Learning Disabilities: A Contemporary Journal 22(2), 87-122, 2024

Criterion Validity of Early Numeracy Curriculum-Based Measurement: A Meta-Analysis

Soyoung Park*
University of Central Florida, USA

Gena Nelson
University of Oregon, USA

Jaehyun Shin
Gyeongin National University of Education, South Korea

Ben Clarke
Madison A. Cook
Joanna Hermida
University of Oregon, USA

The study examines the validity of four commonly used early numeracy curriculum-based measurement (EN-CBM) tools in relation to math criterion measures. Additionally, the investigation examines the reporting quality of the included studies as related to important features of assessing CBM's technical adequacy. For inclusion in the meta-analysis, research had to investigate and report on the criterion validity of EN-CBM—oral counting, numeral identification, quantity discrimination, missing number—administered to students in preschool, kindergarten, or Grade 1. Twenty-two studies published between 1997 and 2023 met inclusion criteria, with 272 correlations and 4,130 total participants. A meta-analysis with robust variance estimates for correlated and hierarchical effects was conducted to estimate the average weighted correlations between the four EN-CBM tasks and math outcome measures and to identify potential moderating variables (grade level, CBM type, criterion measure type, correlation lag time). Results indicated the average weighted correlation across all EN-CBM was significantly correlated with criterion measures (r =.48; 95% CI [0.430, 0.528]). The average weighted correlation between EN-CBM and criterion measures was significantly smaller for EN-CBM oral counting and state test criterion. Results indicated that grade level and administration lag time did not significantly predict relations between EN-CBM and criterion measures. The quality coding results revealed many opportunities for improving information reporting across all indicators (attrition [M = 0.63], EN-CBM reliability [M = 1.48], scoring reliability [M = 1.24], and administrator training [M = 1.60]). Results are discussed in terms of implications for practitioners who administer EN-CBM in school settings.

Keywords: **curriculum-based measurement, early numeracy, meta-analysis**

*Please send correspondence to: Soyoung Park, PhD, School of Teacher Education, University of Central Florida, 12494 University Blvd., Orlando, FL 32816, United States of America, Email: soyoung.park@ucf.edu.

Introduction

Early Math Screening and Declining U.S. Scores

Given the 2022 U. S. Nation's Report Card recorded a significant decline (-11 points) in math scores from the 2019 assessments, along with persistent low math achievement over the last two decades, there has been a strong emphasis on early identification of students at risk in math (National Center on Education Statistics [NCES], 2022). Of the 29 education systems participating in the Trends in International Mathematics and Science Study (TIMSS) in both 2011 and 2019, the United States was the only one to experience a widening score gap in math between high- and low-performing students (Stephens et al., 2022). Within the multi-tiered system of support (MTSS) framework in schools, screening students early in their academic development is especially important to assess the achievement levels of those who are or may be at risk of math difficulties (MD). To identify students who may be at risk, measures need to be technically adequate, easy to administer, and implemented at regular intervals throughout the academic year (Dong et al., 2023; Gersten et al., 2012). Given its brevity, ease of administration, and strong technical adequacy, curriculum-based measurement (CBM; Deno, 1985) is one of the most common approaches to screening students in early grades (VanDerHeyden et al., 2017). CBM is standardized and efficient, making it an optimal tool within the MTSS framework for making instructional decisions (Fuchs et al., 2021). Early numeracy curriculum-based measurement (EN-CBM) is the most commonly used method of CBM to screen early grade students (kindergarten and Grade 1) for risk in math. Screening across the school year allows educators to identify students who may benefit from intervention and subsequently evaluate students' progress (Dong et al., 2023; Fuchs & Vaughn, 2012); however, this requires that schools have access to technically adequate screening tools to identify risk, especially at school entry.

To determine technical adequacy, screening tools such as CBM and EN-CBM should be assessed for criterion validity. Criterion validity is a key aspect of CBM as it demonstrates how accurately the test measures its intended outcome (Ketterlin-Geller et al., 2019). To demonstrate how accurately CBM assesses their intended outcome, researchers can examine the relation between CBM tasks and other math assessments, such as norm-referenced tests or state tests (Lee & Lembke, 2016). However, establishing criterion validity of CBM, particularly in math, has been challenging because of the multifaceted nature of math skills (Nelson et al., 2023). The technical adequacies of math CBM, such as reliability and validity, have been reported to be weaker (Foegen et al., 2007; Nelson et al., 2023) compared to reading CBM (Clements et al., 2023); an equivalent general outcome measure in CBM for math, like oral reading fluency for monitoring general reading proficiency over time, has never been found (Miciak et al., 2024). To address this issue, CBM can include broader math content to increase its correlation with criterion measures, but the downside is it may be less sensitive to changes in student performance over time (Nelson et al., 2023). Thus, the technical adequacy of single points in time, which refers to Stage 1 research in CBM (Fuchs & Fuchs, 2004), has been most extensively studied to measure CBM's validity and screening accuracy (Foegen et al., 2007; Nelson et al., 2023). Despite these efforts, no meta-analysis has been conducted to synthesize the relation

between EN-CBM and its criterion measures. Given the recent uptick in individual studies reporting on the validity of EN-CBM (Nelson et al., 2023), it is an appropriate time to quantitatively synthesize the latest knowledge on its criterion validity. Thus, the purpose of this meta-analysis is to examine how students' performance on EN-CBM relates to their performance on math achievement criterion measures. Specifically, this meta-analysis focuses on oral counting, numeral identification, quantity discrimination, and missing number forms of EN-CBM. In the sections that follow, we discuss the importance of early numeracy skills development, screening practices in schools, and previous research that has guided the current meta-analysis.

Early Numeracy Curriculum-Based Measurement (EN-CBM)

EN-CBM is used to screen students in preschool, kindergarten and Grade 1 for risk in math by measuring their early numeracy. Early numeracy performance for kindergarteners and first graders has commonly consisted of the four domains of oral counting, numeral identification, quantity discrimination, and missing number. These domains have also been studied as a potential indicator of concurrent and future risks of MD. Hence, researchers have most commonly built screening tools around those skills (Methe et al., 2011b). Oral counting, the ability to verbally count a string of numbers, was measured by Koponen et al. (2019) in kindergarten as predictive of calculation fluency and word problem-solving skills in Grade 4 as well as broad math performance in Grade 7. In addition, oral counting was found to be a strong predictor of later arithmetic fluency (Aragón et al., 2016; Lê & Noël, 2021; Locuniak & Jordan, 2008) and later strategy sophistication (Chu et al., 2018). Students who demonstrated MD (i.e., chronic low achievement in math; Nelson & Powell, 2018) in Grade 1 showed more difficulty with producing the correct count sequence (Geary et al., 2000). Other researchers have also linked lower counting skills in kindergarten with difficulty with number knowledge and mental arithmetic in Grade 1 (Desoete & Grégoire, 2006). Likewise, numeral identification, the ability to identify numbers that correspond with a given quantity, is also a key measure of EN-CBM. A longitudinal study of students from Grades 1 to 4 found that on a task of numeral identification, students with persistent MD performed more slowly compared to students without MD (Vukovic & Siegel, 2010). Chard et al. (2005) investigated number identification with kindergarten students to predict MD and confirmed its predictive and concurrent validity. Quantity discrimination refers to comparing quantities with their symbolic representations. Rouselle and Noel (2007) examined quantity discrimination and compared the instructional efficacy of presenting quantities as numerals compared to sets (e.g., as dots) for at-risk students. They found students with MD were able to distinguish sets and compare the numerosity of the sets just as well as students without MD. However, students with MD were slower and less accurate in comparing numerals than students without MD. Reigosa-Crespo et al. (2012) found efficiency in enumerating sets of dots and number comparison was associated with arithmetical competence and developmental dyscalculia in second to ninth grade students. Finally, missing number refers to the ability to identify a missing number in a given series. Ability to identify missing numbers was found to be a strong predictor of Grade 3 math domains as measured by a state test, such as numbers and operations, algebra, and data analysis (Kiss et al., 2019). Measures that focus on number sequence can

also distinguish between students with persistent MD and students without persistent MD (Vukovic & Siegel, 2010). Failure to acquire one or more of these four early numeracy skills may result in difficulty acquiring more advanced skills, as they are precursors to understanding formal mathematics (Chard et al., 2005; Clarke & Shinn, 2004; Lembke et al., 2008; NMAP, 2008; NRC, 2009).

Meta-Analyses on EN-CBM and Criterion Measures

Criterion validity is important, but to date there has been no meta-analysis conducted to synthesize the relation between EN-CBM and its criterion measures. However, two relevant meta-analyses addressed CBM and criterion measures in math but differed in their focus on math skills and grade bands (i.e., Codding et al., 2023; McCulloch, 2010). McCulloch (2010) reviewed 16 studies focused on early math CBM (EM-CBM) of counting, comparing and ordering quantities, adding/subtracting, mixed skills, and readiness concepts—defined as knowledge about shapes and colors, size, and pattern recognition. The author specifically referred to the measures as EM-CBM due to their focus on math broadly from preschool through Grade 2, as opposed to just early numeracy. McCulloch found a moderate relation between EM-CBM and norm-referenced math achievement measures ($r = .49$). While McCulloch's contribution to summarizing the evidence on the concurrent validity coefficients of EM-CBM is valuable, the findings were limited, as McCulloch only examined EM-CBM criterion validity in relation to norm-referenced math achievement tests; the author did not consider other criterion measures, such as state tests or researcher-developed measures. In addition, McCulloch did not report results related to the four EN-CBM tools (other than counting) that are the focus of the current meta-analysis.

A more recent meta-analysis conducted by Codding et al. (2023) examined the validity of CBM-Mathematics (CBM-M) in relation to math outcomes from Grades 2 to 8. Codding et al. included 29 studies focused on two types of CBM (computation or conceptual knowledge and applications) and their criterion measures. The average correlation was $r = .584$ [95% CI: 533, .635]; the authors concluded scores were strongly correlated with criterion measures. The implications of this study are important, revealing that these two CBM tools are used for screening students in Grades 2 to 8, with emerging evidence of their construct validity and utility in schools. However, Codding et al.'s study was limited to Grades 2-8, unlike the current meta-analysis, which examines early numeracy in preschool to Grade 1. The current meta-analysis differs from and builds on the two previous meta-analysis studies by focusing on four domains of early numeracy and specific grade bands from kindergarten to Grade 1.

Potential Moderators of EN-CBM and Criterion Measures

Previous findings across meta-analyses are mixed regarding the significance of various moderators of the relation between CBM and criterion measures. The moderators that have been investigated previously with other types of CBM may have practical implications for early numeracy screening methods in schools. Given the recent uptick in individual studies reporting on the validity of EN-CBM (Nelson et al., 2023), it is an appropriate time to quantitatively synthesize the current evidence base and examine the moderating role of grade level, CBM type, criterion measure type,

and correlation lag time. In addition, considering the recent focus on reporting quality in the field (Appelbaum et al., 2018; Park & Nelson, 2022; Cook et al., 2018; Cummings et al. 2023; Page et al., 2021), it is important to assess the quality of reporting related to attrition, reliability of the EN-CBM, scoring reliability, and administrator training in the included studies. The following section highlights the possible moderators from previous meta-analyses for examining the relation between EN-CBM and criterion measures.

