correlations and confidence intervals for this effect is shown in Fig. 3.

Comparison of individual and group-administered national tests

A statistically significant estimate of the z -transformed correlation coefficient on the individually administered tests was found ($n=13$), $\gamma_0=1.18$, $t(81)=20.10$, $p<.001$. This confirmed evidence that R-CBM was a significant predictor of state-specific tests of reading standards. There was also a significant negative estimate, reflecting a significant decrease in the magnitude of the correlation, related to the group-administered tests

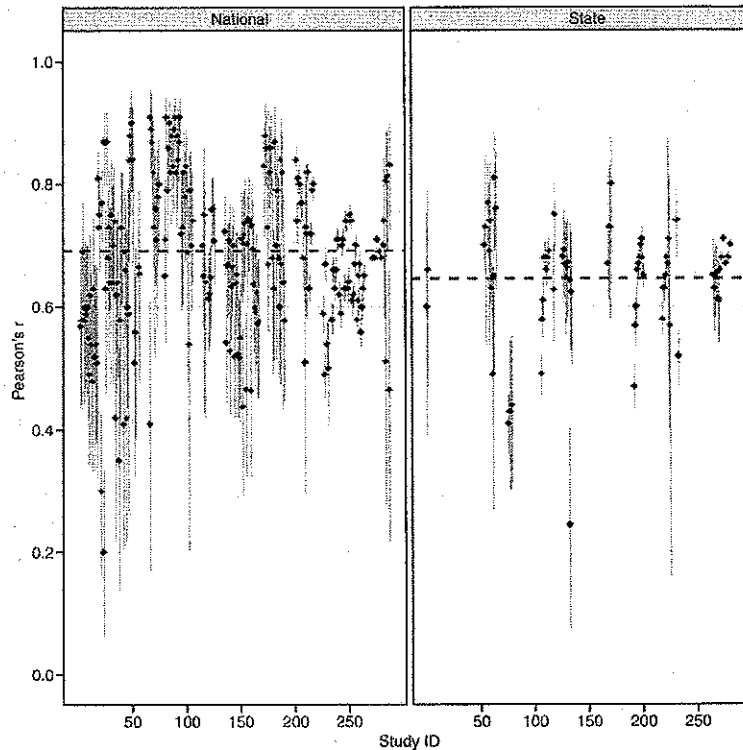


Fig. 3. Forest plot of state/national main effect. Note: Filled symbol is Pearson's r , vertical line is the 95% confidence interval, and the dashed horizontal line is the mean.

($N=70$), $\gamma_1 = -0.29$, $t(139) = -4.59$, $p < .001$. Significant variation was also found between studies above both sampling error and the test type moderator variable, $\chi^2(81) = 1571.53$, $p < .001$. Thus, the expected correlation coefficient of the individually administered tests of .83 was significantly higher than the correlation coefficient of .71 with group administered tests. In addition, significant variability between studies remained. One explanation for these results is that individually administered tests may have higher reliability estimates than group-administered tests, and therefore, due to the lack of correction for unreliability, might have produced higher correlation coefficients. A forest plot of the correlations and confidence intervals for this effect is shown in Fig. 4.

R-CBM and Total Reading Scores by grade

A statistically significant estimate of the z-transformed correlation coefficients for the 3rd grade students was found ($n=57$), $\gamma_0 = 0.87$, $t(147) = 34.02$, $p < .001$. This results confirmed evidence that R-CBM was a significant predictor of third grade reading outcomes. There were no statistically significant differences (all $p > .007$) for the first grade ($n=22$; $p = .28$), second grade ($n=34$; $p = .53$), fourth grade ($n=24$; $p = .29$), and fifth grade

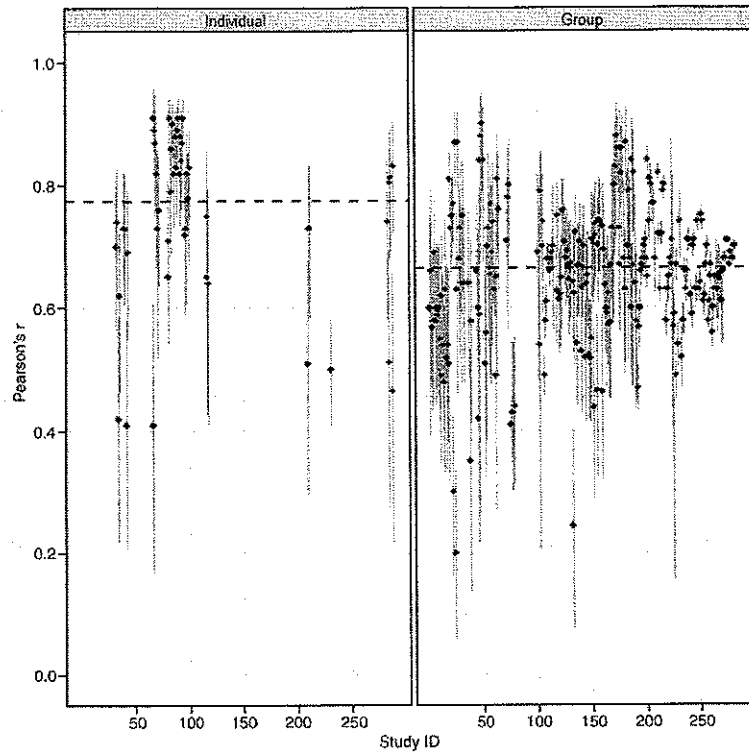


Fig. 4. Forest plot of individual/group main effect. Note: Filled symbol is Pearson's r , vertical line is the 95% confidence interval, and the dashed horizontal line is the mean.

($n=15$; $p=.04$) students. These results suggested that the magnitude of the correlation between grades was not significantly different. The forest plot of the correlations and confidence intervals for this effect is shown in x Fig. 5. Even though no significant differences between grades were found, there was still significant between-studies variation, $\chi^2(147)=3,026.64$, $p<.001$. The expected correlation coefficient across grades 1 to 5 was .70, and significant variability suggested other potential moderating variables might be found.

