VERIFYING AND LOOKING INTO DATA: VALIDITY OF MATHEMATICS CURRICULUM BASED MEASURES

A Dissertation

by

COREY JAMES PELTIER

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Kimberly J. Vannest
Committee Members,	Jennifer B. Ganz
	Patricia S. Lynch
	Oimon Kwok
	Monica Neshyba
Head of Department,	Shanna Hagan-Burke

December 2017

Major Subject: Special Education

ABSTRACT

The mathematical performance of U.S. students has drawn attention from the field of education as well as the public sector. An integral component of the nationwide initiative to improve mathematics instruction is using data for decision-making. However, data is only useful if it is reliable and valid, which requires technically sound measures. This dissertation includes two articles: (a) a literature review on the criterion validity of mathematics curriculum-based measures and (b) a correlational study analyzing the criterion validity of a mathematics curriculum-based measure.

The first study is a review of the literature that administered mathematics curriculum-based measures (m-CBMs) and examined the criterion validity of the scores. The review includes 40 articles that met the following criteria: (a) published in a peer-review journal, (b) administered a m-CBM with school age students, (c) reported quantitative data regarding the validity of scores, and (d) was published in English. Variables were identified and coded that may moderate the validity of scores produced, these variables included the mathematical focus of the measure and administration protocol (i.e., timing, paper pencil/computer, proctor, and grouping [i.e., classwide, small group, individual]). Results suggest concepts and applications m-CBMs yielded the strongest validity coefficients to standardized measures of mathematics performance for students in upper elementary and middle school. Scores from numeracy measures indicate evidence of criterion validity to standardized measures of mathematical achievement for early elementary students. There was no evidence the proctor or grouping moderate the validity; a mismatch between the administration format or the m-CBM and the criterion measure may affect the validity.

The second article analyzes the criterion validity of a computer adaptive m-CBM used for universal screening purposes. Data from 1195 students in third through eighth grade attending four schools located in the rural Southern U.S. were included. Correlational analyses were used to identify the predictive and concurrent validity of the computer adaptive m-CBM to the end-of-year state assessment. Multiple linear regression analyses were used to identify whether student demographic variables (i.e., gender, race, free and reduced meals, limited English proficiency, special education, Section 504) moderated the validity. Results suggest the m-CBM had strong criterion validity to the end-of-year state assessment across grades. Validity coefficients were strongest to the major content domain and the weakest to the additional and supporting content. Moderator analyses reveal that the demographic variables: gender, SPED, FARMS, Section 504, and LEP moderated the criterion validity of m-CBM.

DEDICATION

For my son, Harper Chief and my daughter, Rye Colleen.

ACKNOWLEDGEMENTS

I would like to thank my committee chair and advisor, Dr. Kimberly Vannest, for her unwavering support throughout my studies. In addition, I extend my sincerest gratitude to my committee members: Drs. Jennifer Ganz, Oi-man Kwok, Patricia Lynch, and Monica Neshyba for their expertise and guidance. Lastly, I would like to thank the faculty of the undergraduate Special Education Program. I have enjoyed working alongside you throughout my program; you have made my experience at Texas A&M memorable.

CONTRIBUTORS AND FUNDING SOURCES

This work was supervised by a dissertation committee consisting of Drs. Kimberly Vannest [advisor and chair], Jennifer Ganz [EPSY], Patricia Lynch [EPSY], Oimon Kwok [EPSY], and Monica Neshyba [TLAC].

The existing data sets were obtained by Dr. Kimberly Vannest and provided in a de-identified form from a regionally proximal school district. There are no outside funding contributions to acknowledge related to the research and compilation of this document.

All work for the dissertation was completed independently by the student.

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

Roughly 5 – 10% of students are expected to have persistent low achievement in mathematics (Berch & Mazzocco, 2007; Geary, 2011). Data from the National Assessment of Educational Progress (2016) further substantiates these concerns with 60% of fourth- and 67% of eighth-graders failing to demonstrate expected proficiency. In addition, The Office of Special Education and Rehabilitation Services' annual report to Congress presented data for students with disabilities on end-of-year state assessments that was concerning. The median percentage of students served under Part B of the IDEA in third through fourth grade who demonstrated proficiency ranged from 27-36% and 17-20% for sixth grade through high school (Office of Special Education and Rehabilitation Services, U.S. Department of Education, 2016). The hierarchical nature of mathematics curricula makes these data concerning; students displaying early deficits are likely to continue tp struggle and have poor mathematical outcomes (Geary, 2011; Murphy, Mazzocco, Hanich, & Early, 2007). Furthermore, the mathematical achievement of post-secondary adults has been shown to be a strong indicator of career outcomes for both typically developing students (Bynner & Parsons, 1997; Koedel & Tyhurst, 2012; Rivera-Batiz, 1992; Snow, Burns, & Griffin, 1998) and students identified with disabilities (Benz, Lindstrom, & Yovanoff, 2000; Reynolds, Elksnin, & Brown, 1996; Test et al., 2009). Thus, the National Mathematics Advisory Panel (2008) released a final report and concluded eloquently, "America has genuine opportunities for improvement in mathematics education" (p. xiii).

Experts from the field as well as leading organizations advocated for initiatives that focus on the improvement of mathematics performance. A core component of these initiatives is collecting and using data to inform administrative and instructional decision-making (Deno, 2003; Fuchs, Fuchs, Eaton, Hamlet, & Karns, 2000; Mandinach, 2012; Marsh, Pane, & Hamilton, 2006; National Council of Teachers of Mathematics, 2000; National Mathematics Advisory Panel, 2008). The emphasis on data, in addition to research in the field of community health and medicine (Noam & Hermann, 2002; Pearce, 1996), lead to the adoption of multi-tiered prevention models (e.g., Multi-Tiered Systems of Support [MTSS], Positive Behavior Interventions & Supports [PBIS], Response to Intervention [RtI]). Within an MTSS framework, collecting and using data are integral for screening and monitoring student progress after receiving instruction (e.g., Chard, Harn, Horner, Simmons, & Kame-enui, 2008; Hollenback, 2007; Seethaler & Fuchs, 2011; Shinn, 1998; Sugai & Horner, 2009).