Grade Level

Each grade has different standards for end-of-year expectations in math. Therefore, it is important that researchers and practitioners alike understand whether the relation between EN-CBM and later math achievement is affected by the students' grade level. In their meta-analysis, Codding et al. (2023) reported scores on math computation (MCOMP) and math concepts and applications (MCAP) had a stronger relation with criterion measures in middle and high school compared to elementary school. McCulloch (2010) reported stronger correlations for Grade 1 and Grade 2 EM-CBM compared to preschool. In contrast, two previous meta-analyses focused on CBM in reading reported grade level was not a significant moderator of the relation between CBM and criterion measures (Reschly et al., 2009; Shin & McMaster, 2019).

EN-CBM Type

Although CBM as a screening tool is intended to be brief in administration and scoring, schools who are struggling with access to resources and staff time may need to consider administering fewer EN-CBMs overall as part of their universal screening process. Thus, it may also be important to identify whether a specific EN-CBM has a stronger relation with later math achievement. Codding et al. (2023) reported the relation between a CBM and criterion measure was significantly larger for MCAP than MCOMP. McCulloch (2010) reported EM-CBM focused on counting yielded significantly lower correlations than other measures (i.e., comparing and ordering quantities, adding/subtracting, mixed skills, readiness concepts). In contrast, in a meta-analysis of CBM for reading, Shin and McMaster (2019) reported the type of CBM—oral reading fluency or maze reading—did not significantly moderate the relation between the CBM and criterion.

Criterion Measure Type

The scores from administering a universal screening measure are used to predict the likelihood a student will later pass or fail some kind of criterion. Students who are predicted to fail the criterion are typically identified to receive intervention. In later grades, state tests often serve as a criterion; however, most states do not administer state tests in math to children in preschool through Grade 1. Researchers and practitioners may use other criterion measures, including norm-referenced achievement tests, researcher-developed measures, and other CBM tools. Codding et al. (2023) reported the criterion measure was not a significant moderator of the relation between Grades 2–8 math CBM and criterion measures. Whereas, in a meta-analysis of CBM reading for students in Grades 1–10, Reschly et al. (2009) reported

norm-referenced tests yielded significantly higher correlations with the criterion than state tests did.

Administration Lag Time

Schools often screen students for academic difficulty during the fall, winter, and spring. The scores from universal screening are used to determine a student's level of risk, typically for end-of-year failure on a criterion measure. Screening conducted in the fall has a greater lag time to the end of the year, compared to screening conducted in the winter. Therefore, it is important meta-analyses investigate whether lag time affects the overall strength of the relation between EN-CBM and math achievement to provide recommendations about EN-CBM across different screening periods. In a meta-analysis focused on Grades 2–8 MCOMP and MCAP, Codding et al. (2023) reported statistically larger concurrent correlations (i.e., shorter lag time), as compared to predictive correlations. In contrast, in a meta-analysis of EM-CBM, McCulloch (2010) reported statistically larger predictive correlations on average, compared to concurrent correlations. Authors of meta-analyses focused on reading CBM have reported similar larger average concurrent correlations than predictive correlations (Reschly et al., 2009; Shin & McMaster, 2019) and mixed findings (January & Klingbeil, 2020).

Quality of Studies in Meta-Analysis Research

Since the start of the 21st century, many organizations and professions have given attention to the quality of research and have established their own set of reporting standards (Appelbaum et al., 2018; Cook et al., 2009; Gersten et al., 2005; Thompson et al., 2005). These efforts were based on the consensus that researchers should provide sufficient details—such as their positionality, clear constructs, and robust methodological procedures—to ensure transparency and applicability of review findings (Cook et al., 2018; Cummings et al., 2023; Page et al., 2021). Drawing from the established quality indicators in the field, several researchers have investigated and reported on the quality of reporting within educational research, including in measurement research (Park & Nelson, 2022), intervention research (Jitendra et al., 2016), and meta-analysis research (Nelson et al., 2022).

Quality of reporting should also be examined in criterion validity studies, as some aspects of study reporting quality may impact readers' interpretation or generalization of the results. For example, according to experts, information gathered from the administration of CBM is "only as good as the corresponding data collection and scoring procedures" (Gersten et al., 2005, p. 160). Other experts have stated it is critical for all authors to examine the reliability of the data for their sample, given that participants across studies vary (Thompson & Vacha-Haase, 2000). Yet, among the previous meta-analyses focused on CBM research, only McCulloch (2010) investigated issues of quality. McCulloch investigated examiner training and scoring reliability and reported differences in training did not yield significantly different correlations with quality, whereas variations in how procedural integrity was reported yielded significantly different correlations.

Purpose and Research Questions

With the recent increase in the number of studies focused on EN-CBM (Nelson et al., 2023), now is an appropriate time to conduct a comprehensive review of EN-CBM criterion validity studies. Performing a meta-analysis allows researchers to report on the convergence of evidence on a topic (Siddaway et al., 2019). A meta-analysis allows the summary of results by calculating an overall average weighted effect size (ES) across studies and exploring study- and sample-related sources of possible heterogeneity in the ESs (Cooper et al., 2019). This approach allows the summary of results by calculating an overall ES estimate and the examination of potential sources of heterogeneity in the ESs through moderator analyses (Pustejovsky & Tipton, 2022). Meta-analyses also provide consumers of research with a report of findings across the literature base (Park et al., 2024). The current meta-analysis quantitatively synthesizes the criterion validity of EN-CBM. The results of the meta-analysis will inform screening practices in schools, and the results of the study reporting quality analysis will inform areas of need in future research. The research questions were as follows:

1. What are the average weighted correlations between EN-CBM and math achievement criterion measures?
2. What variables moderate the relation between students' performance on EN-CBM and criterion measures (i.e., grade level, CBM type, criterion measure type, and correlation lag time)?
3. What is the study reporting quality of the included studies, and does study reporting quality moderate the relation between EN-CBM and criterion measures?

Method

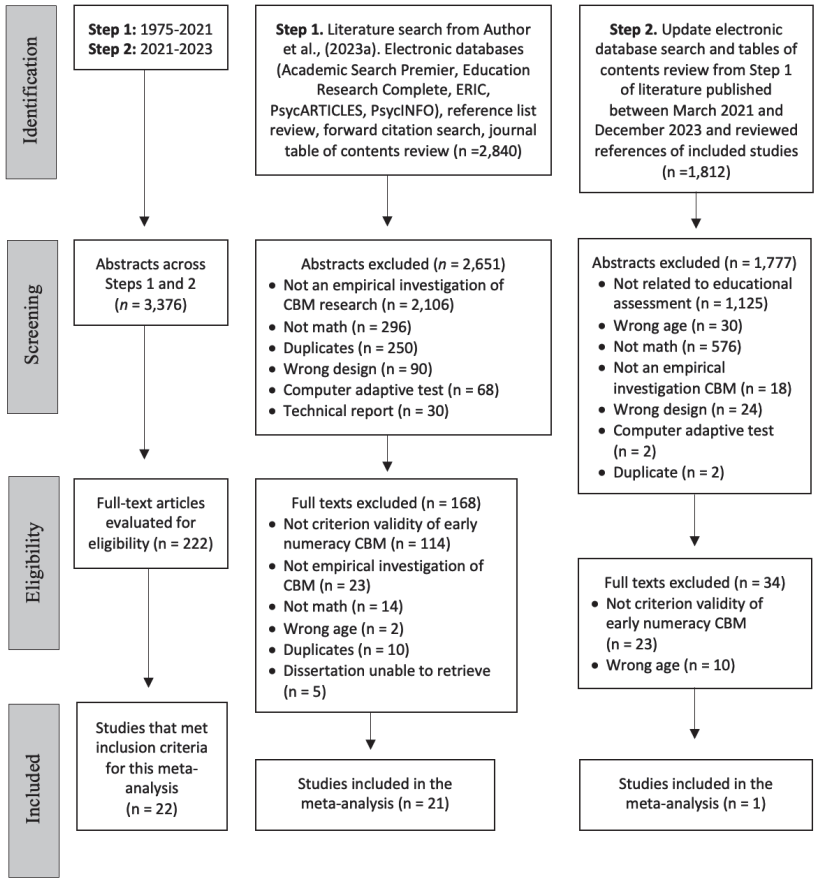
Literature Search Strategies

The literature search for this meta-analysis was conducted as part of a literature search for a broader CBM-M project (Nelson et al., 2023) and was expanded for this meta-analysis. See Figure 1 for a PRISMA diagram that outlines the number of articles retrieved and reviewed at Steps 1 and 2 and reasons for exclusion. Collectively, the search included literature published from 1975 to December 2023.

Step 1: Original Search

The first step involved selecting studies from a corpus of studies identified as part of a literature search in which members of our research team updated the Foegen et al. (2007) review of CBM-M (Nelson et al., 2023). To conduct a comprehensive review of the literature, (Nelson et al., 2023) first conducted an electronic search using Academic Search Premier, Education Research Complete, Educational Resources Information Center, PsycARTICLES, and PsycINFO. Nelson, G., Kiss, A. J., Codding, R. S., McKevett, N. M., Schmitt, J. F., Park, S., & Hwang, J. (2023). searched the databases using the Boolean string: (CBM OR "curriculum based measure*" OR "curriculum-based measure*" OR "general outcome measure" OR "progress monitoring" OR "mastery measure*" OR "curriculum-based assessment*" OR "curriculum based assessment*") AND (math* OR "numeracy" OR "problem solving" OR geometry OR

computation OR algebra OR "proportional reasoning" OR "basic facts" OR fractions OR "concepts and applications" OR "number sense" OR readiness). Nelson, G., Kiss, A. J., Codding, R. S., McKevett, N. M., Schmitt, J. F., Park, S., & Hwang, J. (2023) also searched the online table of contents for two journals, *Assessment for Effective Intervention* and *School Psychology Review*. Nelson, G., Kiss, A. J., Codding, R. S., McKevett, N. M., Schmitt, J. F., Park, S., & Hwang, J. (2023) identified these two journals as having the greatest number of articles that met inclusion. We also reviewed the reference lists of previous studies that included a review of screening and progress monitoring math measures (Christ et al., 2008; Foegen et al., 2007; Gersten et al., 2012; Lembke et al., 2012; Peltier, 2017; Stecker et al., 2005).



Note. From Moher et al. (2009). For more information, visit https://doi.org/10.1371/journal.pmed.1000097

Figure 1. **PRISMA Diagram of Literature Search Procedures and Results**

Finally, we used the Web of Science forward citation search function to find any studies that cited Foegen et al. (2007). Nelson, G., Kiss, A. J., Codding, R. S., McKevett, N. M., Schmitt, J. F., Park, S., & Hwang, J. (2023) originally searched the literature from 1975 to 2021. The authors identified 2,840 studies and then screened the titles and abstracts of all studies. During abstract screening, 2,651 studies were excluded.