Length of time

A statistically significant estimate of the z-transformed correlation coefficient for within an academic year was found (current grade) ($n=126$), $\gamma_0=0.88$, $t(152)=51.60$, $p<.001$. This confirmed prior evidence that R-CBM was a significant predictor of reading skills when the criterion test was taken within the same academic year. There was a statistically significant difference for the across-academic-years ($n=28$) estimates, $\gamma_1=-0.14$, $t(152)=-3.58$, $p=.001$. This negative estimate suggested that there was a significant decrease in the magnitude of the correlations when the time span between R-CBM and criterion

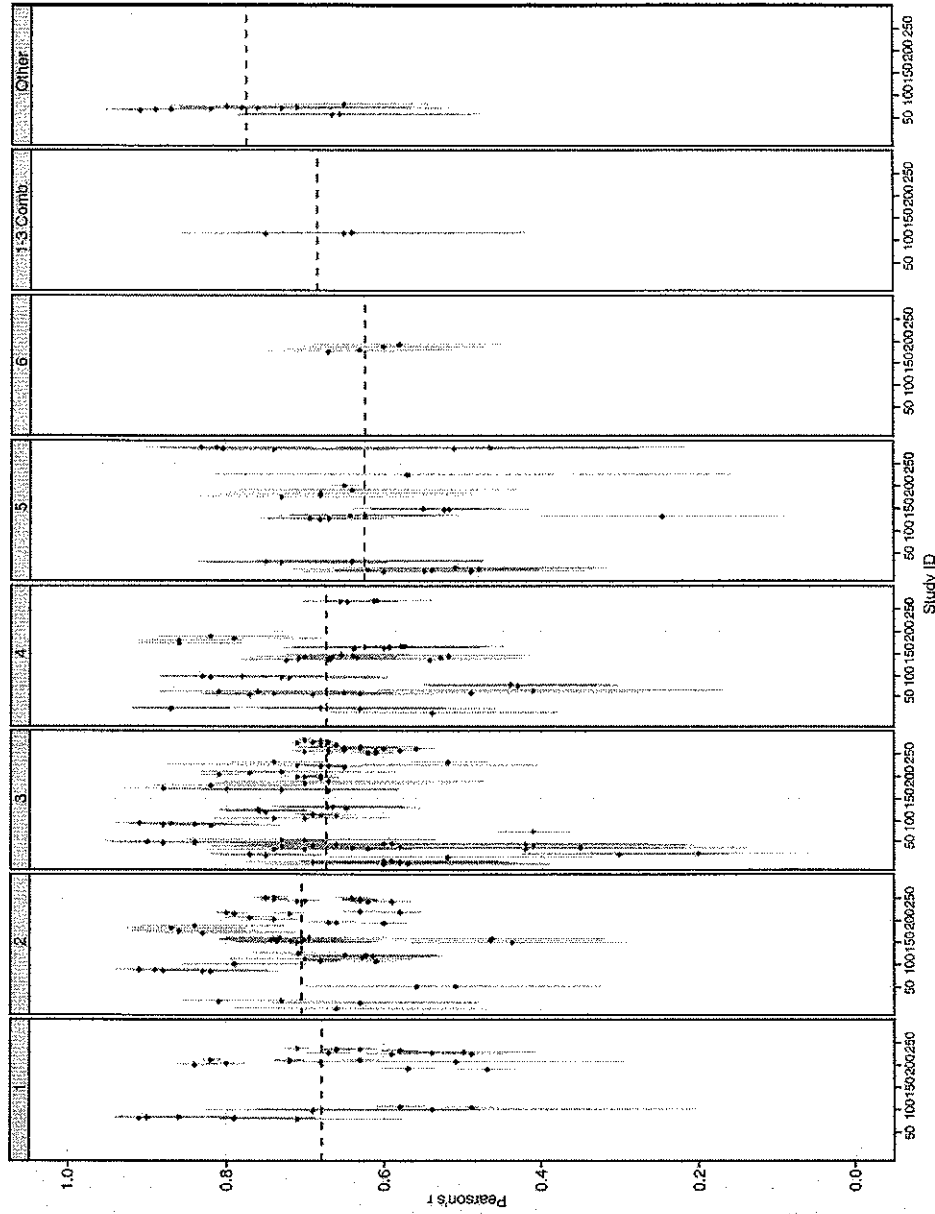


Fig. 5. Forest plot of reading score by grade main effect. Note: Filled symbol is Pearson's r , vertical line is the 95% confidence interval, and the dashed horizontal line is the mean.

measures were obtained, which comports to the common finding that correlations tend to decrease in magnitude when the time span between measurement occasions increases. Statistically significant variability was also found between studies, $\chi^2(152)=2,544.05$, $p<.001$. Thus, the expected correlation coefficient for the current grade was .71. A significantly lower correlation coefficient of .63 (which included across-year tests taken at least 1 year after the R-CBM) was found for all across-year studies. These results were expected because correlation coefficients tend to decrease in magnitude as the time between assessments increases. The forest plot of the correlations and confidence intervals for this effect is presented in Fig. 6.

Individual and group-administered reading subtest scores

A statistically significant estimate of the z-transformed correlation coefficient for the Comprehension subtests ($N=72$) was found, $\gamma_0=0.80$, $t(131)=31.01$, $p<.001$. This result indicated that R-CBM was a significant predictor of reading comprehension. There were no statistically significant differences ($p>.02$) found for the Vocabulary ($n=27$; $p=.96$) and