To collect usable data, schools adhering to an MTSS framework use curriculum-based measurement systems. Jitendra, Dupuis, and Zaslofsky (2014) reiterated this sentiment stating, "to provide meaningful assessments of student progress, educators need valid and reliable, curriculum-based formative assessment measures" (p. 241). Research has demonstrated that mathematics curriculum based measures (m-CBMs) administered for the purpose of universal screening or progress monitoring produce valid and reliable scores (Deno, 1985; Fuchs & Deno, 1991; Fuchs, Fuchs, & Courey, 2005). Furthermore, recent work has suggested teachers and pre-service teachers can effectively interpret and use data obtained from m-CBMs if provided guidance and training (Espin, Miura Wayman, Deno, McMaster, & Rooij, 2017; Keuning, Van Geel, & Visscher, 2017; Wagner, Hammerschmidt-Snidarich, Espin, Steifert, & McMaster, 2017; van den Bosch, Espin, Chung, & Saab, 2017). Although initial work has demonstrated the technical adequacy of certain m-CBMs (e.g., Christ, Scullin, Tolbize, & Jiban, 2008; Foegen, Jiban, & Deno, 2007), additional research is needed with particular attention to variables that may moderate the reliability and validity of the scores.

This dissertation investigated how the validity of m-CBMs is moderated when they are administered in various environments to diverse student populations. The purpose for the first study was to synthesize the current data-based literature on the validity of m-CBMs. The review addressed the following research questions:

(a) What empirical evidence regarding the criterion validity of m-CBMs is available across age ranges of students?

(b) How does the validity of m-CBMs vary across the mathematical focus of m-CBM and administration protocol (i.e., timing, paper pencil/ versus computer, proctor, and grouping [i.e., classwide, small group, individual])?

(c) How does the validity of m-CBMs vary depending on the measure set as the criterion (i.e., mathematical achievement measures, end of year state examinations)?

The second study assessed the validity of an m-CBM administered for the purpose of universal screening and benchmarking within an MTSS framework. Correlational analyses were used to analyze the validity of the m-CBM to the end of year state assessment; multiple regression analyses were used to identify how validity was moderated by demographic variables of participants. The study addressed the following research questions:

(a) What is the predictive and concurrent validity of a computer adaptive m-CBM (i.e., i-Ready Diagnostic) to an end-of-year state assessment (i.e., LEAP 2025)?

(b) How does the predictive and concurrent validity of the computer adaptive m-CBM vary across time (fall, winter) to the domains of the end-of-year state assessment (i.e., major content, expressing mathematical reasoning, modeling & application, additional & supporting content)?

(c) Is the predictive or concurrent validity moderated by demographic variables of participants: gender, race, free and reduced meals, limited English proficiency, special education, and Section 504?

VALIDITY OF MATHEMATICS CURRICULUM-BASED MEASURES: CURRENT EVIDENCE AND FUTURE DIRECTIONS

Introduction

Data from the National Assessment of Educational Progress (2016) and the Program for International Students Assessment (2015) suggest the mathematics performance of U.S. students is below expectations. The performance trends have raised concern, which resulted in leading organizations (e.g., Council for Exceptional Children, National Council of Teachers of Mathematics) addressing the mathematics instruction students are receiving. One component of improvement initiatives encouraged teachers to engage in data-based decision-making within an MTSS framework (Ball & Christ, 2012; Council for Exceptional Children, 2009; National Mathematics Advisory Panel, 2008).

Data-Based Decision-Making

In the field of special education, the use of data to inform instructional programming is not new. Deno and Mirkin (1977) defined Data-Based Program Modification (DBPM) as an approach relying on practitioners' use of empirical data to adapt the instructional environment and content based on individualized needs. Core components underlying DBPM incolves practitioners hypothesizing that an instruction program, curriculum, or strategy would lead to student effects, followed by collecting and analyzing corroborating data (Deno & Mirkin, 1977). Historically, the reliance on data is integral to special education and is more recently emphasized in general education.

One reason for the increased emphasis on data in general education was the passage of No Child Left Behind (2002), which placed a priority on data-based decision-making in the general education environment (Datnow & Hubbard, 2016; Mandinach, 2012; Schildkamp, Ehren, & Lai, 2012). In addition, the Individuals with Disabilities Education Improvement Act (IDEA, 2004) updated the evaluation process for students with specific learning disabilities. IDEA (2004) allowed states to use a data driven process to monitor students' responses to research-based interventions and determine students' eligibility for special education services. Therefore, states were allowed to use a Multi-Tiered System of Support (MTSS) framework to identify students with specific learning disabilities. In addition, up to 15% of funding reserved for IDEA (2004) Part B can be allocated to early intervention services. These pieces of education legislation have placed an onus on schools to adopt data-based intitiatives for students with and without disabilities.