Step 2: Updated Search

Nelson et al., (2023) only included studies published between 2005 and 2021. The current meta-analysis conducted an updated search, which expands upon the original search (Step 1) conducted by Nelson, G., Kiss, A. J., Codding, R. S., McKevett, N. M., Schmitt, J. F., Park, S., & Hwang, J. (2023) by searching for studies across a greater time frame (1975 to 2023). First, the current meta-analysis returned to the 189 full texts published between 1975 and 2005 in Step 1 and found three additional studies in Step 2. Second, given the length of time that had passed between the original literature search (March 2021) and the current meta-analysis, we extended the literature search to include the period from April 2021 to December 2023. This included updating the electronic search of the academic databases and reviewing the online tables of contents of the selected journals. The original search was limited to reviewing the tables of contents for two journals, *Assessment for Effective Intervention* and *School Psychology Review*. In this current meta-analysis, we expanded the review to include the tables of contents for the following journals: *Journal of Educational Psychology, Journal of Psychoeducational Assessment, Journal of School Psychology, Psychology in the Schools,* and School *Psychology*. Third, we added one literature search strategy. We reviewed the reference lists of all included articles to determine if any studies might have been missed by other methods. We identified an additional 544 articles to review for abstract screening; from the screening process, we identified 33 articles for full text review; one study met inclusion criteria. Thus, through our collective literature search procedures, we identified 22 studies for inclusion in the current meta-analysis.

Eligibility Criteria

We applied five inclusion criteria to determine eligibility for the current meta-analysis. First, the abstract, purpose, or research question stated the focus of the study was related to CBM research (i.e., technical adequacy of the static score). Studies that reported information about CBM within a different overarching purpose were excluded. For example, we excluded studies that were *primarily* focused on reporting on the effectiveness of an intervention that also happened to report the correlation between EN-CBM and a criterion (e.g., Doabler et al., 2022). The primary focus of the study was determined by reviewing the study's purpose statement(s) and research question(s). This inclusion criterion was retained from the broader CBM-M review (Nelson et al. 2023) in which we were interested in studies that were focused on specifically examining the criterion validity of EN-CBM. Other meta-analyses focused on CBM have applied the same criteria to exclude intervention studies (Codding et al., 2023; January & Klingbeil, 2020; Kilgus et al., 2014; McCulloch, 2010). Moreover, a focus of this paper is coding included studies for quality related to information reported in criterion validity studies. Studies that were not intentionally

focused on investigating or reporting the technical properties of EN-CBM may not have reported information that is often included in studies focused on criterion validity. Second, the study included at least one of the four common EN-CBM. Oral counting was defined as measures that asked students to count verbally as high as possible without error. Missing number was defined as asking students to identify a missing number in a sequence of numbers, such as "8, __, 10." Numeral identification was defined as rapid and accurate detection of individual numerals presented visually in a random order. Quantity discrimination was defined as asking students to identify the larger or smaller value of two numerals. Studies that focused on other EN-CBM (e.g., matching quantities) were excluded. Third, the study reported individual concurrent or predictive correlations between EN-CBM and performance on a math criterion measure. Studies that reported only ranges of correlations across several measures were excluded (e.g., Reid et al., 2006). Fourth, the study included participants in preschool, kindergarten, or Grade 1 at the time of CBM administration and reported correlations separately for different grade levels. Fifth, the results of the study were published in a peer-reviewed journal or as a dissertation. We excluded technical reports from inclusion to avoid bias in seeking technical reports of which we were already aware (i.e., familiarity bias; Borenstein et al., 2011) and to avoid missing technical reports that might not be captured with electronic searches. Nelson, G., Kiss, A. J., Codding, R. S., McKevett, N. M., Schmitt, J. F., Park, S., & Hwang, J. (2023) also excluded technical reports in their search.

Literature Search Training and Reliability

The second author developed the abstract screening tool based on the inclusion criteria and trained research team members to screen titles and abstracts. Training involved clarifying and adding examples and details to the abstract screening tool and independently screening a set of abstracts for practice. Following the initial abstract screening, the research team met a second time to discuss any discrepancies from the practice screening. The second abstract training focused on refining the research team's understanding of the inclusion criteria. All abstracts were screened by two independent coders and reliability of the screening process was 92.0%. Any abstract that was identified by only one author for full-text review was subsequently reviewed by a third research team member to determine whether it should receive full-text review.

To conduct the full-text review, the second author trained members of the research team to use the full-text review protocol. The purpose of the full-text review training was to refine the team's knowledge of the inclusion criteria and complete independent full-text review practice with a set of articles. Discrepancies from the full-text practice were discussed via email. Then, all full texts were independently screened by two research team members. Reliability of the full-text review process was 85.1%. Any full text with a disagreement was discussed between coders in a virtual meeting to come to an agreement about inclusion.

Coding Manual

The second author drafted the coding manual (https://scholarworks.boisestate.edu/sped_facpubs/145/) and included variables related to basic study infor-

mation, participant demographics, EN-CBM, criterion measures, correlations, and study reporting quality. To identify the codes included in the coding manual, the author examined coding manual descriptions from previous meta-analyses focused on reading CBM (January & Klingbeil, 2020; McCulloch, 2010; Shin & McMaster, 2019). The author then determined which codes were needed to answer the research questions that were the focus of the current meta-analysis. For example, we aligned the development of the coding manual to specific moderators we planned to investigate. The code response options in the final coding manual represented a mix of forced response and open response codes.

For basic descriptive information, we coded the year of publication, publication type, journal name, and study location**.** For participant demographics, we coded sample-level demographics, including sample size, age or grade, disability or risk, gender, race or ethnicity, socioeconomic status, and emergent bilingual status. For EN-CBM information, we coded each EN-CBM for measure name and type, administration time in minutes, grade level and time of year of EN-CBM administration, such as the fall, winter, or spring. For the criterion measure information, we coded the specific criterion measure name and whether the measure was a state test, norm-referenced achievement measure, another CBM tool, researcher-developed measure, or other type of test. We also recorded the grade level and time of year of the criterion measure administration. For information related to the correlational evidence of criterion validity, we recorded each correlation between the EN-CBM and criterion measure and the lag time to determine whether the correlation was concurrent or predictive. If correlations were presented for different groups of students (e.g., students with disabilities vs. without disabilities), we recorded them separately.

In addition, we evaluated each study's methodological rigor using a rubric to assess the following quality indicators: (a) attrition, (b) EN-CBM reliability, (c) EN-CBM scoring reliability, and (d) test administrator training. These quality indicators were selected based on recommendations from previous research about quality indicators used to evaluate methodological rigor, specific focus on assessment indicators such as reliability (McCulloch, 2010; Park & Nelson, 2022; riskofbias.info, 2023; Thompson et al., 2005; University of South Australia, n.d.). In addition to extracting general assessment indicators, the quality indicators "procedural integrity" and "examiner training" were also extracted from McCulloch (2010). Along with selecting quality indicators often reported in assessment and outcome measures such as information about reliability (Gersten et al., 2005; Nelson et al., 2022; The Council for Exceptional Children, 2014), the addition of quality indicators from McCulloch was necessary to capture a more complete and relevant picture of study quality.

After we set up the four quality indicators, we assessed the quality of reporting using the following procedure. First, criteria for rating each quality indicator was established. Each quality indicator was then assessed and coded as follows: high quality (3), acceptable quality (2), low quality (1), and very low quality (0). Table 1 provides additional details about the criteria for each possible score. Next, the scores for each quality indicator were averaged across studies (for example, all studies' attrition scores were averaged) to determine overall study reporting quality using the 0-3 scale.

Table 1. ***Study Reporting Quality Related to CBM Studies***

Study Reporting Quality Category	Associated Level of Quality	Description/Criteria
Attrition Reported and Level	High quality	Attrition less than 10%
	Acceptable quality	Attrition between 10-19%
	Low quality	Attrition between 20-29%
	Very low quality	(1) Attrition equal to or greater than 30%; or (2) Attrition Not Reported
EN-CBM Reliability	High quality	Reported and all >.80
	Acceptable quality	(1) Reported and mixed results (some >.80 and some >.70, range included values in .70); or (2) reported and between .70 and .79
	Low quality	(1) Reported and mixed results (some values or range included values in with one greater than <.70 and others less than .70); or (2) reported and between .60 and .69
	Very low quality	(1) Reported and <.60; or (2) Not reported for the *study sample*
Scoring and Administration Reliability	High quality	Scoring reliability reported >.90, and procedural fidelity of administration was documented during training or during test administrations and >.90
	Acceptable quality	Scoring reliability reported >.90, training information reported but no specific IRR information for administration provided (or indicated it was checked but the values were not provided)
	Low quality	(1) Scoring reliability reported and >.90% but no training/procedural information for data collectors documenting reliability to the administration process; or (2) Training was documented with or without IRR but no scoring reliability of protocols
	Very low quality	No scoring or procedural fidelity reliability reported, No training information

Note. Attrition was coded at the study level and applied to each correlation. EN-CBM Reliability and Scoring and Administration Reliability was coded at the correlation level (as many studies reported separate statistics for different measures).

For brevity, we focus on how studies could receive a score of high quality (3) and very low quality (0); refer to Table 1 for details about acceptable quality (2) and low quality (1) scores. We coded whether studies reported attrition, as well as the level of attrition. High quality scores meant attrition was reported as less than 10% and very low quality referred to either (a) attrition was not at all reported or (b) attrition was equal to or greater than 30%. We coded all available EN-CBM reliability information including internal consistency, alternate form, test–retest, and split-half reliabilities. Studies that reported reliability as .80 or greater for the reliability estimates provided in the study received a high-quality score. Very low quality was used to refer to studies that either (a) failed to report any information about reliability for the study sample or (b) reported all reliabilities as less than .60. We also coded whether studies reported if data were double-checked for accurate scoring. High-quality studies reported that a subset of tests or score sheets were double-scored to determine EN-CBM scoring reliability, and very low-quality studies either reported EN-CBM scoring reliability below .70 or did not report any information about EN-CBM scoring reliability. Finally, we coded studies for information related to test administrator training and test administrator's procedural fidelity. Training and procedural fidelity were grouped together given many studies reported on monitoring procedural fidelity as a component of training. High-quality studies met three criteria: They reported (a) data collectors were trained, (b) procedures were in place to monitor test administrators' fidelity to test protocols, and (c) the reliability of the procedural fidelity. Very low-quality studies failed to provide any information about training or monitoring procedural fidelity. Each correlation from each study received a reporting quality score in each of the three areas, as well as an overall average score.

Coding Procedures

After the second author drafted the coding manual, the main coder reviewed the manual and provided feedback prior to coding any practice articles. Then, the author trained the main coder on each of the variables and provided examples of how studies provided information. Next, the author and the main coder coded three practice articles independently to discuss challenges with the coding manual. Once the coding manual was revised, the main coder coded all articles and had weekly check-in meetings with the author until coding was completed. The main coder was majoring in psychology and enrolled in an undergraduate directed research course; they had two semesters of previous experience coding articles for systematic reviews. The three practice articles were part of the 22 included studies and were re-coded as part of the independent coding after the training.