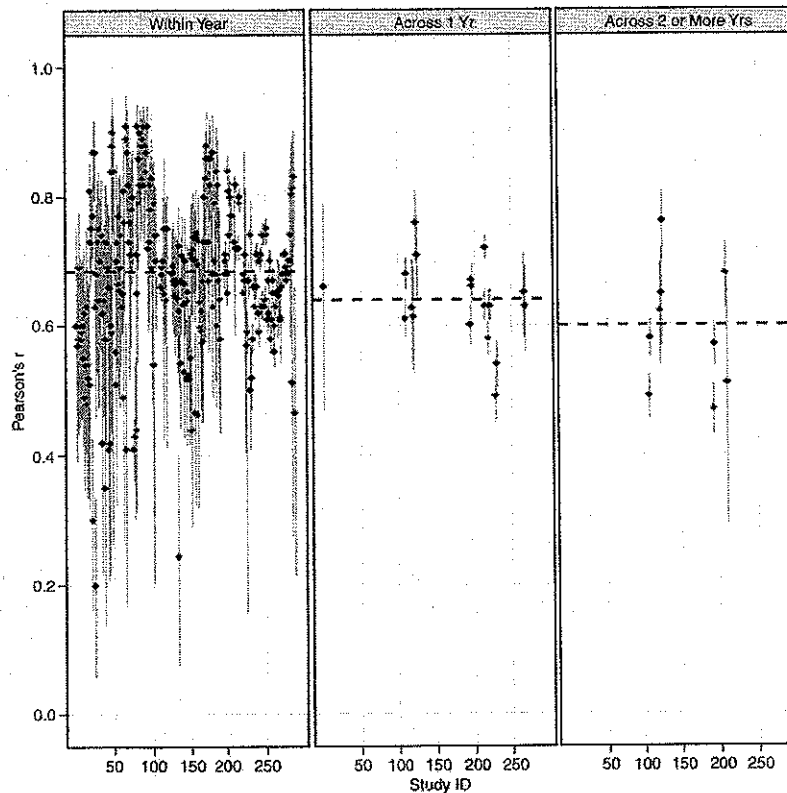


Fig. 6. Forest plot of length of time main effect. Note: Filled symbol is Pearson's r , vertical line is the 95% confidence interval, and the dashed horizontal line is the mean.

Decoding ($n=25$; $p=.10$) subtests, which indicated that R-CBM tended to correlate as highly with these domains or reading skills as it did with Comprehension. However, there was a statistically significant increase found for the Word Identification ($n=11$) subtests, $\gamma_2=0.36$, $t(131)=4.71$, $p<.001$, which suggested that R-CBM tends to be more highly correlated with isolated word reading skills than Comprehension, Vocabulary, and Decoding measures. Statistically significant between studies-variation was also found, $\chi^2(131)=1,109.19$, $p<.001$. Thus, the expected correlation coefficient for Comprehension, Vocabulary, and Decoding subtests were all found to be .66 and was .82 with Word Identification subtests. The forest plot of the correlations and confidence intervals for this effect is presented in Fig. 7.

Publication bias

The results of the trim and fill and file drawer analysis are shown in the right-hand columns of Table 3. All results are expressed in terms of Pearson's r for ease of interpretation (recall the trim and fill analysis is based on Fisher's z). For the trim and fill analysis, the number of missing studies is 0 in a number of instances, indicating a relatively symmetric funnel plot and evidence of minimal bias. For the cases in which the number of missing studies is greater than 0, there is very little difference in the estimated mean r between the initial and imputed funnel plots, with the largest difference being 0.04 for the score type of total score. The small mean difference indicates that in general, the funnel plots are relatively symmetric and represent minimal bias.

The failsafe N (N_f) for the file drawer analysis is shown in the second to last column of Table 3. There is some debate as to what criterion should be used to evaluate N_f . A stringent criterion indexes minimal bias as the case when $N_f > 5N + 10$ (Mullen, Muellerleile, & Bryant, 2001). Based on this criterion, there were seven instances in which $N_f < 5N + 10$, indicating potential bias. However, even in the most extreme case, which is Grade 6 ($5N + 10 - N_f = 9$), N_f was not much smaller than the cutoff criterion. A less stringent criterion indexes minimal bias, as in the case when $N_f > N$ (McDaniel, Rothstein, & Whetzel, 2006). Table 3 shows that in all instances $N_f > N$, providing evidence of minimal bias.

Discussion

There are over three decades of research on the psychometric properties of CBM scores. CBM, originally designed to evaluate the effectiveness of instruction and intervention with individual students, is widely used in education for the purposes of screening, benchmarking, goal-setting, and program evaluation, as well as individual progress monitoring across general, remedial, and special education. One purpose of this study was to quantitatively summarize the correlational evidence of the association between scores derived from R-CBM, the most widely used and researched of the CBM measures, and scores from other standardized measures of reading achievement. A second purpose was to examine potential moderating variables of the correlation coefficients between R-CBM and criterion measures of reading achievement as a function of characteristics of students (grade level, demographics) and the criterion measures (test source, administration format, type of score). The findings from this meta-analysis reveal a very clear picture. The association