Curriculum-Based Measures in MTSS

Within an MTSS framework, curriculum-based measures (CBMs) are integral to the data-based decision-making process (Deno, 2003; Fletcher & Vaugh, 2009; Fuchs & Fuchs, 2006; Hughes & Dexter, 2011; Keller-Margulis, Shapiro, & Hintze, 2008). CBMs are administered to gather data and to inform professionals when making both high- (e.g., special education eligibility, grade retention) and low-stakes (e.g., alter instruction, receive supplemental instruction) decisions (Blankenship, 1985; Fuchs, Deno, & Marston, 1983; Good III, Simmons, & Kame'enui, 2001; Marston, Mirkin, & Deno, 1984; Shinn, 1989; Stecker, Fuchs, & Fuchs, 2005; Ysseldyke et al., 1983). Within an

MTSS framework, CBMs are administered for two purposes: universal screening and progress monitoring (Gersten et al., 2009). Universal screeners are typically administered three times a year (i.e., fall, winter, spring) with the purpose of identifying students who are not responding to research-based instruction. Progress monitoring probes are typically administered weekly or bi-weekly with the purpose of monitoring student response to intensive research-based instruction and making informed instructional decisions based on the student performance. Universal screeners and progress monitoring probes have different purposes and participants; thus, practitioners need to consider the validity of CBM data when making data-based decisions.

The technical adequacy of CBMs must be considered because the data obtained are used for data-based decision-making. However, the validity of scores obtained from CBMs is often overlooked or assumed valid. Score validity is essential in MTSS; invalid scores ultimately will lead to invalid decisions (Ball & Christ, 2012). Invalid low-stakes decisions are problematic; however, for high-stakes decisions, schools may be culpable of educational malpractice. Therefore, considering the validity of scores produced by CBMs is an integral step for the research community to undertake (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Kane, 1992; Messick, 1995).

Literature on Validity of Mathematics Curriculum-Based Measures

Two recent literature reviews (Christ et al., 2008; Foegen et al., 2007) have analyzed the technical adequacy of mathematics curriculum-based measures (m-CBMs). Christ et al. (2008) reviewed nine studies that reported on the validity of m-CBMs targeting computation. The studies included participants across grades (i.e., elementary [n = 9], middle school [n = 3], high school [n = 1]). The authors concluded the current literature on m-CBMs targeting computation have empirical evidence supporting their validity for screening-type decisions. However, the authors caution that the criterion validity for computation measures will lack strong correlation to assessments of overall mathematics achievement because language demands, spatial reasoning, and conceptual understanding of concepts and applications are not captured by the measures.

Foegen et al. (2007) included 16 studies that analyzed the validity of m-CBMs. A majority of the studies (i.e., 10 of 16) included elementary age participants; five studies included participants in middle school and one study included participants in pre-kindergarten. No studies included participants in high school. The correlation between static m-CBMs and criterion measures are moderate (i.e., 0.50-0.70). M-CBMs that measured students understanding of concepts and applications had stronger correlation with standardized measures of mathematical achievement than strictly computation measures. Although based on a small sample, three studies (Fuchs, Fuchs, Karns et al., 2000; Fuchs et al., 2003; Jitendra, Sczesniak, & Deatline-Buchman, 2005) administered m-CBMs targeting word problem solving and reported moderate to strong correlation to standardized measures of mathematical achievement.

The reviews provided preliminary information regarding m-CBMs; however, there are three limitations for the usability of the results. First, Christ et al. (2007) only examined computation m-CBMs. Second, both reviews examined the validity of m-CBMs; however, neither study thoroughly examined administration protocol that may violate the criterion validity of m-CBMs. Lastly, both systematic searches are a decade old. Given both studies urged researchers to conduct additional research, the legislative changes in the IDEA (2004), and the increased use of MTSS systems nationwide an updated systematic search is warranted.

Validity: Considerations and Clarifications

Scores obtained from a measure are not inherently interpretable. Two criteria must be met for scores to be useable. First, scores must be *reliable,* meaning scores consistently measure the same construct. Second, scores must be *valid*, meaning the scores measure the intended construct. Scores can be reliable and not valid; but not vice versa. A common misconception is that reliable assessments exist (Vacha-Hasse, Henson, & Caruso, 2002). An assessment cannot be reliable; this property is reserved for the scores themselves (Thompson, 1994; Thompson & Vacha-Haase, 2000). It is imperative that practitioners consider the reliability and validity of a score on a case-by-case basis when engaging in data-based decision-making.

For a score to be interpretable the validity of the score must be considered. Typically, validity is classified into three categories: content, construct (see Gulliksen & Wilks, 1950), and criterion (Brualdi, 2002; Messick, 1989, 1993, 1996). Although necessary, these three types of validity provide insufficient information (Messick, 1993). Scores are interpreted to guide decision-making; therefore, an argument must be constructed using validity information from the other types to decide if and how the data should be interpreted and if it can be used. Messick (1993) defined this as consequential validity; however, he did not intend it to be interpreted as a fourth type of validity

> …to appraise how well a test does its job, one must inquire whether the potential and actual social consequences of test interpretation and use are not only supportive of the intended testing purposes, but at the same time are consistent with other social values. However, this form of evidence should not be viewed in isolation as a fourth validity type, say, of "consequential validity" (p. 11).

Quantitative and qualitative information should be considered when identifying the potential consequences of administering and using assessment data for decision-making (Messick, 1989, 1996; Shepard, 1997). Messick (1993) addressed this issue stating, "validation combines scientific inquiry with rational arguments to justify (or nullify) test interpretation and use" (p. 2). To construct an argument all three aspects of validity should be considered.

Factors Affecting Validity

In an attempt to control for external variables and to reduce the likelihood of invalidated scores, assessment procedures are standardized. Despite the standardization, there are still factors that may affect score validity. Student characteristics may influence score validity. A few potential characteristics are: test anxiety (Cassady & Johnson, 2002), language proficiency (Winter, Kopriva, Chen, & Emick, 2006; Young et al., 2006), disabilities (Elliott & Roach, 2007; Kettlerlin-Geller, Alonzo, Braun-Monegan, & Tindal, 2007). To reduce the likelihood of invalid scores, the measure should be constructed with validity embedded (DiBello & Stout, 2007). One factor is *response sets* defined as "any tendency causing a person consistently to give different responses to test items than he would when the same content is presented in a different form" (Cronbach, 1946, p. 476). Some practical examples are tests that allow students to omit test items (Grandy, 1987), Likert-scales (Peabody, 1962), multiple or unlimited answers, use of dichotomous answers (e.g., true/false, yes/no, [Schriesheim & Hill, 1981]), assessments that measure fluency, and computer administration (Karkee, Kim, & Fatica, 2010; Kim & Huynh, 2007; Kingston, 2008; Mead & Drasgow, 1993). Furthermore, a variety of factors during the administration of the assessment affect score validity: fidelity of standardized protocol (Lievens & Patterson, 2011), familiarity of proctor (Derosa & Patalano, 1991), and the provision of accommodation (Phillips, 1994).