Interrater Agreement

We randomly selected seven studies, representing 29.0% of the total correlations, to double code to determine the reliability of the coding. The second author trained two authors who were doctoral students in a school psychology program to serve as second coders. Discrepancies in coding were discussed and authors came to a consensus on the final code prior to analysis via a discussion with the main coder. Interrater agreement was calculated as (agreements / [agreements + disagreements] × 100). Mean interrater reliability was 90.2%, and with the exception of one code,

the reliability of the codes ranged from 80% to 100%. One code, sample size for correlations, had an average agreement of 70%. When the second author reviewed the discrepancies, it was noticed that some studies had different sample sizes for specific correlations compared to the overall study sample. The low reliability was also related to systematic errors. In other words, if a study included 10 correlations and the wrong sample size was applied to all correlations, the discrepancy was counted 10 times. Because sample size is used as part of determining the weight of the correlation in the meta-analysis, the author checked all sample sizes across all 22 studies to ensure accuracy prior to conducting analyses.

Meta-Analytic Procedures

Calculating Average Weighted Effect Sizes

Prior to the main analyses, we extracted the Pearson correlation coefficients (*r*) and the corresponding sample sizes from each study. Then, we conducted Fisher's *r*-to-*z* transformation to convert all correlation coefficients to *z* scores due to the slight skewness of the correlations. In this meta-analysis, the majority of studies included more than one ES. For example, one study could have included multiple correlations due to including various grade levels, different CBM, and different criterion measures. In addition, many studies included correlations between the EN-CBM and outcome measures across multiple time points, for example correlations for fall, winter, and spring CBM administration time points. Correlations within each study were not independent (Lipsey & Wilson, 2001; Pustejovsky & Tipton, 2022). To address dependent ES estimates, we used a meta-analysis with robust variance estimates (RVE; Hedges et al., 2010) for correlated and hierarchical effects (CHE; Pustejovsky & Tipton, 2022). The RVE method is a preferred approach for considering studies with dependent ES estimates because it accounts for dependency between ESs within a single study, even when the exact dependence structure between ES estimates is unknown (Hedges et al., 2010; Tipton & Pustejovsky, 2015; Pustejovsky & Tipton, 2022). Previous working models have typically assumed that dependent ESs are either correlated or hierarchically related; however, a recent extension of the CHE model offered a better match to the data structure, allowing for both types of dependencies compared to the previous correlated effects model (Pustejovsky & Tipton, 2022). In our meta-analysis, we conducted the RVE to compute the structure of variances among ESs by using a covariance matrix which assumed a correlation of $\rho = 0.8$ among ESs clustered in studies (Pustejovsky & Tipton, 2022). We assumed a 0.8 correlation for RVE and conducted follow-up sensitivity analyses to examine the impact of varying correlations, ranging from 0.4 to 0.8. The results indicated these varying correlation values did not change the ES estimates, indicating the robustness of our findings. Then, the structure of variances was used in a two-level random effects model, with random intercepts for studies and individual effects (Borenstein et al., 2011). We estimated tau-squared statistics and I-squared statistics to assess heterogeneity between studies. Tau-squared statistics is the estimate of the variance in ESs between studies in a random-effects meta-analysis. I-squared statistics represent the percentage of the true variability in a set of ESs that can be attributed to between-study heterogeneity

(Borenstein et al., 2011). To estimate the parameters of the meta-analysis, we applied the restricted maximum likelihood method (Viechtbauer, 2010).

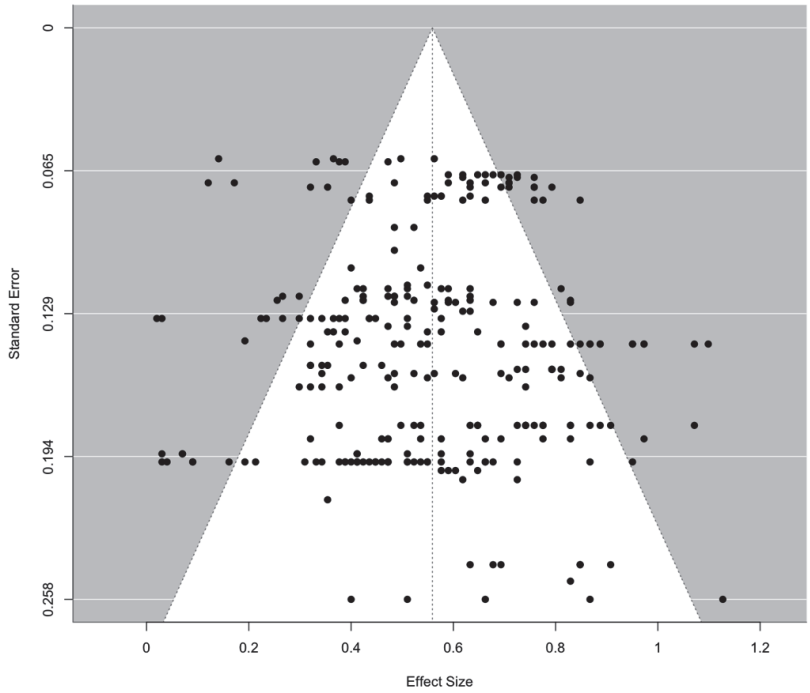
Heterogeneity and Moderator Analyses

Next, we assessed which variables moderate the relation between students' performance on EN-CBM and criterion measures. We conducted moderator analyses to investigate which moderators might explain the between-study heterogeneity on the main effect. We conducted all moderator analyses for EN-CBM and criterion measures separately. No missing data were found for the moderator variables of: grade level, type of CBM, type of criterion measure, and study reporting quality. However, we found a small number of unreported cases for the correlation lag time (n = 4) and the administration time of year (n = 30) within the total sample size of 272 cases. Only completed cases were included in the analysis. After identifying significant moderators from the separate moderator analyses, we built models to see whether a complex model (including multiple moderators) accounted for more variance compared to a simple model. After conducting the analyses, we converted the z scores back to Pearson correlation coefficients for ease of interpretation (Beretvas & Pastor, 2003; Shin & McMaster, 2019).

We then examined the reporting quality of each study and whether study reporting quality moderates the relation between EN-CBM and criterion measures. Each correlation received a rating: not reported or very low quality, low quality, acceptable quality, and high quality on each aspect of study reporting quality—attrition, EN-CBM reliability, scoring reliability, and test administrator training and procedural fidelity. Then, scores across the three aspects were averaged to indicate an overall study reporting quality score (0–3). Thus, each of the three aspects of study reporting quality had equal weight in the overall average quality score. The overall quality score was examined as a moderator.

Publication Bias

Finally, we assessed publication bias using both a funnel plot and Egger et al.'s (1997) regression statistic. A funnel plot shows the relation between ESs and their SE by plotting the effect sizes on the x-axis and standard errors on the y-axis (Sterne et al., 2011). In the absence of publication bias, the plot should form a roughly symmetrical upside-down funnel shape, hence the name "funnel plot." We did not find any asymmetry based on a visual analysis of the funnel plot (see Figure 2).



Note. The symmetrical effect sizes on this funnel plot revealed that there was no publication bias.

Figure 2. **Funnel Plot**

In addition, we used Egger et al.'s regression statistic to test asymmetry in the funnel plot (1997). This statistical test examines the correlation between effect sizes and their standard errors. A significant slope or intercept suggests the presence of publication bias, typically indicated by a *p*-value below .05. Based on Egger et al.'s regression statistic ($t = .84$, $df = 263$, $p = .40$), we did not find any significant asymmetry in ESs. Because significant asymmetry was not found using either the funnel plot or Egger et al.'s regression statistic, we concluded there was no publication bias within this data set (Cooper et al., 2019). Therefore, the included studies provided the representative overall ESs and did not require adjustment for the potential publication bias. We performed all analyses using the clubSandwich (Pustejovsky, 2023) and the metafor packages (Viechtbauer & Viechtbauer, 2015) in R version 4.2.2.

Results

Descriptive Results

We included 22 studies in this meta-analysis published between 1997 and 2023. The majority of the research was conducted in the United States ($k = 21$); one investigation was conducted in Spain. The studies included students in preschool ($k = 3$), kindergarten ($k = 16$), and Grade 1 ($k = 13$), with many studies including students from more than one grade. The total number of participating students was 4,130. The majority of participants were White (47.31%), followed by participants identified as Black (16.03%), Hispanic (3.92%), Asian American (2.69%), other (0.97%), American Indian (0.17%), and multiracial (0.05%). Studies failed to report the race/ethnicity for 28.86% of participants. In addition, 40.14% of participants were identified as female, 40.29% as male, and 19.56% were not reported. Regarding students' linguistic background, 3.41% of students were identified as emergent bilinguals, 63.58% were not identified as emergent bilinguals, and authors failed to report information about 36.42% of participants. Regarding special education services, 4.31% of participants were identified as receiving services, 27.17% were identified as not receiving services, and authors failed to report information for 68.52% of participants. Regarding income level, 23.17% of participants were identified as low socioeconomic status (SES), 43.05% were not identified as low SES, and authors failed to report SES for 33.80% of participants. Too few studies provided information across different demographics to investigate variables as moderators.

We extracted 272 correlations ($M = 12$ per study). Table 2 provides a summary of descriptive statistics for the key variables (average weighted r (SD), frequency, and percentage). The correlations represented preschool ($n = 8$), kindergarten ($n = 132$), and Grade 1 ($n = 132$). The correlations represented the relation between quantity discrimination ($n = 91$), numeral identification ($n = 69$), missing number ($n = 64$), and oral counting ($n = 48$) measures and a criterion measure. Half of the correlations were concurrent (50.7%). The majority of correlations represented the relation between the EN-CBM and a norm-referenced math achievement test (52.9%), followed by state tests (18.4%), other CBM (16.9%), and researcher measures (11.8%).

Table 2. ***Descriptive Statistics of Key Variables***

Variable	*r*	SD	*N*	%
CBM grade level				
Preschool	.39	.16	8	2.94
Kindergarten	.49	.14	132	48.53
First grade	.50	.18	132	48.53
CBM type				
Number identification	.51	.15	69	25.37
Quantity discrimination	.52	.16	91	33.46
Oral counting	.41	.16	48	17.65
Missing number	.51	.15	64	23.53
Criterion type				
State test	.43	.16	50	18.38
Other CBM	.48	.21	46	16.91
Norm referenced	.51	.13	144	52.94
Researcher developed	.53	.17	32	11.76
CBM time of year				
Fall	.52	.15	90	33.8
Winter	.48	.14	32	12.0
Spring	.50	.16	96	36.1
Averagea	.44	.18	24	9.0
Variable	*r*	SD	*N*	%
Criterion grade level				
Preschool	.58	.23	8	2.94
Kindergarten	.48	.15	80	29.41
First grade	.49	.16	126	46.32
Second grade	.52	.17	10	3.68
Third grade	.48	.15	44	16.18
Fourth grade	.57	.14	4	1.47
Correlation lag time				
Concurrent	.52	.15	138	50.74
Predictive	.46	.16	130	47.79
Not reported	.64	.08	4	1.47
Study reporting quality				
Very low	.47	.17	103	51.91
Low	.55	.13	110	48.40
Acceptable	.39	.17	43	0.49
High	.46	.09	16	0.08

a. One study averaged correlations across screening periods.