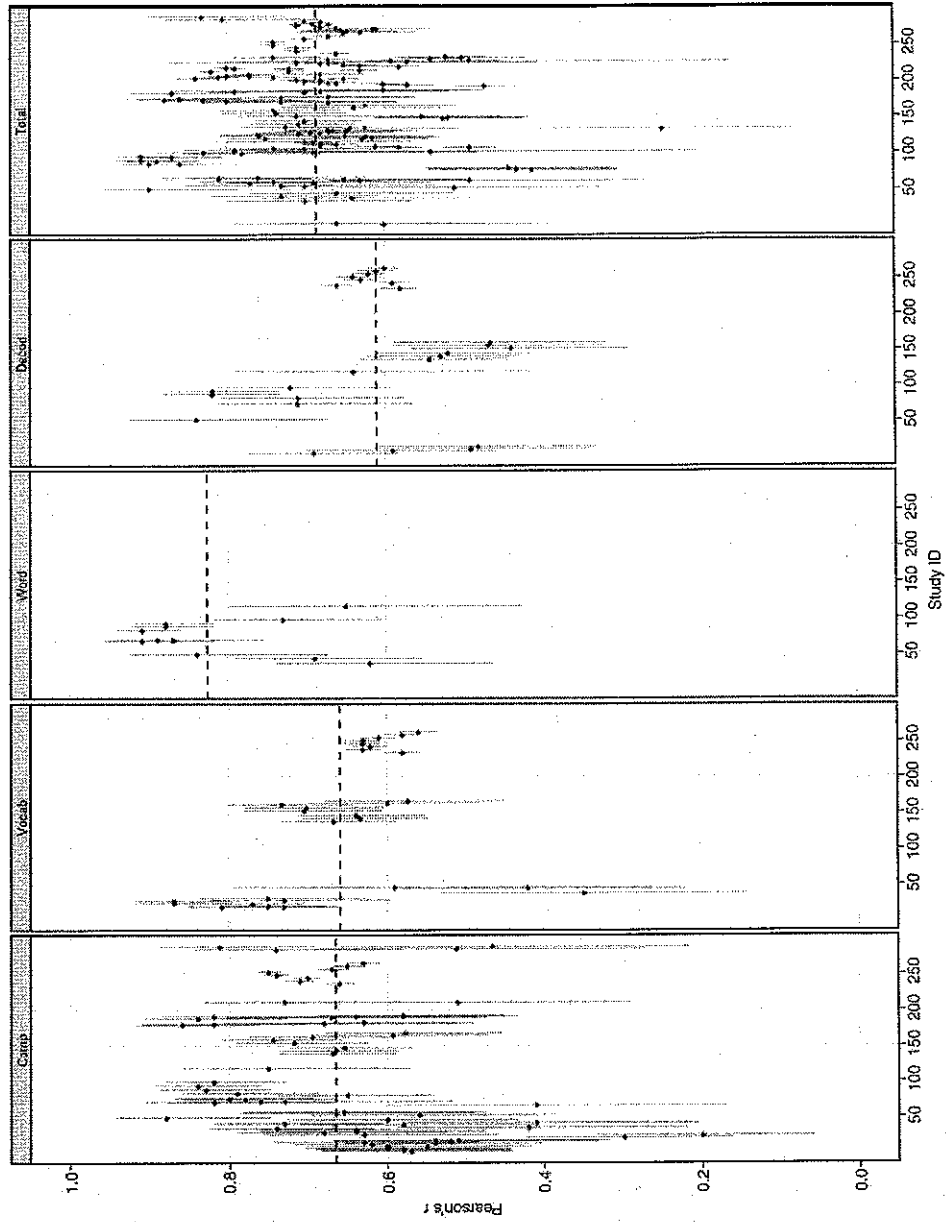


Fig. 7. Forest plot of type of reading score main effect. Note: Filled symbol is Pearson's r , vertical line is the 95% confidence interval, and the dashed horizontal line is the mean.

between scores from R-CBM probes and those derived from other standardized tests of reading achievement is moderately high (weighted average $r=.67$), indicating R-CBM scores function as reasonably good indicators of how well students are likely to perform across a wide range of reading achievement tests. This finding is remarkable when one considers the low cost in terms of resources, both time and financial, required to administer and score R-CBM probes.

The relationship between R-CBM probe scores and scores from other tests of reading achievement varied as a function of the source of test and administration format. The overall correlation coefficient between R-CBM scores and those from individually administered achievement tests was higher than that between R-CBM and group-administered achievement tests. The higher correlation may be due to higher reliability and validity coefficients found for scores derived from individually administered, nationally available and normed tests of achievement, or it may reflect the similarity in administration format. These results provide further support for the use of R-CBM with individual students.

As expected, significant differences were found among state and national group-administered tests of student achievement, with a higher correlation among R-CBM and national tests than with state-specific tests. One reason for this difference may be that the national tests are developed to gauge general reading achievement whereas state-specific tests are designed to assess specific grade-level standards. However, significant between-study variability remained, indicating there are other moderating variables to be discovered. One likely difference in the national and state-specific tests is the higher quality of development, representativeness of the samples, and other technical characteristics of the national tests. Given the varying difficulty levels and quality in the state-specific achievement tests (Peterson & Hess, 2005; Wallis & Steptoe, 2007), quality may account for this between-study variability. Although lower than the correlation with national tests, the association with state-specific tests was significant and moderately strong, which provides additional support for the practice of using R-CBM in general education for screening and identifying students at-risk for future low performance on state standards tests. However, these results also highlight the need for tests with an accumulation of reliability and validity evidence to support their use in high-stakes assessments and decisions about students, educators, schools, and districts.

Another notable finding is that the range of correlations between R-CBM scores and those of various reading subtests was relatively small. There were no differences among score types, with the exception of Word Identification. Higher correlations were found among scores from R-CBM probes and scores on subtests of Word Identification. The magnitude of the difference between these correlation coefficients and those derived with R-CBM probe scores and scores from other reading subtests might be a spurious result given the relatively small number of Word Identification subtest correlations contained in the meta-analysis ($n=11$ vs. 25, 27, and 72 for Decoding, Vocabulary, and Comprehension, respectively). Of these 11 correlations, eight were from the Woodcock Reading Mastery Tests (original and revised versions; Woodcock, 1973, 1987), two were from the Woodcock–Johnson III Tests of Achievement (Woodcock et al., 2001) and one was from the original Stanford Achievement test (Madden, Gardner, Rudman, Karlsen, & Merwin, 1973). The fact that no significant differences were evident in the correlations between R-CBM and tests of Comprehension, Decoding, and Vocabulary adds credence to the ability

of R-CBM scores to act as a general outcome measure of overall reading ability and is contrary to one of the primary criticisms of R-CBM—that it appears to measure decoding rather than comprehension or other reading skills (Hamilton & Shinn, 2003). Further, it is consistent with conclusions by Fuchs and colleagues (2001) and other research suggesting that R-CBM is an indicator of reading comprehension (Fuchs & Fuchs, 1986; Shinn, Good, Knutson, Tilly, & Collins, 1992), as well as overall reading proficiency.

The question of how consistent the strength of the relationship is between scores from R-CBM probes and scores from other standardized tests of reading achievement across the elementary grades has not been entirely answered in the present study. Previous research found disparate results with respect to this question (e.g., Hosp & Fuchs, 2005; Jenkins & Jewell, 1993; Kranzler et al., 1999). One hypothesis from research that indicated a decline in the correlation across grades is that there is a change in what the criterion achievement tests measure as students progress through elementary school (Jenkins & Jewell, 1993). However, our results do not indicate that a significant decline occurs in the criterion validity of R-CBM across grades 1 through 6. We cannot be certain of our conclusion on this issue, however, in that there were insufficient data to examine these correlations as a function of both grade-level and type of score.

R-CBM is increasingly used to predict performance on high-stakes assessments. A logical extension of this practice is to predict performance across years in order to inform early intervention efforts. Early identification of those at risk for obtaining non-passing scores on high stakes assessments is particularly desirable given the evidence that suggests many reading difficulties could be prevented (Snow, Burns, & Griffin, 1998; Torgesen, 2000) and the stability of reading difficulties from the second or third grade forward (Juel, 1988). Data from this study showed declines in the magnitude of the correlation across years. However, predictions across two or more academic years were still significant and moderately high, supporting the use of R-CBM as an indicator, or benchmark, of future performance on high-stakes assessments.