Current Study

Almost a decade has passed since the most recent literature reviews on m-CBMs (Christ et al., 2008; Foegen et al., 2007). The IDEA (2004) allowed districts and schools to use MTSS systems to identify and support students at-risk for learning disabilities, which lead to an increase in MTSS systems nationwide (Klingner & Edwards, 2006; Zirkel & Thomas, 2010). Researchers examined the utility of CBMs within MTSS frameworks for screening and progress monitoring purposes, particularly in mathematics because of the limited amount of research compared to reading. The purpose of this review was to extend the work of Christ et al. (2008) and Foegen et al. (2007) by updating the literature search and identifying how validity may be moderated by alterations in the administration protocol. This review addressed the following research questions:

1. What empirical evidence regarding the criterion validity of m-CBMs is available across age ranges of students?

2. How does the validity of m-CBMs vary across the mathematical focus of m-CBM and administration protocol (i.e., timing, paper pencil/ versus computer, proctor, and grouping [i.e., classwide, small group, individual])?
3. How does the validity of m-CBMs vary depending on the measure set as the criterion (i.e., mathematical achievement measures, end of year state examinations)?

Method

Search the Literature

To conduct the literature review the following steps were followed: (a) formulate the problem, (b) search the literature, (c) gather information from studies, (d) analyze and integrate the outcomes of studies, (e) interpret the evidence, and (f) present the results (Cooper, 2016). First, an electronic systematic search was conducted through the following databases: Academic Search Complete, Education Resources Information Center, and PsycINFO. Ancestral searches of the two recent literature reviews on m-CBMs (Christ et al., 2008; Foegen et al., 2007) were conducted. The following key terms were used: *assess** AND *(curriculum based measure* OR *CBM)* AND *math** AND *valid*.* The limiter peer-review was also used. The decision to include only peer-reviewed articles was made to increase the likelihood of retaining reliable and valid data to answer the research questions. Reports produced by companies who publish CBM systems (e.g., AIMSweb) were excluded to decrease the likelihood of bias due to conflicts of interest. Dissertations were also excluded. The literature search was conducted in March 2017.

To be included in the review an article needed to meet the following criteria:

- The article was published in a peer-review journal; dissertations and reports were excluded
- The study administered at least one CBM focusued on a construct of mathematics
- The study administered at least one criterion measures focused on a construct of mathematics
- The study reported quantitative data on the criterion validity of scores from an m-CBM
- The article was published in English

The initial search, including the ancestral search, yielded 2,457 articles. A total of 39 articles met the inclusion criteria and were included in the review.

Gather Information from Studies

A coding guide was created to extract relevant information (Cooper, 2016). Data extracted from studies were categorized into three sections: source description, research methods and procedures, and substantive issues. Source descriptions included descriptive information pertaining to the studies. The following data were extracted: publication date, journal, and authors. Research methods and procedures provided pertinent information regarding the study design, measures, and procedures used by the researchers. The following data were extracted regarding research methods and procedures: approach used to develop the m-CBMs (i.e., robust indicators, curriculum sampling; see Fuchs, 2004), focus of the m-CBM, format of administration (i.e., paper/pencil, computer, performance task), proctor, grouping (i.e., classwide, small group, individual), timing, criterion measure(s) selected, and criterion validity coefficients. Substantive issues provided insight into potential variables that varied across studies and allowed for moderator analyses. The following data were extracted regarding substantive issues: grade of participants and type of criterion validity evidence reported (see Appendix A for descriptions of coded variables).

Interrater reliability. I coded all included studies. To assess interrater reliability a recent graduate and four current graduate students from a special education program served as secondary coders. Each coder received a one-on-one in person training before coding. Training consisted of a brief description and example of types of validity reported in articles (i.e., content, construct, criterion) and an explanation of CBMs and their application in an MTSS framework. Coders were provided sample m-CBM measures used in related articles. Coders were then provided a coding sheet with descriptions of the variables being extracted. Coders were encouraged to ask questions for clarification; training sessions lasted for approximately 45 minutes. Coders were given a training article with a criterion of 80% interrater agreement (Gwet, 2014). After meeting the criterion, coders were randomly assigned studies to code. A sample of 21 (53.8%) studies was included in the interrater reliability analyses. Interrater reliability was calculated by counting the number of agreements and dividing by the total number of opportunities to agree. For the training articles, mean agreement was 91.6% (range = 86 – 100%). There were 6 disagreements concerning the following extracted variables: development of m-CBM, grouping, proctor, timing, measures, and results.

Disagreements were discussed and clarification was provided. For the remaining articles, mean agreement was 92.00% (range = 86 – 100%). A total of 23 disagreements were found concerning the following variables: development of m-CBM (n = 10), focus of m-CBM (n = 5), sample size (n = 3), grouping used for administration (n = 3), proctor (n = 1) and validity results (n = 1). All disagreements were resolved.

Results

Current Literature on m-CBMs

Manuscripts. The 39 identified studies represented 20 different author teams and were published in peer-reviewed journals between the years 1992-2016. Manuscripts were published in 19 academic journals from the fields of educational psychology, educational assessment, special education, mathematics education, and computer education. The studies were conducted in: the United States (n = 34), Germany (n = 2), and the Netherlands (n = 1).