Average Weighted Correlations Between EN-CBM and Criterion

Based on 272 correlations across 22 studies, the average weighted correlation between the EN-CBM and the criterion measures was r =.48; 95% CI [0.430, 0.528]. Also, the estimated was 2268.814 with = 0.01, indicating significant variability across the studies.

Moderator Analyses

To determine the source of the heterogeneity, we conducted moderator analyses (Table 3) in which we investigated grade level, CBM administration time of year, type of CBM, type of criterion measure, and correlation lag time. We review the non-significant results first. Grade level did not moderate the overall association between EN-CBM and the criterion measures ($F(2, 269)$ = 0.81, p = .44). The heterogeneity statistic of was .0073. Administration time of year did not moderate the overall association between EN-CBM and the criterion measures ($F(2, 215)$ = 0.67, p = .51). The heterogeneity statistic of was .0026. Compared to the concurrent correlation (r = .49), the predictive correlation was slightly smaller, though not to a statistically significant level ($F(2, 263)$ = 0.95, p = .33). The heterogeneity statistic of was .005.

In contrast, both type of CBM and type of criterion measure were significant moderators. CBM type moderated the overall association between EN-CBM and the criterion measures ($F(3, 268)$ = 6.17, p = < .001). The heterogeneity statistic of was .0071. Compared to the average weighted correlation of oral counting (r = .40, intercept), the average weighted correlations for the other CBM (i.e., numeral identification, quantity discrimination, missing number) were significantly larger (all p < .0001). Similarly, criterion types moderated the overall association between EN-CBM and the criterion measures ($F(3, 268)$ = 9.37, p = < .001). The heterogeneity statistic of was .0076. Specifically, compared to the average weighted correlation between EN-CBM and a state test (r = .31; intercept), the average weighted correlations between EN-CBM and the other criterion types were significantly larger for norm-referenced tests (p <.001), other CBM (p <.001), and researcher-developed measures (p <.001).

Study Reporting Quality Results

We also determined reporting quality did not moderate the overall association between EN-CBM and the criterion measures ($F(3, 268)$ = 2.039, p = .109). The heterogeneity statistic of was .0036. Although results of the moderator analysis were nonsignificant, we further explored the study quality results to provide recommendations to improve future reporting of CBM studies.

Table 3. ***Moderator Analyses***

		Correlation			Heterogeneity		Test of Moderators			
Moderator	*k*	*Est.*	*p*	95% CI	*Q(df)*	*p*	*f*	*df1*	*df2*	*p*
Grade level							0.81	2	269	.444
Intercept[a]	126	.477	<.0001	[0.419, 0.531]						
Kindergarten	132	.021	.42	[−0.030, 0.071]						
Preschool	8	− .121	.38	[−0.373, 0.147]	2253.11 (269)	<.0001				
Time of Year							.667	2	215	.514
Intercept[b]	90	.504	<.0001	[0.453, 0.551]						
Spring	96	−.019	.46	[−0.071, 0.032]						
Winter	32	−.040	.28	[−0.112, 0.032]	1942.92 (215)	<.0001				
Type of CBM							6.175	3	268	<.0001
Intercept[c]	45	.401	<.0001	[0.334, 0.464]						
MN	64	.125	<.0001	[0.060, 0.189]						
NI	66	.110	<.0001	[0.048, 0.171]						
QD	91	.121	<.0001	[0.057, 0.183]	1942.92 (268)	<.0001				
Criterion type							9.373	3	268	<.0001
Intercept[d]	50	.305	<.0001	[0.195, 0.407]						
Norm-referenced	144	.212	<.0001	[0.099, 0.320]						
Other CBM	46	.230	<.0001	[0.139, 0.317]						
RD	32	.306	<.0001	[0.154, 0.445]	2178.75 (268)	<.0001				

Table 3. ***Moderator Analyses (continued)***

Moderator	Correlation				Heterogeneity		Test of moderators			
	k	*Est.*	*p*	95% CI	*Q(df)*	*p*	*f*	*df1*	*df2*	*p*
Correlation lag time							0.953	1	260	.330
Intercept[e]	136	.497	<.0001	[0.447, 0.543]						
Predictive	126	−.031	.330	[−0.092, 0.031]	2196.80 (260)	<.0001				
Reporting Quality							2.040	3	268	.109
Intercept[f]	46	.448	<.0001	[0.363, 0.526]						
Low	136	.083	.563	[−0.035, 0.199]						
Acceptable		.014		[−0.123, 0.150]	2247.30 (268)	<.0001				
High		−.093		[−0.271, 0.092]						

Note. CBM = curriculum-based measure; MN = missing number; NI = numeral ID; QD = quantity discrimination; RD = researcher developed.
[a] Intercept = grade 1.
[b] Intercept = fall.
[c] Intercept = oral counting.
[d] Intercept = state test.
[e] Intercept = concurrent.
[f] Intercept = very low.

Table 4. ***Summary of Study Reporting Quality Results***

Study Reporting Quality Scores	Average Score	*n* (*N* = 272)	%
Attrition	0.63		
High quality (3)		12	4.4
Acceptable quality (2)		63	23.2
Low quality (1)		8	2.9
Very low quality (0)		189	69.5
Reliability of EN-CBM	1.48		
High quality (3)		108	39.7
Acceptable quality (2)		34	12.5
Low quality (1)		11	4.0
Very low quality (0)		119	43.8
Scoring Reliability	1.24		
High quality (3)		112	41.2
Acceptable quality (2)		0	0.0
Low quality (1)		0	0.0
Very low quality (0)		160	58.8
Training and Procedural Fidelity	1.60		
High quality (3)		94	34.6
Acceptable quality (2)		32	11.8
Low quality (1)		89	32.7
Very low quality (0)		57	21.0
Overall Average Score	1.24		
High quality[a] (2.75–3.0)		16	5.9
Acceptable quality (2.0–2.74)		43	15.8
Low quality (1.0–1.99)		110	40.4
Very low quality (0–0.99)		103	37.9

Note.
a. High quality includes 2.75–3.0 instead of only 3.0 to allow for studies that scored a 3 on 3 aspects and a score of 2 on one aspect to be deemed high quality.

Table 4 provides a summary of the results across the studies and correlations. We considered attrition, EN-CBM reliability, EN-CBM scoring reliability, and test administrator training and procedural fidelity and then averaged the scores for each quality indicator across studies. Higher scores on a scale of 0 to 3, indicate the data used in the current systematic review had greater study reporting quality as reported by the original studies. Attrition had the lowest overall study reporting quality score ($M = 0.63$). Sixteen of 22 studies received a score of 0 in this study reporting quality category; however, it is important to note 69% of these studies ($k = 11$) received a score of 0 because the authors did not report any information related to attrition. EN-CBM reliability had an average score of 1.48. The reason for this score was that 119 correlations out of 272 correlations had a score of 0, since authors did not report any information about EN-CBM reliability. When we considered only the correlations from studies that reported information about EN-CBM reliability, the average score was much higher ($M = 2.63$ out of 3). Scoring reliability had an average score of 1.24. Here, 160 out of 272 correlations received a score of 0 because the study authors did not provide any information related to scoring reliability. In fact, when these "not reported" cases were excluded, the average scoring reliability quality score was 3.0 out of a maximum score of 3.0. Test administrator training and procedural fidelity had an average score of 1.60; however, when we removed the instances in which authors did not report any information, the average quality score improved ($M = 2.02$). Overall, the average study reporting score was low ($M = 1.24$), primarily due to authors failing to report information associated with the aspects related to study quality that were included in the coding rubric.

Discussion

As early numeracy skills are the strongest predictor of later math achievement, there has been significant attention given to screening early numeracy skills as early as possible to provide necessary support within the MTSS framework (Fuchs et al., 2021). By doing so, it is essential educators have access to screening tools that can accurately assess student competence for key early numeracy skills (Dong et al., 2023; VanDerHeyden et al., 2017). The most common early numeracy CBM implemented in the schools were categorized into four domains: oral counting, numeral identification, quantity discrimination, and missing number. Researchers have investigated these four domains in relation to predicting later achievement (Clarke & Shinn, 2004; Lembke et al., 2008; Methe et al., 2011b). Although criterion validity is important in CBM, there is a lack of research in math, especially through meta-analyses that converge evidence on this topic. Given recent developments (Nelson, G., Kiss, A. J., Codding, R. S., McKevett, N. M., Schmitt, J. F., Park, S., & Hwang, J. 2023), a new meta-analysis is needed to synthesize the latest knowledge on the criterion validity of EN-CBM. This study makes a unique contribution to the literature with a meta-analytic investigation of the relation between students' scores on EN-CBM and broader math performance as measured by a criterion. The results of this study have implications for research and practice related to screening for difficulty with early numeracy concepts at school entry. Moreover, the results of our analysis of study reporting quality offer several recommendations for improving future research.

What Is the Average Weighted Correlation Between EN-CBM and Criterion Measures?

Our first research question aimed to determine the strength of the relation between EN-CBM and math achievement measures. Across 22 studies, we included 272 correlations between the EN-CBM and each study's chosen criterion measure, with an average weighted correlation of $r = .48$ ($p < .05$). When considering metrics for interpreting criterion validity evidence (e.g., Foegen et al., 2007), the results of this meta-analysis indicate, at best, a modest relation between EN-CBM and criterion measures. In fact, according to the National Center on Intensive Intervention (NCII; 2020), a correlation less than .60 would not meet expectations for evidence of validity for an academic screening tool. Yet, our findings aligned with the previous meta-analysis on criterion validity of early math CBM (McCulloch, 2010). Despite differences in methodology and inclusion criteria, McCulloch (2010) reported a similar overall moderate relation between the EM-CBM (counting, comparing and ordering, adding/subtracting, mixed skills, readiness concepts) and norm-referenced criterion measures ($r = .49$). In contrast, Codding et al. (2023) conducted a meta-analysis of MCOMP and MCAP in Grades 2–8 and reported average weighted correlations of .53 and .65, respectively, in relation to criterion measures. This suggests researchers have more work to do in the area of developing early math and early numeracy CBM for the purposes of screening in the earliest grades.

Our findings on the weighted effect size of the correlations between EN-CBM and criterion measures demonstrate a weaker relation when comparing studies examining the validity of comprehensive early numeracy screeners (Jordan et al., 2010; Purpura et al., 2015). For example, the Number Sense Brief by Jordan et al. (2010) predicted failure at the end of Grade 3 on state tests and reported good area under the curve results (0.78–0.88). The Preschool Early Numeracy Skills Screener by Purpura et al. (2015) had a strong correlation with the Test of Early Mathematics Ability ($r = .73$) and an overall classification accuracy of 82.2%. The main reason for this is because each of the EN-CBM in the current meta-analysis targets a narrow early numeracy skill; whereas screeners include a broader range of skills for diagnostic measures, as they provide more accurate information in terms of prediction for later achievement.