We were unable to examine correlations between R-CBM scores and scores of other tests of reading achievement as a function of student demographic characteristics. There appears to be a great deal of variability in the samples of students with which R-CBM has been used (Table 1). However, examinations of the extent to which scores from tests function similarly for individuals of different backgrounds or groups is crucial to establishing the validity of inferences drawn from any test (AERA et al., 1999) and the disparate results of studies that examined predictive bias for R-CBM and tests of reading achievement and comprehension do not allow firm conclusions to be drawn (e.g., Hintze et al., 2002; Klein & Jimerson, 2005; Kranzler et al., 1999).

Finally, the potential effects of publication bias on the meta-analytic results were explored through two separate analyses. A fill-and-trim analysis (Duval, 2005; Duval & Tweedie, 2000) was used to check for systematic publication bias based on funnel plots like the one shown in Fig. 1. The results showed that average Pearson r values adjusted for asymmetry in the funnel plots were very close to the original values. This provides evidence that the funnel plots were reasonably symmetrical, which is indicative of minimal publication bias. Furthermore, we conducted an analysis to address the “file drawer problem” that non-statistically significant results tend not to be published (Rosenthal, 1979). The number of file drawer studies was estimated and evaluated based on two different stringency criteria. The

results show that even when using the more stringent standard, there was no substantial evidence of extensive bias. Taking the bias results collectively, we feel there is relatively strong evidence our meta-analysis results are not unduly tainted by publication bias.

Limitations and future directions

It is necessary to note limitations of this study, as well as directions for future research. A significant limitation was our inability to correct for errors in measurement of R-CBM probes and other reading tests as well as restriction of range in scores. The data necessary to make adjustments to correlation coefficients due to range restriction and instrument unreliability were not available in most of the articles included in this meta-analysis. This finding is disturbing in light of the fact that even the minimum requirement of reporting means and standard deviations of the R-CBM instruments and other standardized tests of reading achievement was missing in a substantial number of studies. In addition, the lack of reporting of reliability in the samples used within the reported studies runs counter to recent recommendations on technical reporting for research in psychology (Wilkinson & Task Force on Statistical Inference, 1999). Reliability could have been computed for criterion measures directly from the item responses. In addition, several studies included multiple R-CBM passages, typically three, in each administration. An estimate of the reliability could have been easily computed by taking the intercorrelations between passages or running a simple factor analysis and computing the reliability directly. This lack of reliability estimates for the diverse samples also precludes any attempts to conduct reliability generalizability studies (Vacha-Haase, 1998).

Another limitation that reflects the current status of this literature is the varying number of correlations included in each study; therefore, some studies and samples contributed several correlations to the meta-analysis, whereas others may have only reported one or two correlations. There is no way to correct for the uneven distribution of the number of correlation coefficients across studies, but it should be noted in the interpretation of these results. In addition, there are numerous potential moderating variables of the association between R-CBM probe scores and scores from other standardized measures of reading achievement. This study focused on characteristics of students (e.g., grade level) and the criterion measures. Future research may also address other potential moderating variables related to R-CBM, such as whether one or three passages were given, source of passages (e.g., publishing company, curriculum), the psychometric aspects of different passage sets (Betts, Pickart, & Heistad, 2009; Christ & Ardoin, 2009), and the use of specific grade-level or universal passages, among other things. In addition, criterion measures were grouped according to what the tests themselves purported to measure. Additional research may examine the stimulus items on these tests and group them according to types of passages or tasks that are used or forms of comprehension (e.g., literal, inferential). This type of analysis in conjunction with examination of grade-level data may illuminate the decline across grade levels observed in some studies.

Finally, future research may also address other limitations to this research, such as empirical examinations of bias for students of varying socioeconomic, language, and racial-ethnic backgrounds. Another important area of research will examine the use of CBM with students who are non-native English speakers. Evidence is accumulating that

scores from R-CBM are a valid indicator of reading achievement and sensitive to growth for English Learner (EL) students in English (Baker & Good, 1995; Kung, 2007; Ramirez & Shapiro, 2006; Wiley & Deno, 2005). Given the growing EL population in schools across the U.S. (U.S. Department of Education, 2006), increased accountability for the performance of EL students through initiative and legislations such as No Child Left Behind, and poor educational outcomes for students who are not native English speakers (e.g., Federal Interagency Forum on Child and Family Statistics, 2007; Perie, Grigg, & Donahue, 2005), research examining whether these measures are reliable, valid, sensitive to growth, and of high utility when used in conjunction with a Problem-Solving Model (Deno, 2005) with this population is needed.

Conclusion

CBM was designed to provide educators with a set of tasks that were reliable, valid, low-cost, and time-efficient indicators of student achievement in core academic areas. In reading, there is remarkable consistency in the relationship between R-CBM and other standardized measures of reading achievement across decades, samples, and various achievement tests. These results are extraordinary when one considers the brevity, availability, and low-cost of R-CBM. Educators should have great confidence in their use of R-CBM as an indicator of students' overall reading achievement.

References²

- 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.
- Alonso, J., & Tindal, G. (2003). *Analysis of reading fluency and comprehension measures for sixth grade students (Tech. Rep. No. 24)*. Eugene, OR: University of Oregon.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- *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. (1995). Curriculum-Based Measurement of English reading with bilingual Hispanic students: A validation study with 2nd grade students. *School Psychology Review, 24*, 561–578.
- *Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame'enui, E. J., et al. (2008). Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. *School Psychology Review, 37*, 18–37.
- *Barger, J. (2003). *Comparing the DIBELS oral reading fluency indicator and the North Carolina end of grade reading assessment (Tech. Rep.)*. Asheville, NC: North Carolina Teacher Academy.
- Begg, C. B., & Berlin, J. A. (1988). Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society, 151*, 419–463.
- Betts, J., Pickart, M., & Heistad, D. (2009). An investigation of the psychometric evidence of CBM-R passage equivalence: Utility of readability statistics and equating for alternate forms. *Journal of School Psychology, 47*, 1–17.

² References marked with an asterisk indicate studies included in the meta-analysis.