Participants. The studies included participants across grades: pre-kindergarten (n = 4), early-elementary (K-2; n = 21), upper-elementary (3-5; n = 13), middle school (6-8; n = 10), high school (9-10, n = 1), and post-secondary (n = 1). A majority of the criterion validity coefficients used data from participants with and without disabilities; however, three studies provided criterion validity information unique to students identified with disabilities.

M-CBMs. Studies administered measures using a curriculum sampling (n = 17) and robust indicators approach (n = 26). In addition, studies administered m-CBMs with varying foci: numeracy (n = 18), basic facts (n = 8), computation (n = 18), concepts and applications (n = 10), word problems (n = 4), estimation (n = 3), teacher rating of performance (n = 1), and algebra (n = 1). A time constraint was a component of the administration protocol for a majority of the m-CBMs administered (n = 31, no time constraint n = 8). The administration protocol included a variety of grouping arrangements: classwide (n = 23), small group (n = 1), and individual (n = 10). Both classwide and individual groupings were used in four studies. One study did not report the administration grouping and one study reported the measure could be administered in small group or individually. A majority of the m-CBMs were administered either paper/pencil or via performance tasks (n = 35); computer administered measures (n = 6) were prevalent in more recently published studies (i.e., 5 published 2011 or later). Teachers (n = 9) or researchers (n = 24) served as proctors for a majority of the studies. However, a few studies used both teachers and researchers (n = 2) or trained assessors (n = 1) as proctors. One study used a computer program with no proctor and three studies did not report who administered the measures.

Criterion measures and type of criterion reported. Studies selected various measures as the criterion by which to report concurrent and predictive validity results: standardized achievement tests (n = 25), end of year state assessment (n = 11), teacher rating/ranking (n = 10), m-CBMs used in previous research (n = 6), and grades (n = 4). Studies that reported only one type of criterion validity addressed either: concurrent (n = 15) or predictive (n = 9) validity. A total of 15 studies reported data on both concurrent and predictive validity. See Tables 1-3 for coding results by study.

Potential Moderators of Criterion Validity

M-CBM focus. *Numeracy.* M-CBMs focused on numeracy comprise the largest literature base. Measures classified within numeracy used the following types of tasks: number recognition/production, counting, filling in a missing number in a sequence or pattern, and discriminating between two sets of quantities. Thirteen studies reported concurrent validity coefficients to standardized measures of mathematics achievement. When a standardized measures of overall mathematics achievement was selected as the criterion studies reported concurrent validity coefficients ranging from 0.03-0.72. A majority of the studies that administered numeracy m-CBMs reported concurrent validity coefficients above 0.50. Salaschek and Souvignier (2014) administered m-CBMs comprised of numeracy tasks to second grade students and reported validity coefficients to their national standardized mathematics exam ranging from 0.54-0.62. When adding in an m-CBM focused on computation, validity coefficients increased minimally to 0.59-0.63. Foegen (2008) administered two novel numeracy tasks requiring complex discrimination and filling in a missing number with middle school students. Concurrent validity coefficients to a standardized assessment of mathematical achievement ranged from 0.52-0.60 across grades for complex discrimination tasks and 0.46-0.67 for missing number tasks.

Table 1

Characteristics of Participants and m-CBM Construction

Study	**N**	**Grades**	**Develop**	**m-CBM Focus**
		Early Elementary		
Allinder (1992)	NR	2	CS	computation
Baglici (2010)	62	K-2	RI	numeracy
Betts (2009)	2180	K-2	RI	numeracy
Chard (2005)	919	K-1	RI	numeracy
Clarke (2004)	52	1	RI	numeracy
Daly (1997)	30	1	RI	numeracy
Eckert (2006)	33	2	RI	basic facts, computation
Floyd (2006)	163	3-6 yrs.	RI	numeracy
Fuchs (1994)	46	2	CS	concepts & applications
Fuchs (2000)	NR	2	CS	word problems
Fuchs (2007)	170	1-2	RI CS	basic facts, computation, concepts & applications, numeracy,
Ginsburg (2016)	280 (K) 297 (1) 338 (2)	K-2	RI CS	K-1: basic facts, numeracy 2: basic facts, computation, concepts & applications, numeracy
Kettler (2013)	136 (1) 142 (2)	1-2	CS	computation, teacher rating
Klinkenberg (2011)	334 (K) 529 (1) 681 (2)	K-2	RI	computation
Laracy (2016)	419	PreK-K	RI	numeracy
Lee (2012)	137	K-1	RI	numeracy
Lee (2016)	280 (K) 297 (1) 338 (2)	K-2	RI	computation, concepts & applications, numeracy
Methe (2008)	64	K	RI	numeracy
Polignano (2012)	40	PreK	RI	numeracy
Salaschek (2013)	148	1	RI	computation, numeracy
Salaschek (2014)	414	2	RI	computation, numeracy
Seethaler (2011)	180-193	K	CS	computation
VanDerHeyden (2001)	40	K	RI	numeracy
VanDerHeyden (2004)	53	PreK	RI	numeracy
		Upper Elementary		
Allinder (1992)	NR	3-5	CS	computation
Fuchs (1994)	49 (3) 45 (4)	3-4	CS	concepts & applications