Taken all together, the current meta-analysis contributes to the ongoing scholarly debate on the extent to which CBM should include broader math content to increase the capability to support screening decisions with criterion measures, despite the risk of reduced sensitivity to student performance changes (Nelson, G., Kiss, A. J., Codding, R. S., McKevett, N. M., Schmitt, J. F., Park, S., & Hwang, J. (2023). Our findings of the current meta-analysis provide up-to-date evidence on EN-CBM and criterion measures, suggesting moderate positive effects and defining areas of needed improvement in future tools. However, compared to reading CBM (Dong et al., 2023), the evidence for math CBM still shows weaker technical properties. For this reason, scholars in the field highlight the need for better CBM in math, similar to "oral reading fluency" in reading CBM (Miciak et al., 2024). However, such measures remain undeveloped due to the complexity of math skills (Nelson et al. 2023). In addition, a significant drawback to the current literature on work with EN-CBM is the lack of information on diagnostic accuracy (Nelson et al. 2023). Work in develop-

ing and investigating the technical adequacy of early numeracy measures is relatively limited (Nelson, G., Kiss, A. J., Codding, R. S., McKevett, N. M., Schmitt, J. F., Park, S., & Hwang, J. (2023), especially compared to its early literacy counterpart (Methe et al., 2011b). Thus, continued research is necessary to strengthen evidence for math CBM measures.

What Variables Moderate the Relation Between EN-CBM and Criterion Measures?

The results of the moderator analysis indicate there is no statistically significant difference in the magnitude of the correlation between measures administered in preschool, kindergarten, and Grade 1. In contrast, McCulloch (2010) reported the EM-CBM included in that study yielded statistically significant differences, with measures in later grades yielding stronger correlations. Codding et al. (2023) also reported significantly higher correlations between MCOMP/MCAP and criterion measures for students in Grades 6–8 than for students in Grades 2–5. Inconsistent results between the current meta-analysis and previous meta-analyses may be attributed to differences in methodology and the specific skills measured by CBM.

The results of the moderator analysis indicate that EN-CBM of oral counting yielded a significantly smaller average weighted correlation compared to numeral identification, quantity discrimination, and missing number. A smaller average weighted correlation for oral counting was expected. Previous researchers have noted reciting the oral count sequence is an informal math skill most children develop prior to school entry, often as a result of spontaneous interactions with their environment (Purpura et al., 2013). In contrast, formal math knowledge refers to tasks that are often taught as part of a formal curriculum and that require students to use and understand conventional numerical notation (Purpura et al., 2013). Numeral identification, quantity discrimination, and missing number are examples of formal math skills that require students to understand the relation between representations of numbers (Purpura & Lonigan, 2013). Previous researchers have reported skills in numeral identification mediate the relation between informal math knowledge, such as counting, and formal math knowledge measured with calculation and number combination tasks (Purpura et al., 2013).

The results of the moderator analysis indicate state tests had a statistically significant smaller average weighted correlation with EN-CBM compared to all other criterion measures. In contrast, Codding et al. (2023) did not find criterion measure type to be a significant moderator of the relation between MCOMP/MCAP and criterion measures. The smaller average correlation for state tests may be attributed to the fact that in the included studies, only the correlations between state tests and EN-CBM represented an administration lag time greater than 12 months. As more longitudinal research is conducted with EN-CBM, researchers may further investigate lag time as a moderator of the relation between EN-CBM and criterion measures.

The results of the moderator analysis indicated the magnitudes of concurrent and predictive correlations were not statistically different. In our study, we investigated ordinal categories of concurrent (lag time of 1 month or less) versus predictive (any correlation with more than 1 month of lag time), which was based on the overall lag time but provided fewer degrees of freedom in the meta-regression. Further, in a follow-up analysis, no significant difference was found in lag time

according to different benchmarks (i.e., 1 month, 2–6 months, 7–12 months, more than 12 months). Conversely, McCulloch (2010) found a significant relation between predictive or concurrent correlations and the relation between EM-CBM and norm-referenced math achievement assessments. It is possible that findings differed between the present study and McCulloch as a result of inclusion criteria, meta-analysis methodology, and slight differences in the way concurrent and predictive correlations were defined. In addition, Codding et al. (2023) reported correlations that represented a lag time of 5–7 months between MCOMP/MCAP and a criterion measure were significantly smaller than concurrent correlations. Although correlations that represented lag times of 1–3 months, 3–5 months, and more than 1 year were also smaller, differences were not statistically significant (Codding et al., 2023). Finally, our results indicate the time of year (fall, winter, spring) that EN-CBM are administered does not impact the relation between EN-CBM and criterion measures. The findings in the current study show time of administration was not a moderator of the strength of the relation between the EN-CBM and criterion measures, which does not contradict the notion that educators need not wait to use EN-CBM to identify students in need of supplemental math support.

Implications of Study Reporting Quality

Study reporting quality did not significantly moderate the relation between EN-CBM and criterion measures. The findings of our study align with previous meta-analyses showing quality indicators did not moderate the main effect (McCulloch, 2010). Various factors could contribute to the lack of significant results, and it is worth considering previous findings related to quality indicators. For example, (Nelson et al., 2022) found only 84% of quality indicators were met across outcome measures, highlighting the need for multiple measures and providing validity information. Similar trends were observed in our coding results. In several instances, studies failed to report information that allowed our research team to evaluate aspects of quality. Our coding manual emphasized evaluating the quality of the reported information; thus, we encourage readers to avoid equating these findings with actual study or methodological quality.

In addition to the moderator analysis of study reporting quality, the results of the quality coding highlight several areas of growth for future researchers conducting and reporting the results of CBM-related research. In particular, studies and correlations scored relatively better on the aspect of EN-CBM reliability, as approximately half of correlations (52.2%) either (a) reported reliability estimates greater than .80 (38.3%), or (b) reported a mix of reliability estimates with some greater than .80 and some between .70 and .79 (12.8%). When considering only the studies that reported information about EN-CBM reliability that our team could code, the average score surpassed the threshold for acceptable quality and neared a high-quality rating, which is encouraging. Other meta-analyses have noted similar limitations of previous research on CBM that failed to report reliability of the included measures (Reschly et al., 2009). It is critical for all authors to examine the reliability of the data for their sample (Thompson & Vacha-Haase, 2000), especially given the diversity in participants across studies (Vacha-Haase, 1998). Future studies, especially studies focused on measurement and reporting criterion validity of the scores derived

from those measures, should either report sample-specific reliability or provide information about the previously reported reliability estimates and the comparability of sample characteristics and score distributions (Thompson et al., 2005).

Trends were somewhat discouraging for the other quality indicators: attrition, scoring reliability, and test administrator training and fidelity. For attrition, results indicated more than two-thirds of the correlations (69.5%) or studies (72.7%) had very low study reporting quality related to attrition. For scoring reliability, some studies did not provide this information, but those that did reported values above a high threshold of .90. For test administrator training and fidelity, the overall rating for this category was acceptable when considering only the studies that reported information we could code.

Despite the less-than-ideal scores for study reporting quality across the three categories we coded, our results indicate that low scores were generally attributable to authors' failure to report information aligned with our researcher-developed coding rubric. The lack of a standard format in reporting across the studies in this meta-analysis impedes the ability to draw consistent conclusions from the data; it also weakens the dependability of the findings. The field would benefit from a set of standardized reporting parameters, so that data are more easily comparable across studies. Based on the results of the current meta-analysis and quality review, we recommend that future researchers consider reporting attrition and its reasons (Appelbaum et al., 2018), double-scoring tests, rectifying discrepancies in scores, and reporting on their processes related to such effects (Appelbaum et al., 2018), and detailing report training and fidelity of test administration (Gersten et al., 2005). More broadly, however, we recommend that researchers of criterion validity studies follow the recommendations for reporting results of quantitative studies as outlined by the American Psychological Association Journal Article Reporting Standards (JARS) documents. Specifically, readers can find a helpful table of recommended information in the JARS–Quant Table 1 (Appelbaum et al., 2018).

Limitations and Future Research

While the research base on early numeracy is growing, it is relatively limited compared to the work done to date in other areas (e.g., early literacy; Methe et al., 2011b). Combined with the lack of diagnostic accuracy work, implementation in school settings should be done with an understanding of where the field stands in best practices in using EN-CBM for decision-making. However, the results of this meta-analysis are limited in several ways.

When considering the implications of the findings in this study, results must be contextualized in terms of the settings in which the studies were conducted. Most of the correlations (97.0%) used in the analysis represented the administration of EN-CBM in kindergarten or Grade 1. This meta-analysis is limited because few preschool studies met inclusion criteria; thus, readers should use caution as they interpret the results of this study as they relate to younger, preschool students. Future studies should continue to report on the criterion validity of EN-CBM used with younger students. This is especially important given previous research indicates gaps in early numeracy knowledge and skills prior to school entry (Burchinal et al., 2011).

Moreover, the studies included in this meta-analysis did not provide adequate sample demographics to allow us to explore student characteristics as moderators of the relation between EN-CBM and criterion measures. Future investigations of these relations require that researchers of primary studies report sample-level demographics—specifically, students' gender, special education status, socioeconomic status, and English language proficiency (Appelbaum et al., 2018). In a 2022 U.S. NCES report, 36% of Grade 4 students performed at or above proficiency in math; this number decreased for at-risk populations, including students who were eligible for free or reduced lunch (20%), were receiving special education services (16%), or were English learners (14%). The lack of detailed demographic data is problematic on two different levels. First, limited demographic data in the included primary studies limit practitioners' ability to understand how results of individual EN-CBM studies generalize students in their own classrooms (Cook & Cook, 2017). The lack of demographic data presented in CBM studies is highly problematic given a primary use of EN-CBM in schools is to screen for difficulty and monitor student progress in math. National experts recommend that practitioners consider whether the sample used to validate a tool, such as an EN-CBM, is representative of the population with which the tool will be used (NCII & National Center on Improving Literacy, n.d.). Second, limited demographic data in individual studies limit the ability of meta-analysts to examine evidence by student population to report on a summary of findings. We urge researchers to include diverse samples of students and report disaggregated results in their studies so that we may further investigate the nature of the relation between EN-CBM and criterion measures with the students who are most at risk.

This meta-analysis focused only on the four EN-CBM tools that are used for screening in preschool through Grade 1. We did not explore the criterion validity of other early numeracy or early math CBM, such as measures of geometry or measurement. Our choice was driven by the emphasis in the existing literature on the four most common EN-CBM measures, thus limiting the data available to us for this meta-analysis. The study was also limited by the fact that little research has focused on criterion validity for other early math CBM tools (e.g., measurement [one study, six correlations], patterns [two studies, four correlations]). Until more scholars do so, meta-analysis will be limited in their ability to examine the convergence of evidence on this topic. Readers should note that our choice to focus on the four common EN-CBM does not imply that schools should disregard the use of other early numeracy or early math screening tools (Methe et al., 2011a). Future research should continue to explore the use and validity of other early numeracy and early math measures as screeners.