- *Buck, J., & Torgesen, J. (2003). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test (FCRR Tech. Rep. No. 1)*. Tallahassee, FL: Florida State University.
- *Carlisle, J. F., Schilling, S. G., Scott, S. E., & Zeng, J. (2004). *Do fluency measures predict reading achievement? Results from the 2002–2003 school year in Michigan's reading first schools (Tech. Rep. No. 1, Evaluation of Reading First in Michigan)*. Ann Arbor: University of Michigan.
- Christ, T., & Ardoin, S. (2009). Curriculum-based measurement of reading: Passage equivalence and selection. *Journal of School Psychology, 47*, 55–75.
- *Colon, E. P., & Kranzler, J. H. (2006). Effect of instructions on Curriculum Based Measurement of Reading. *Journal of Psychoeducational Assessment, 24*, 318–328.
- *Crawford, L., Tindal, G., & Steiber, S. (2001). Using oral reading rate to predict student performance on statewide assessment tests. *Educational Assessment, 7*, 303–323.
- CTB/McGraw-Hill. (1983). *Comprehensive Test of Basic Skills*. Monterey, CA: Author.
- CTB/McGraw-Hill. (1985). *California Achievement Test*. Monterey, CA: Author.
- CTB/McGraw-Hill. (2003). *TerraNova second edition: California Achievement Tests technical report*. Monterey, CA: Author.
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219–232.
- Deno, S. L. (1992). The nature and development of curriculum-based measurement. *Preventing School Failure, 36*, 5–10.
- Deno, S. L. (2005). Problem-solving assessment with Curriculum-based Measurement (CBM). In R. Brown-Chidsey (Ed.), *Problem-solving based assessment for educational intervention*. New York: Guilford Press.
- Deno, S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual*. Reston, VA: Council for Exceptional Children.
- *Deno, S. L., Mirkin, P., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*, 36–45.
- *Deno, S. L., Mirkin, P., Chiang, B., & Lowry, L. (1980). Relationships among simple measures of reading and performance on standardized achievement tests. *Institute for Research on Learning Disabilities Tech. Rep. No. 20*. Minneapolis, MN: University of Minnesota.
- Deno, S., Reschly, A. L., Lembke, E., Magnussen, D., Callender, S., Windram, H., et al. (2009). A school-wide model for progress monitoring. *Psychology in the Schools, 46*, 44–55.
- Dunn, L. M., & Markwardt, F. C. (1970). *Peabody Individual Achievement Test*. Circle Pines, MN: American Guidance Service.
- Duval, S. J. (2005). The “trim and fill” method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment, and adjustments* (pp. 127–144). Chichester, UK: Wiley.
- Duval, S. J., & Tweedie, R. L. (2000). A non-parametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89–98.
- Espin, C. A., Shin, J., & Busch, T. W. (2005). Curriculum-based measurement in the content areas: Vocabulary-matching as an indicator of social studies learning. *Journal of Learning Disabilities, 38*, 353–363.
- Federal Interagency Forum on Child and Family Statistics. (2007). *America's children: Key national indicators of well-being, 2007*. Federal Interagency Forum on Child and Family Statistics, Washington, DC: U.S. Government Printing Office Retrieved on September 6, 2007 from www.childstats.gov
- Field, A. (2001). Meta-analysis of correlation coefficients: A Monte Carlo Comparison of fixed- and random-effects methods. *Psychological Methods, 6*, 161–180.
- Foegen, A., Espin, C. A., Allinder, R. M., & Markell, M. A. (2001). Translating research into practice: Preservice teachers' beliefs about Curriculum-Based Measurement. *Journal of Special Education, 34*, 226–236.
- Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57*, 488–500.
- *Fuchs, L. S., Deno, S. L., & Marston, D. (1982). *Use of aggregation to improve the reliability of simple direct measure of academic performance, Vol. IRLD-RR-94*. (pp. 33): University of Minnesota, Minneapolis Institute for Research on Learning Disabilities.
- Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). Effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal, 21*, 449–460.

- Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation on student achievement. *Exceptional Children, 53*, 199–207.
- Fuchs, L. S., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom based assessment. *School Psychology Review, 28*, 659–671.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement using a reading maze task. *Exceptional Children, 58*, 436–450.
- Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*, 239–256.
- Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*, 20–29.
- Geisinger, K. F., Spies, R. A., Carlson, J. F., & Plake, B. S. (Eds.). (2007). *The seventeenth mental measurements yearbook* Lincoln, NE: Buros Institute of Mental Measurements.
- Glenn, D. (2007). Consultants in reading program had conflicts, report says. *Chronicle of Higher Education, 53* (29), A16.
- *Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance of 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.
- Goodman, K. S. (2006). *The truth about DIBELS: What it is, what it does*. Portsmouth, NH: Heinemann.
- Greenwood, C. R., Dunn, S., Ward, S. M., & Luze, G. J. (2003). The Early Communication Indicator (ECI) for infants and toddlers: What it is, where it's been, and where it needs to go. *The Behavior Analyst Today, 3*, 383–388.
- Grimes, J., & Tilly, W. D. (1996). Policy and process: Means to lasting educational change. *School Psychology Review, 25*, 465–476.
- 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.
- Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.
- Hedges, L., & Vevea, J. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486–504.
- *Hintze, J., Callahan, J., Matthews, W., Williams, S., & Tobin, K. (2002). Oral reading fluency and prediction of reading comprehension in African American and Caucasian elementary school children. *School Psychology Review, 31*, 540–554.
- *Hintze, J. M., Shapiro, E. S., Conte, K. L., & Beslie, 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., & Silbergitt, 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.
- *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.
- Hunter, J., & Schmidt, F. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Beverly Hills, CA: Sage.
- Hunter, J., & Schmidt, F. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*, 275–292.
- *Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children, 59*, 421–432.
- Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80*, 437–447.
- Kamii, C., & Manning, M. (2005). Dynamic Indicators of Basic Early Literacy Skills (DIBELS): A tool for evaluating student learning? *Journal of Research in Childhood Education, 20*, 75–90.
- Kaminitz-Berkooza, I., & Shapiro, E. S. (2005). The applicability of Curriculum-Based Measurement to measure reading in Hebrew. *School Psychology International, 26*, 494–519.
- Kaminski, R. A., & Good, R. H. (1996). Towards a technology for assessing basic early literacy skills. *School Psychology Review, 25*, 215–227.
- Kaminski, R. A., & Good, R. H. (1998). Assessing early literacy skills in a problem-solving model: Dynamic Indicators of Basic Early Literacy Skills. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 113–142). New York: Guilford Press.