Table 1 *Continued*

Fuchs (2000)	NR	3-4	CS	word problems
Fuchs (2003)	412	3	CS	word problems
Ginsburg (2016)	337	3	RI CS	basic facts, computation, concepts & applications, numeracy
Jiban (2007)	38 (3) 55 (5)	3, 5	RI	basic facts
Jitendra (2005)	77	3	CS	computation, word problems
Jitendra (2014)	136	3	CS	word problems
Kettler (2013)	135	3-5	CS	computation, teacher rating
Klinkenberg (2011)	529 (3) 513 (4) 574 (5)	3-5	RI	computation
Lee (2016)	337	3	RI	computation, concepts & applications, numeracy
Shapiro (2015)	82-92 (3) 71-84 (4) 64-74 (5)	3-5	CS	computation, concepts & applications
Thurber (2002)	207	4	CS	basic facts, computation
		Middle and High School		
Codding (2015)	249	7	RI	computation
Codding (2016)	408	6-8	RI, CS	basic facts, concept & application
Foegen (2000)	105	6	RI	basic facts, estimation
Foegen (2001)	100	6-8	RI	computation, estimation
Foegen (2008)	563	6-8	RI CS	basic facts, computation, concepts & applications, estimation, numeracy
Helwig (2002)	171	8	CS	concepts & applications
Helwig & Tindal (2002)	193	8	CS	concepts & applications
Hosp (2014)	41	post-sec	CS	computation, concepts & applications
Johnson (2012)	189 (7) 168 (8) 123 (10)	7 8 10	RI	algebra
Klinkenberg (2011)	416 (6) 75 (Sec)	6-Sec	RI	computation

Note. CS = curriculum sampling, RI = robust indicators, Sec = Secondary

Table 2

Characteristics of m-CBM Administration

Study	**Timed**	**Grouping**	**Format**	**Procto r**	**Purpose**
			Early Elementary		
Allinder (1992)	Y	CW	Paper	T	PM
Baglici (2010)	Y	I	Perf	R	US
Betts (2009)	N	I	Perf	TA	US
Chard (2005)	Y	I	Perf	R	US
Clarke (2004)	Y	I	Perf	R	US
Daly (1997)	Y	1	Paper, Perf	R	US
Eckert (2006)	Y	CW	Paper	R	US
Floyd (2006)	Y	I	Perf	R	US
Fuchs (1994)	Y	CW	Paper	T	PM
Fuchs (2000)	N	CW	Paper	R	US
Fuchs (2007)	Y	CW	Paper	R	PM, US
Ginsburg (2016)	Y	CW	Comp, Paper, Perf	R	US
Kettler (2013)	Y	CW	Paper, Perf	T	US
Klinkenberg (2011)	Y	CW[a]	Comp	T[b]	PM
Laracy (2016)	Y	I	Perf	T	US
Lee (2012)	N	I	Perf	R	US
Lee (2016)	Y	CW, I	Paper, Perf	R	US
Methe (2008)	Y	I	Perf	R	US
Polignano (2012)	Y	I	Perf	R	US
Salaschek (2013)	N	CW, I	Comp	T	PM
Salaschek (2014)	N	CW	Comp	T	PM
Seethaler (2011)	Y	CW	Paper	T, R	PM
VanDerHeyden (2001)	Y	CW	Paper	R	US
VanDerHeyden (2004)	Y	I	Comp	R	US
			Upper Elementary		
Allinder (1992)	Y	CW	Paper	T	PM
Fuchs (1994)	Y	CW	Paper	T	PM
Fuchs (2000)	Y	CW	Paper	R	US
Fuchs (2003)	N	CW	Paper	R	US
Ginsburg (2016)	Y	CW	Comp, Paper, Perf	R	US
Jiban (2007)	Y	CW	Paper	R	US
Jitendra (2005)	Y	CW	Paper	R	PM

Table 2 *Continued*

Jitendra (2014)	Y	CW	Paper	R	PM
Kettler (2013)	Y	CW	Paper, Perf	T	US
Klinkenberg (2011)	Y	CW[a]	Comp	T[b]	PM
Lee (2016)	Y	CW, I	Paper, Perf	R	US
Shapiro (2015)	Y	I, SG	Comp, Paper	R	PM
Thurber (2002)	Y	SG	Paper	R	US
		Middle and High School			
Codding (2015)	Y	CW	Paper	R	US
Codding (2016)	Y	CW	Paper	R	US
Foegen (2000)	Y	NR	Paper	T, R	PM
Foegen (2001)	Y	CW	Comp	T	US
Foegen (2008)	Y	CW	Paper	T	US
Helwig (2002)	N	CW	Paper	NR	US
Helwig & Tindal (2002)	N	CW	Paper	NR	PM, US
Hosp (2014)	Y	CW[c]	Paper	R	US
Johnson (2012)	Y	CW	Paper	T	PM
Klinkenberg (2011)	Y	CW[a]	Comp	T[b]	PM

Note. I = individual administration, CW = class wide administration, SG = small group administration, T = teacher, R = researcher, TA = trained assessor, PM = progress monitoring, US = universal screening
[a]The first two probes were administered classwide, students accessed the remaining independently in school and at home.
[b]The teacher administered the first two probes, no proctor was used for the remaining.
[c]Make ups were administered individually

Twelve studies administered numeracy m-CBMs and reported predictive validity coefficients to standardized measures of mathematical achievement. Predictive validity coefficients ranged from 0.02-0.70. Missing number and quantity discrimination tasks serve as better predictors for mathematical achievement as students progress through grades. Foegen (2008) reported predictive validity coefficients ranging from 0.53-0.58 for complex discrimination tasks and 0.48-0.60 for missing number tasks. Fuchs et al.

(2007) administered m-CBMs focused on number identification and counting to first and second grade students. When using the intercept, validity coefficients to a standardized measure of computation were 0.34 for number identification and counting and 0.39 for problem solving. However, when slope was used, coefficients shrunk to -0.11 for number identification and counting and to -0.19 for problem solving. Salaschek and Souvignier (2014) administered m-CBMs focused on numeracy and reported predictive validity coefficients to their national standardized mathematics exam ranging from 0.66-0.69. When adding in an m-CBM focused on computation as a predictor, the validity coefficients increased slightly to 0.72-0.77.