The literature search for this meta-analysis was conducted as part of a broader CBM literature review process, which required that studies identified for inclusion state a focus on investigating the technical properties of CBM (Nelson, G., Kiss, A. J., Codding, R. S., McKevett, N. M., Schmitt, J. F., Park, S., & Hwang, J. (2023). In other words, we did not include all studies that might have reported correlations between scores on EN-CBM and a criterion measure; we included studies only when authors *intentionally* sought to investigate the EN-CBM criterion validity as reported in their purpose statement or research questions. Despite this, the results of this re-

view still provide readers with critical information as they select EN-CBM for screening purposes. Future researchers may consider conducting a meta-analysis on a larger scale to determine if the results are consistent with the results of the current review.

Conducting future diagnostic accuracy studies on EN-CBM is critical because screening decisions in schools are dichotomous; students are identified as at risk or not at risk or as needing intervention or not needing intervention. While the importance of classification accuracy has been extensively covered (e.g., Klingbeil et al., 2019; VanDerHeyden et al., 2017; 2021), the lack of diagnostic work in EN-CBM has precluded the possibility of conducting a meta-analysis with this focus at this time (e.g., Kilgus et al., 2014).

Implications for Practice

The findings of this study have three key implications for practice. First, the results of the current study demonstrate that EN-CBM have moderate levels of evidence to support their use as measures of risk and achievement for math in early grades. District staff who are responsible for selecting early math screening tools may need to weigh the findings of this meta-analysis against other assessment factors. Although EN-CBM may have weaker relations with criterion measures than comprehensive measures of early math (Jordan et al., 2010), they are cost efficient, brief to administer, require little staff training, and are sensitive to student growth over time (Nelson et al., 2023). Schools may want to consider the costs (e.g., direct costs, material prep, training) associated with administering assessments alongside the technical adequacy of the measures (Paly et al., 2022).

Second, to effectively identify at-risk students who can benefit most from early numeracy interventions, studies of diagnostic accuracy are crucial. While evidence of criterion validity is necessary to determine if a measure has the potential to be used for universal screening, diagnostic accuracy is a specific type of evidence necessary to confirm if the measure reliably differentiates between risk for math difficulty and no risk (Kilgus et al., 2014). The lack of diagnostic accuracy studies in the current literature on EN-CBM (Nelson et al., 2023) limits researchers' and practitioners' understanding of the utility of these tools in real-world educational settings. Therefore, there is a critical need for further investigation into the diagnostic accuracy of EN-CBM, along with the development of more robust screening tools. Following these efforts, meta-analysts can then evaluate and report on the convergence of evidence of diagnostic accuracy of EN-CBM, potentially offering additional insight for research and practice.

Finally, when considering the screening process for students in early elementary grades, it is imperative to consider both the importance of engaging in early intervention for students at risk for difficulties in math and the need for students to develop discrete early numeracy skills related to the development of number sense. For this reason, the results of this study suggest that regardless of the time of year a school screens in early math, the measures of EN-CBM tools should perform similarly in relation to criterion measures. It is important to note criterion validity is only one aspect of assessment that educators need to consider as they select screening tools. Another form of validity evidence could be content-oriented evidence, which provides information on the cognitive complexity of the test content and its suit-

ability for all members of the intended population (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Educators will also benefit from examining other sources of evidence for screening tools, including reliability, classification accuracy, and other usability features, such as administration time, format, and scoring (NCII, n.d.). Research on direct contrasts related to use (e.g., the cost of administration, training requirements, and ease of administration and scoring) would offer the field greater context when selecting measures.

Conclusion

National achievement trends in the United States indicate a need for schools to identify students who may be at risk for math failure, even in the earliest years of schooling. The results of this meta-analysis show that the currently available EN-CBM tools have moderate levels of technical adequacy for predicting performance on criterion measures. Because EN-CBM have several notable strengths related to their efficiency, usability, and affordability, it is imperative that scholars continue to research the development and criterion validity of similar and potentially new approaches to math screening in the earliest years of school. Research on the technical adequacy of early numeracy measures is relatively limited, when compared to early literacy measures. Hence, ongoing research is essential to enhance the validation of math CBM measures. In addition, future research is needed to investigate the relation between EN-CBM and criterion measures according to student characteristics such as disability, emergent bilingual status, and socioeconomic status. Finally, researchers who are conducting CBM-related research should consider aspects of their studies that can improve the overall study reporting quality. Doing so will increase researchers' and practitioners' confidence in study results and the overall utility of EN-CBM as screening tools.

References

Aragón, E., Navarro, J. I., Aguilar, M., Cerda, G., & García-Sedeño, M. (2016). Predictive model for early math skills based on structural equations. *Scandinavian Journal of Psychology*, *57*(6), 489–494. https://doi.org/10.1111/sjop.12317

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Beretvas, S. N., & Pastor, D. A. (2003). Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement*, *63*(1), 75–95. https://doi.org/fjw4hk

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (Eds.). (2011). *Introduction to meta-analysis* (2nd ed.). Wiley.

Burchinal, M., McCartney, K., Steinberg, L., Crosnoe, R., Friedman, S. L., McLoyd, V., Pianta, R., & NICHD Early Child Care Research Network. (2011). Examining the black–white achievement gap among low-income children using the NICHD study of early child care and youth development. *Child Development*, *82*(5), 1404–1420. https://doi.org/10.1111/j.1467-8624.2011.01620.x

Chard, D. J., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention*, *30*(2), 3–14. https://doi.org/fngshd

Christ, T. J., Scullin, S., Tolbize, A., & Jiban, C. L. (2008). Implications of recent research: Curriculum-based measurement of math computation. *Assessment for Effective Intervention*, *33*(4), 198–205. https://doi.org/10.1177/153450840731

Chu, F. W., Rouder, J., & Geary, D. C. (2018). Children's early understanding of number predicts their later problem-solving sophistication in addition. *Journal of Experimental Child Psychology*, *169*, 73–92. https://doi.org/10.1016/j.jecp.2017.12.010

Clarke, B., & Shinn, M. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review*, *33*(2), 234–248. https://doi.org/jmb5

Clemens, N. H., Lee, K., Liu, X., Boucher, A., Al Otaiba, S., & Simmons, L. (2023). The relations of kindergarten early literacy skill trajectories on common progress monitoring measures to subsequent word reading skills for students at risk for reading difficulties. *Journal of Educational Psychology, 115(8),* 1045–1069. https://doi.org/10.1037/edu0000814

Codding, R. S., Nelson, G., Kiss, A. J., Shin, J., Goodridge, A., & Hwang, J. (2023). A meta-analysis of the relations between curriculum-based measures in mathematics and criterion measures. *School Psychology Review*, 1–16. https://doi.org/10.1080/2372966X.2023.2224055

Cook, B. G., & Cook, L. (2017). Do research findings apply to my students? Examining study samples and sampling. *Learning Disabilities Research & Practice*, *32*(2), 78–84. https://doi.org/10.1111/ldrp.12132

Cook, B. G., Lloyd, J. W., Mellor, D., Nosek, B. A., & Therrien, W. J. (2018). Promoting open science to increase the trustworthiness of evidence in special education. *Exceptional Children*, *85*(1), 104–118. https://doi.org/10.1177/0014402918793138.

Cook, B. G., Tankersley, M., & Landrum, T. J. (2009). Determining evidence-based practices in special education. *Exceptional Children*, *75*(3), 365–383. https://doi.org/gd23h9

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *The handbook of research synthesis and meta-analysis* (3rd ed.). Russell Sage Foundation. https://doi.org/10.7758/9781610448864

Council for Exceptional Children (CEC) (2014). Council for Exceptional Children: Standards for Evidence-Based Practices in Special Education. *Teaching Exceptional Children*, *46*(6), 206–212. https://doi.org/10.1177/0040059914531389

Cumming, M. M., Bettini, E., & Chow, J. C. (2023). High-quality systematic literature reviews in special education: Promoting coherence, contextualization, generativity, and transparency. *Exceptional Children*, *89*(4), 412–431. https://doi.org/grm7cz

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*(3), 219–232. https://doi.org/10.1177/001440298505200303

Deno, S. L. (2003). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention*, *28*(3-4), 3–12. https://doi.org/10.1177/073724770302800302

Desoete, A., & Grégoire, J. (2006). Numerical competence in young children and in children with mathematics learning disabilities. *Learning and Individual Differences*, *16*(4), 351–367. https://doi.org/c7rjsj

Doabler, C. T., Clarke, B., Kosty, D., Sutherland, M., Turtura, J. E., Firestone, A. R., Kimmel, G. L., Brott, P., Brafford, T. L., Nelson Fien, N. J., Smolkowski, K., & Jungjohann, K. (2022). Promoting understanding of measurement and statistical investigation among second-grade students with mathematics difficulties. *Journal of Educational Psychology, 114(3)*, 560–575. https://doi.org/10.1037/edu0000711

Dong, Y., Dumas, D., Clements, D. H., Day-Hess, C. A., & Sarama, J. (2023). Evaluating the consequential validity of the research-based early mathematics assessment. *Journal of Psychoeducational Assessment*, *41*(5), 575–582. https://doi.org/10.1177/07342829231165812

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. https://www.bmj.com/content/315/7109/629

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education*, *41*(2), 121–139. https://doi.org/dcz4vv

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33(2)*, 188–192. https://doi.org/10.1080/02796015.2004.12086241

Fuchs, L. S., & Fuchs, D. (2004). *What is scientifically based research on progress monitoring?* National Center on Progress Monitoring, American Institute for Research, Office of Special Education Programs.

Fuchs, L. S., Fuchs, D., Compton, D. L., Bryant, J. D., Hamlett, C. L., & Seethaler, P. M. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children*, *73*(3), 311–330. https://doi.org/gjzmxr

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (2021). Bringing data-based individualization to scale: A call for the next-generation technology of teacher supports. *Journal of Learning Disabilities*, *54*(5), 319–333. https://doi.org/hw5p

Fuchs, L. S., & Vaughn, S. (2012). Responsiveness-to-intervention: A decade later. *Journal of Learning Disabilities*, *45*(3), 195–203. https://doi.org/10.1177/0022219412442150

Geary, D. C., Hamson, C. O., & Hoard, M. K. (2000). Numerical and arithmetical cognition: A longitudinal study of process and concept deficits in children with learning disability. *Journal of Experimental Child Psychology*, *77*(3), 236–263. https://doi.org/cmj4nm

Gersten, R. M., Clarke, B., Jordan, N., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children*, *78*(4), 423–445. https://doi.org/drdx

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*(2), 149–164. https://doi.org/10.1177/001440290507100202

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

January, S. A. A., & Klingbeil, D. A. (2020). Universal screening in grades K-2: A systematic review and meta-analysis of early reading curriculum-based measures. *Journal of School Psychology*, *82*, 103–122. https://doi.org/10.1016/j.jsp.2020.08.007

Jordan, N. C., Glutting, J., Ramineni, C., & Watkins, M. W. (2010). Validating a number sense screening tool for use in kindergarten and first grade: Prediction of mathematics proficiency in third grade. *School Psychology Review*, *39*(2), 181–195. https://doi.org/10.1080/02796015.2010.12087772