- Karlsen, B., Madden, R., & Gardner, E. F. (1976). *Stanford Diagnostic Reading Test*. San Antonio, TX: Psychological Corporation.
- Kaufman, A. S., & Kaufman, N. L. (1985). *Kaufman Test of Educational Achievement*. Circle Pines, MN: American Guidance Service.
- *Ketterlin-Geller, L. R., & Tindal, G. (2004). *Analysis of reading fluency and comprehension measures for 3rd grade students (Tech. Rep. No. 22)*. Eugene, OR: University of Oregon.
- *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.
- Konstantopoulos, S., & Hedges, L. (2004). Meta-analysis. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 281–297). Thousand Oaks, CA: SAGE Publishing.
- Koslin, B. L., Koslin, S., Zeno, S. M., & Ivens, 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.
- Kung, S., (2007). Predicting the success on a state standards test for culturally and linguistically diverse students using curriculum-based oral reading Measures. Unpublished doctoral dissertation. University of Minnesota, Minneapolis.
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293–323.
- Lembke, E. S., Foegen, A., Whittaker, T. A., & Hampton, D. (2008). Establishing technically adequate measures of progress in early numeracy. *Assessment for Effective Intervention*, 33, 206–214.
- Linn, R. L. (2005). *Test-based educational accountability in the era of No Child Left Behind. (Tech. Rep.)*. University of California Los Angeles, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- MacGintie, W. H., Kamons, J., Kowalski, R. L., MacGintie, R. K., & McKay, T. (1978). *Gates-MacGintie Reading Tests*, 2nd Ed. Chicago: Riverside Publishing.
- Madden, R., Gardner, E. R., Rudman, H., Karlsen, B., & Merwin, J. C. (1973). *Stanford Achievement Test*. New York: Harcourt Brace.
- Manzo, K. K. (2005). National clout of DIBELS test draws scrutiny. *Education Week*, 25(5), 1–12.
- Manzo, K. K. (2005). States pressed to refashion Reading First grant designs. *Education Week*, 25(2), 1–25.
- Manzo, K. K. (2007). Ed. Department allowed singling out of 'Reading First' products. *Education Week*, 26(26), 13.
- *Marston, D. (1989). A curriculum based measurement approach to assessing academic performance: What is it and why do it. In M. R. Shinn (Ed.), *Curriculum Based Measurement: Assessing Special Children* (pp. 18–78). New York: Guilford.
- *Marston, D., & Deno, S. L. (1982). *Implementation of direct and repeated measurement in the school setting, Vol. IRLD-RR-106*. (pp. 45): University of Minnesota, Minneapolis Institute for Research on Learning Disabilities.
- Marston, D., Muyskens, P., Lau, M., & Canter, A. (2003). Problem-solving model for decision making with high-incidence disabilities. *Learning Disabilities Research and Practice*, 18, 187–200.
- McConnell, S. R., McEvoy, M. A., & Priest, J. S. (2002). "Growing" measures for monitoring progress in early childhood education: A research and development proves for Individual Growth and Development Indicators. *Assessment for Effective Intervention*, 27, 3–14.
- McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case study of four test vendors. *Personnel Psychology*, 59, 927–953.
- *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.
- *McIntosh, A. S., Graves, A., & Gersten, R. (2007). The effects of response to intervention on literacy development in multiple-language settings. *Learning Disability Quarterly*, 30, 197–212.
- Mirkin, P. K., Deno, S. L., Tindal, G., & Kuehnle, K. (1982). Frequency of measurement and data utilization strategies as factors in standardized behavioral assessment of academic skill. *Journal of Behavioral Assessment*, 4, 361–370.
- Morgan, S. K., & Bradley-Johnson, S. (1995). Technical adequacy of curriculum-based measurement for Braille readers. *School Psychology Review*, 24, 94–103.

- Mullen, B., Muellerleile, P., & Bryant, B. (2001). Cumulative meta-analysis: A consideration of indicators of sufficiency and stability. *Personality and Social Psychology Bulletin*, 27, 1450–1462.
- Naglieri, J. A., & Crockett, D. P. (2005). Response to Intervention (RTI): Is it a scientifically proven method? *Communiqué*, 34(2), 38–39.
- National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (NIH Publication No. 00-4769)*. Washington, DC: U.S. Government Printing Office.
- National Research Council. (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academy of Sciences Press.
- Orwin, R. F. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, 8, 157–159.
- Perie, M., Grigg, W. S., & Donahue, P. L. (2005). *The Nation's Report Card: Reading 2005 (NCES 2006-451)*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Washington, D.C.: U.S. Government Printing Office.
- Peterson, P. E., & Hess, F. M. (2005). Johnny Can Read... in some states. *Education Next*, 5, 52–53.
- Powell-Smith, K. A., & Stewart, L. H. (1998). The use of curriculum-based measurement on the reintegration of students with mild disabilities. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 254–307). New York: Guilford Press.
- Prescott, G. A., Balow, I. H., Hogan, T. P., & Farr, R. C. (1984). *Metropolitan Achievement Tests (Mat-6)*. San Antonio, TX: The Psychological Corporation.
- Pressley, M., Hilden, K., & Shankland, R. (2005). *An evaluation of end-grade-3 Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Speed reading without comprehension, predicting little (Tech. Rep.)*. East Lansing, MI: Michigan State University, Literacy Achievement Research Center.
- Raju, N., & Drasgow, F. (2003). Maximum likelihood estimation in validity generalization. In K. Murphy (Ed.), *Validity generalization: A critical review* (pp. 263–286). Mahwah, NJ: Erlbaum.
- Ramirez, R. D., & Shapiro, E. S. (2006). CBM and the evaluation of reading skills of Spanish-speaking English Language Learners in bilingual education classrooms. *School Psychology Review*, 35, 356–369.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, 2nd Ed. Thousand Oaks: Sage.
- Reschly, D. J., & Bergstrom, M. K. (2009). Response to intervention. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (pp. 434–460), 4th Ed. New York: Wiley.
- *Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly*, 42, 546–567.
- *Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology*, 46, 343–366.
- Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, 86, 638–641.
- Salvia, J., & Ysseldyke, J. (2007). *Assessment in special and inclusive education*, 10th Ed. Boston: Houghton-Mifflin.
- Samuels, S. J. (2007). The DIBELS tests: Is speed of barking at print what we mean by reading fluency? *Reading Research Quarterly*, 42, 563–566.
- *Schilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *Elementary School Journal*, 107, 429–448.
- Schmidt, F., & Hunter, J. (2003). History, development, evolution, and impact of validity generalization and meta-analysis methods, 1975–2001. In K. Murphy (Ed.), *Validity generalization: A critical review* (pp. 31–65). Mahwah, NJ: Erlbaum.
- *Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment*, 24, 19–35.
- *Shaw, R., & Shaw, D. (2002). *DIBELS Oral Reading Fluency-Based Indicators of Third Grade Reading Skills for Colorado State Assessment Program (CSAP) (Tech. Rep.)*. Eugene, OR: University of Oregon.
- Shinn, M. R. (1989). Identifying and defining academic problems: CBM screening and eligibility procedures. In M. R. Shinn (Ed.), *Curriculum Based Measurement: Assessing special children* (pp. 90–129). New York: Guilford Press.