Computation. Studies that administered m-CBMs focused on computation reported criterion validity to a variety of measures. When standardized measures focused on computation were selected, the criterion validity coefficients were larger than standardized measures focused on different constructs of mathematical achievement. Seven studies (i.e., Allinder, Fuchs, Fuchs, & Hamlett, 1992; Foegen & Deno, 2001; Fuchs et al., 2007; Ginsburg, Lee, & Pappas, 2016; Hosp, Hensley, Huddle, & Ford, 2014; Jitendra et al., 2005; Thurber, Shinn, & Smolkowski, 2002) administered an m-CBM focused on computation and set the criterion as a standardized achievement test with both a computation subtest and a subtest measuring another construct of mathematical achievement. Concurrent validity coefficients were stronger to the computation subtest than other subtests of mathematical achievement across five studies that reported this data. Similarly, the predictive validity coefficients were larger to the computation subtest than other subtests of mathematical achievement across three of the four studies reporting this data. Studies that analyzed both concurrent and predictive validity found stronger validity coefficients for concurrent than for predictive validity.

Ten studies (i.e., Codding, Petscher, & Truckenmiller, 2015; Foegen, 2008; Hosp et al., 2014; Kettler & Albers, 2013; Klinkenberg, Straatemeier, & van der Maas, 2011; Lee & Lembke, 2016; Salaschek & Souvignier, 2014; Seethaler & Fuchs, 2011; Shapiro, Dennis, & Fu, 2015; Thurber et al., 2002) administered m-CBMs focused on computation and reported concurrent validity to a criterion measure of overall mathematical achievement. Three of the studies reported validity coefficients less than 0.50, whereas six studies reported a coefficient larger than 0.50. Foegen (2008) reported correlation coefficients ranging from 0.59-0.64 for sixth grade students compared to a coefficient of 0.38 for seventh grade students.

Teacher ratings of student performance were selected as the criterion measure in three studies (Eckert, Dunn, Codding, Begeny, & Kleinmann, 2006; Foegen, 2008; Foegen & Deno, 2001). Validity coefficients ranged from 0.09-0.54 when analyzing the concurrent and predictive validity of computation m-CBMs and a teacher rating of student mathematical performance.

Concepts and applications. Ten studies (i.e., Codding, Mercer, Connell, Fiorello, & Kleinert, 2016; Foegen, 2008; Fuchs et al., 1994; Ginsburg et al., 2016; Helwig, Anderson, & Tindal, 2002; Helwig & Tindal, 2002; Hosp et al., 2014; Lee & Lembke, 2016; Shapiro et al., 2015) administered m-CBMs focused on concepts and applications and reported concurrent or predictive validity to a criterion measure of overall mathematical achievement. Concurrent validity coefficients were larger than 0.60 for 6 out of 8 studies. Lee and Lembke (2016) reported validity coefficients of 0.30 for second-graders and 0.55 for third-graders when setting Woodcock Johnson-Broad Math scores as the criterion. Shapiro et al. (2015) reported larger concurrent validity coefficients for the STAR-Math to the state mathematics assessment (0.70-0.88) when compared to validity coefficients for the AIMSweb-Concepts & Applications (0.24-0.49). Three studies administered m-CBMs focused on concepts and applications and reported predictive validity coefficients ranging from 0.40-0.89. Coefficients were larger when the criterion measure was focused on overall mathematics achievement as opposed to more discrete constructs (e.g., problem solving, arithmetic).

Basic facts. Nine studies (i.e., Codding et al., 2016; Foegen, 2000; 2008; Fuchs et al., 2007; Ginsburg et al., 2016; Jiban & Deno, 2007; Lee & Lembke, 2016; Lee, Lembke, Moore, Ginsburg, & Pappas, 2012; Thurber et al., 2002) administered m-CBMs focused on basic facts and reported criterion validity. When a standardized measure of overall mathematical achievement was set as the criterion, concurrent and predictive validity coefficients were between 0.40-0.60 for 6 out of 7 studies. Codding et al. (2016) reported larger predictive validity coefficients when the basic facts intercept (i.e., first probe) was used as the predictor than using the slope. Jiban and Deno (2007) compared a basic facts m-CBM (e.g., 8 – 1 = ?, 4 x 5 = ?) to a Cloze math facts m-CBM (e.g., 8 - ? = 6, 3 x ? = 21) and reported stronger predictive validity for the Cloze math facts m-CBM. Ginsburg et al.'s (2016) findings corroborated this, when administering measures similar to the Cloze math facts, they reported statistically significant predictive validity coefficients. Studies that administered concepts and applications and basic facts m-CBMs reported smaller validity coefficients for basic facts.

Word problems. Four studies (i.e., Fuchs, Fuchs, Karns et al., 2000; Fuches et al., 2003; Jitendra et al., 2005; 2014) administered m-CBMs focused on word problems to analyze the criterion validity to standardized measures of mathematics achievement. Concurrent validity coefficients ranged from 0.58-0.71 for three studies. Two studies (i.e., Jitendra et al., 2005; 2014) reported predictive validity coefficients ranging from 0.38-0.69. Validity coefficients were larger when the standardized assessments focused on overall mathematics achievement or problem solving than on a discrete skill of mathematics achievement (i.e., computation).

Estimation. Three studies (i.e., Foegen, 2000; 2008; Foegen & Deno, 2001) conducted by the same author team administered estimation m-CBMs to analyze the criterion validity to standardized measures of mathematics achievement. Concurrent validity coefficients ranged from 0.45-0.66 when a criterion was selected that measured overall mathematics achievement. Predictive validity coefficients were slightly smaller, evidenced by coefficients ranging from 0.30-0.55 (Foegen, 2008; Foegen & Deno, 2001).

Algebra. Only one study (Johnson, Galow, & Allenger, 2012) administered an m-CBM focused on basic algebra skills with middle and high school students. Predictive validity coefficients comparing these measures to the end of year state assessment ranged from 0.67-0.68 across grades.