Ketterlin-Geller, L. R., Shivraj, P., Basaraba, D., & Schielack, J. (2019). Universal screening for algebra readiness in middle school: Why, what, and does it work?. *Investigations in Mathematics Learning*, *11*(2), 120–133. https://doi.org/kvwv

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of School Psychology*, *52*(4), 377–405. https://doi.org/10.1016/j.jsp.2014.06.002

Kiss, A. J., & Christ, T. J. (2018). Screening for math in early grades: Is reading enough? *Assessment for Effective Intervention*, *45*(1), 38–50. https://doi.org/gk58

Kiss, A. J., Nelson, G., & Christ, T. J. (2019). Predicting third-grade mathematics achievement: A longitudinal investigation of the role of early numeracy skills. *Learning Disability Quarterly*, *42*(3), 161–174. https://doi.org/10.1177/0731948718823083

Klingbeil, D. A., Maurice, S. A., Van Norman, E. R., Nelson, P. M., Birr, C., Hanrahan, A. R., Schramm, A. L., Copek, R. A., Carse, S. A., Koppel, R. A., & Lopez, A. L. (2019). Improving mathematics screening in middle school. *School Psychology Review*, *48*(4), 383–398. https://doi.org/10.17105/SPR-2018-0084.V48-4

Koponen, T., Aunola, K., & Nurmi, J. (2019). Verbal counting skill predicts later math performance and difficulties in middle school. *Contemporary Educational Psychology*, *59*, article 101803. https://doi.org/10.1016/j.cedpsych.2019.101803

Lê, M. L., & Noël, M. P. (2021). Preschoolers' mastery of advanced counting: The best predictor of addition skills 2 years later. *Journal of Experimental Child Psychology, 212,* 105252. https://doi.org/10.1016/j.jecp.2021.105252

Lee, Y. S., & Lembke, E. (2016). Developing and evaluating a kindergarten to third grade CBM mathematics assessment. *ZDM*, *48*(7), 1019–1030. https://doi.org/gf5z6j

Lee, Y. S., Lembke, E., Moore, D., Ginsburg, H. P., & Pappas, S. (2012). Item-level and construct evaluation of early numeracy curriculum-based measures. *Assessment for Effective Intervention*, *37*(2), 107–117. https://doi.org/gkpv

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*(4), 206–214. https://doi.org/10.1177/1534508407313479

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Sage Publications.

Locuniak, M. N., & Jordan, N. C. (2008). Using kindergarten number sense to predict calculation fluency in second grade. *Journal of Learning Disabilities*, *41*(5), 451–459. https://doi.org/10.1177/0022219408321126

McCulloch, L. M. (2010). *A systematic reivew* [sic] *and meta-analysis of criterion-related validity in early mathematics curriculum-based measurement* [Master's Thesis, East Carolina University]. The ScholarShip. http://hdl.handle.net/10342/2685

Methe, S. A., Begeny, J. C., & Leary, L. L. (2011a). Development of conceptually focused early numeracy skill indicators. *Assessment for Effective Intervention*, *36*(4), 230–242. https://doi.org/10.1177/1534508411414150

Methe, S. A., Hojnoski, R., Clarke, B., Owens, B. B., Lilley, P. K., Politylo, B. C., White, K. M., & Marcotte, A. M. (2011b). Innovations and future directions for early numeracy curriculum-based measurement: Commentary on the special series. *Assessment for Effective Intervention*, *36*(4), 200–209. https://doi.org/10.1177/1534508411414154

Miciak, J., Famer, R. L., & VanDerHeyden, A. M. (2024). Academic assessment. In M. I. Axelrod & S. Hupp (Eds.), *Investigating School Psychology. Pseudoscience, Fringe Science, and Controversies* (pp. 86–97). Routledge.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), article e1000097. https://doi.org/10.1371/journal.pmed.1000097

National Center for Education Statistics. (2022). *NAEP report card: Mathematics.* https://www.nationsreportcard.gov/mathematics

National Center on Intensive Intervention. (n.d.). *Academic screening tools chart.* https://intensiveintervention.org/resource/academic-screening-tools-chart

National Center on Intensive Intervention (2020). *Academic screening tools chart rating rubrics.* https://intensiveintervention.org/sites/default/files/NCII_AcademicScreening_RatingRubric_2020-06-30.pdf

National Center on Intensive Intervention & National Center on Improving Literacy. (n.d.). *Sample representativeness.* https://intensiveintervention.org/sites/default/files/Sample_Rep_508.pdf

Nelson, G., Kiss, A. J., Codding, R. S., McKevett, N. M., Schmitt, J. F., Park, S., ... & Hwang, J. (2023). Review of curriculum-based measurement in mathematics: An update and extension of the literature. *Journal of School Psychology, 97*, 1–42. https://doi.org/10.1016/j.jsp.2022.12.001

Nelson, G., Park, S., Brafford, T., Heller, N. A., Crawford, A. R., & Drake, K. R. (2022). Reporting quality in math meta-analyses for students with or at risk of disabilities. *Exceptional Children, 88(2)*, 125–144. https://doi.org/10.1177/00144029211050851

Nelson, G., & Powell, S. R. (2018). A systematic review of longitudinal studies of mathematics difficulty. *Journal of Learning Disabilities*, *51*(6), 523–539. https://doi.org/gbhg7q

Paly, B. J., Klingbeil, D. A., Clemens, N. H., & Osman, D. J. (2022). A cost-effectiveness analysis of four approaches to universal screening for reading risk in upper elementary and middle school. *Journal of School Psychology*, *92*, 246–264. https://doi.org/10.1016/j.jsp.2022.03.009

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, *88*, article 105906. https://doi.org/10.1016/j.ijsu.2021.105906

Park, S., Lee, Y. R., Nelson, G., & Tipton, E. (2024). Four best practices for meta-analysis: A systematic review of methodological rigor in mathematics interventions for students with or at risk of disabilities. *Learning Disability Quarterly, 47(4)*, 234–246. https://doi.org/10.1177/07319487231185133

Park, S., & Nelson, G. (2022). The quality of outcome measure reporting in early numeracy intervention studies. *Psychology in the Schools, 59(9)*, 1721–1736. https://doi.org/10.1002/pits.22726

Peltier, C. J. (2017). *Verifying and looking into data: Validity of mathematics curriculum based measures* [Doctoral Dissertation, Texas A&M University]. OAKTrust. https://hdl.handle.net/1969.1/173120

Purpura, D. J., Baroody, A. J., & Lonigan, C. J. (2013). The transition from informal to formal mathematical knowledge: Mediation by numeral knowledge. *Journal of Educational Psychology*, *105*(2), 453–464. https://doi.org/10.1037/a0031753

Purpura, D. J., & Lonigan, C. J. (2013). Informal numeracy skills: The structure and relations among numbering, relations, and arithmetic operations in preschool. *American Educational Research Journal*, *50*(1), 178–209. https://doi.org/gjkx72

Purpura, D. J., Reid, E. E., Eiland, M. D., & Baroody, A. J. (2015). Using a brief preschool early numeracy skills screener to identify young children with mathematics difficulties. *School Psychology Review*, *44*(1), 41–59. https://doi.org/10.17105/SPR44-1.41-59

Pustejovsky, J. E. (2023). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* (R package version 0.5.9) http://jepusto.github.io/clubSandwich/

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, *23*, 425–438. https://doi.org/10.1007/s11121-021-01246-3

Reid, E. E., Morgan, P. L., DiPerna, J. C., & Lei, P. W. (2006). Development of measures to assess young children's early academic skills: Preliminary findings from a Head Start-university partnership. *Insights on Learning Disabilities*, *3*(2), 25–38.

Reigosa-Crespo, V., Valdés-Sosa, M., Butterworth, B., Estévez, N., Rodríguez, M., Santos, E., Torres, P., Suárez, R., & Lage, A. (2012). Basic numerical capacities and prevalence of developmental dyscalculia: The Havana Survey *Developmental Psychology*, *48*(1), article 123. https://doi.org/10.1037/a0025356

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology*, *47*(6), 427–469. https://doi.org/10.1016/j.jsp.2009.07.001

Riskofbias.info. (2023, April 1). *Risk of bias tools.* https://www.riskofbias.info/

Rouselle, L., & Noël, M. (2007). Basic numerical skills in children with mathematics learning disabilities: A comparison of symbolic vs non-symbolic number magnitude processing. *Cognition*, *102*(3), 361–395. https://www.doi.org/10.1016/j.cognition.2006.01.005

Shin, J., & McMaster, K. (2019). Relations between CBM (oral reading and maze) and reading comprehension on state achievement tests: A meta-analysis. *Journal of School Psychology*, *73*, 131–149. https://doi.org/10.1016/j.jsp.2019.03.005

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology*, *70*, 747–770. https://doi.org/10.1146/annurev-psych-010418-102803

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*(8), 795–819. https://doi.org/10.1002/pits.20113

Stephens, M., Erberber, E., Tsokodayi, Y., & Fonseca, F. (2022). *Changes between 2011 and 2019 in achievement gaps between high- and low-performing students in mathematics and science: International results from TIMSS* (NCES 2022-041). U.S. Department of Education. National Center for Education Statistics, Institute of Education Sciences. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2022041.

Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., ... & Higgins, J. P. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*, article d4002. https://doi.org/10.1136/bmj.d4002

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, *40*(6), 604–634. https://doi.org/10.3102/1076998615606099

Thompson, B., Diamond, K. E., McWilliam, R., Snyder, P., & Snyder, S. W. (2005). Evaluating the quality of evidence from correlational research for evidence-based practice. *Exceptional Children*, *71*(2), 181–194. https://doi.org/10.1177/001440290507100204

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*(2), 174–195. https://doi.org/10.1177/0013164400602002

University of South Australia. (n.d.). *Critical appraisal tools.* https://guides.library.unisa.edu.au/SystematicReviews/CriticalAppraisal

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*(1), 6–20. https://doi.org/10.1177/0013164498058001002

VanDerHeyden, A. M., Broussard, C., & Burns, M. K. (2021). Classification agreement for gated screening in mathematics: Subskill mastery measurement and classwide intervention. *Assessment for Effective Intervention*, *46*(4), 270–280. https://doi.org/10.1177/1534508419882484

VanDerHeyden, A. M., Codding, R. S., & Martin, R. (2017). Relative value of common screening measures in mathematics. *School Psychology Review*, *46*(1), 65–87. https://doi.org/10.1080/02796015.2017.12087608

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/gckfpj

Viechtbauer, W., & Viechtbauer, M. W. (2015). *Package 'metafor'. The comprehensive R archive network*. https://cran.r-project.org/web/packages/metafor/metafor.pdf

Vukovic, R. K., & Siegel, L. S. (2010). Academic and cognitive characteristics of persistent mathematics difficulty from first through fourth grade. *Learning Disabilities Research & Practice*, *25*(1), 25–38. https://doi.org/10.1111/j.1540-5826.2009.00298.x

Declaration of Interest

The authors declare that there are no conflicts of interest.

Ethics Statement

This study did not require human subjects as this is an analysis of previously published studies; we were exempt from an IRB.

Funding Statement

There is no funding to report.