- Shinn, M. R., & Bamonto, S. (1998). Advanced applications of Curriculum-Based Measurement: "Big ideas" and avoiding confusion. In M. R. Shinn (Ed.), *Advanced applications of Curriculum-Based Measurement* (pp. 1–31). New York: Guilford Press.
- *Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*, 459–479.
- Shinn, M. R., Powell-Smith, K. A., Good, R. H., & Baker, S. (1997). The effects of reintegration into general education reading instruction for students with mild disabilities. *Exceptional Children, 64*, 59–80.
- *Sibley, D., Biwer, D., & Hesch, A. (2001). *Establishing Curriculum-Based Measurement Oral Reading Fluency performance standards to predict success on local and state tests of reading achievement (Tech Rep)*. Arlington Heights, IL: Arlington Heights School District 25.
- *Silbergliitt, B., Burns, M. K., Madyun, N. H., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in the Schools, 43*, 527–535.
- *Silbergliitt, B., & Hintze, J. M. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state mandated achievement tests. *Journal of Psychoeducational Assessment, 23*, 304–325.
- Snow, C. E., Burns, S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
- *Sofie, C. A., & Riccio, C. A. (2002). A comparison of multiple methods for the identification of children with reading disabilities. *Journal of Learning Disabilities, 35*, 234–244.
- *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. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research and Practice, 15*, 128–134.
- Swanson, H. L., Trainin, G., Denise, M., Necochea, D. M., & Hammill, D. D. (2003). Rapid Naming, Phonological Awareness, and reading: A meta-analysis of the correlation evidence. *Review of Educational Research, 73*, 407–440.
- Tindal, G. (1989). Evaluating the effectiveness of educational programs at the systems level using curriculum-based measurement. In M. Shinn (Ed.), *Curriculum-based assessment: Assessing special children* (pp. 202–238). New York: Guilford Press.
- *Tindal, G., & Marston, D. (1996). Technical adequacy of alternative reading measures as performance assessments. *Exceptionality, 6*, 201–230.
- Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research and Practice, 15*, 55–64.
- *Uribe-Zarain, X. (2007, February). *Relationship between performance on DIBELS Oral Reading Fluency and performance on the Reading DSTP Year 2005–2006*. Newark, DE: Delaware Education Research & Development Center, University of Delaware.
- U.S. Department of Education, National Center for Education Statistics. (2006). *Public Elementary and Secondary Students, Staff, Schools and School Districts: School Year 2003–04 (NCES 2006-307)*. Retrieved from <http://nces.ed.gov/fastfacts/display.asp?id=96> on September 6, 2007.
- Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*, 6–20.
- *VanDerHeyden, A. M., Witt, J. C., & Naquin, G. (2003). Development and validation of a process for screening referrals to special education. *School Psychology Review, 32*, 204–227.
- *Vander Meer, C. D., Lentz, F. E., & Stollar, S. (2005). *The relationship between Oral Reading Fluency and Ohio Proficiency Testing in Reading (Tech. Rep.)*. Eugene, OR: University of Oregon.
- Wallis, C., & Steptoe, S. (2007, May 24). How to fix No Child Left Behind. *Time* Retrieved January 28, 2009, from <http://www.time.com/time/magazine/article/0,9171,1625192-1,00.html>
- Wayman, M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on Curriculum-Based Measurement in reading. *Journal of Special Education, 41*, 85–120.
- Wesson, C., Deno, S. L., & King, R. (1984). Direct and frequent measurement of student performance: If it's good for us, why don't we do it? *Learning Disability Quarterly, 7*, 45–48.
- *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.

- Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.
- Williams, K. T. (2001). *Technical manual: Group Reading Assessment and Diagnostic Evaluation*. Circle Pines, MN: American Guidance Service.
- *Wilson, J. (2005). The relationship of Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency to performance on Arizona Instrument to Measure Standards (AIMS). *Technical Report* Tempe, AZ: Tempe School District No. 3.
- Wilson, M., Schendel, J. M., & Ulman, J. E. (1992). Curriculum-based measures, teachers' ratings, and group achievement scores: Alternative screening measures. *Journal of School Psychology*, 30, 59–76.
- Woodcock, R. W. (1973). *Woodcock Reading Mastery Tests*. Circle Pines, MN: American Guidance Service.
- Woodcock, R. W. (1987). *Woodcock Reading Mastery Tests, Revised*. Circle Pines, MN: American Guidance Services Inc.
- Woodcock, R. W., & Mather, N. (1990). *Woodcock–Johnson Tests of Achievement - Revised*. Ailen, TX: DLM Teaching Resources.
- Woodcock, R. W., McGrew, K. S., & Mathew, N. (2001). *Woodcock–Johnson III Test of Achievement*. Itasca, IL: Riverside Publishing.
- Yeh, C. (1992). The use of passage reading measures to assess reading proficiency of Chinese elementary school students. Unpublished doctoral dissertation. University of Minnesota.
- Yell, M., Deno, S. L., & Marston, D. E. (1992). Barriers to implementing Curriculum-Based Measurement. *Diagnostique*, 18, 99–105.