Administration protocol. *Format of administration.* A majority of studies administered m-CBMs using either pencil/paper or performance tasks (i.e., 36 studies); however, 6 studies used computers. A finding worth noting was 10 studies administered m-CBMs and a criterion measure via differing formats. Jitendra et al. (2005) administered a word problem measure and standardized measure of overall mathematics achievement via paper/pencil and reported predictive validity coefficients ranging from 0.45-0.65. Smaller predictive validity coefficients (0.38-0.45) were found when Jitendra et al. (2014) administered a similar word problem measure via paper/pencil and via a computer administered mathematics achievement test. Helwig, Anderson et al. (2002) administered a computation measure via paper/pencil and compared it to a computer adapted assessment of mathematics achievement, reporting concurrent validity coefficients ranging from 0.61-0.80. Validity coefficients appeared to be moderated by disability status; the concurrent validity coefficient for general education students was 0.80 compared to 0.61 for students identified with a disability. Seven additional studies reported concurrent validity coefficients ranging from 0.29-0.84 and predictive validity coefficients ranging from 0.30-0.72 when administering an m-CBM and a mathematical achievement assessment with differing administration formats.

Proctor. Researchers (n = 24) served as proctors in a majority of the studies (Teachers = 10). Teachers and researchers administered m-CBMs with varying foci: computation, basic facts, concepts and applications, estimation, numeracy, and teacher ratings. The type of proctor did not appear to moderate the validity coefficients reported across studies.

Grouping. Studies administered m-CBMs using various groupings: 23 classwide, 11 individual, 2 small group, 4 both classwide and individual. The most prevalent m-CBM using individual administration was numeracy (n = 11) due to the nature of the tasks. Only 5 studies administered m-CBMs focused on numeracy using classwide administration, and two of these studies used computers with interactive displays. Two studies (VanDerHeyden et al., 2004; VanDerHeyden, Witt, Naquin, & Noell, 2001) administered similar m-CBMs focused on numeracy with differing groupings (individual, classwide) and reported similar concurrent validity coefficients. Two studies (i.e., Hosp et al., 2014; Thurber et al., 2002) administered m-CBMs focused on computation, concepts and applications, and basic facts in small groups and reported concurrent validity coefficients in the same range as studies that administered similar m-CBMs classwide. At this time, there is no evidence to suggest grouping serves as a moderator to the validity of scores.

Criterion measure. A variety of standardized measures were selected as the criterion: overall mathematics achievement, fluency, computation, concepts and applications, end of year state assessments, national assessments, and math readiness.

Overall mathematics achievement. A variety of m-CBMs were administered to identify the validity coefficients to overall mathematics achievement. Coefficients varied depending on the focus of the m-CBM: concepts and applications (0.30-0.87), numeracy (0.03-0.80), computation (0.14-0.70), basic facts (0.37-0.60), estimation (0.34-0.59), word problems (0.37-0.45), and teacher ratings (0.27-0.44). Scoring of computation m-CBMs provided differential validity coefficients, when scoring problems correct (0.70) validity coefficients were slightly larger than when scoring digits correct (0.67). the same was not true for concepts and applications m-CBMs, when scoring problems correct the validity coefficient was 0.81 and when scoring points the validity coefficient was 0.80. In addition, m-CBMs focused on computation that used the intercept (0.61) reported slightly larger validity coefficients than those using the slope (0.49). For middle school students, the use of complex quantity discrimination tasks (0.52-0.60) had similar validity coefficients to missing number tasks (0.46-0.67).

End of year state assessment. The end of the year state assessments was used as the criterion in 11 studies. Coefficients ranged based on the focus of the m-CBM: concepts and applications (0.24-0.89), computation (0.26-0.75), algebra (0.67-0.68), teacher rating (0.41-0.48), and basic facts (0.43-0.46). Validity coefficients for basic facts and concepts and applications were larger when the intercept was used compared to the slope. Regarding basic facts, Cloze math facts (e.g., 4 x ? = 20) had larger validity coefficients than standard basic fact (e.g., 4 x 5 = ?) probes. One study (i.e., Shapiro et al., 2015) compared a computerized concepts and applications measure to a paper/pencil measure and reported larger validity coefficients for the computer measure (0.70-0.88) than for the paper/pencil measure (0.24-0.61).

Concepts and applications. Many of the studies that administered standardized assessments of overall mathematics achievement also reported data on subtests focused on concepts and applications. Validity coefficients ranged across the focus of m-CBMs: concepts and applications (0.60-0.87), computation (0.33-0.88), numeracy (0.29-0.83), estimation (0.30-0.64), and basic facts (0.32-0.55). Similar to measures of overall mathematics achievement, computation measures that scored problems correct (0.47) reported slightly larger validity coefficients than digits correct (0.40). M-CBMs focused on concepts and applications reported similar validity coefficients whether problems (0.71) or points (0.70) were scored correctly. Studies that administered m-CBMs focused on numeracy reported similar validity coefficients across task types: counting tasks (0.29-0.72), number identification (0.36-0.72), quantity discrimination (0.38-0.79), and missing number (0.68-0.71). Worth noting was one study (Floyd, Hojnoski, & Key, 2006) that reported smaller validity coefficients for counting (0.29-0.33), number identification (0.36), and quantity discrimination (0.38); these were substantially smaller coefficients than those other studies reported. Floyd et al. (2006) did not administer a missing number task, which could explain why missing number tasks have larger validity coefficients.

Computation and math fluency. Subtests focused on computation were used as the criterion by many of the studies. Validity coefficients ranged across the focus of the m-CBMs: computation (0.34-0.93), concepts and applications (0.40-0.77), basic facts (0.14-0.67), and numeracy tasks related to number identification and counting (0.34). Hosp et al. (2014) selected the Woodcock Johnson-Math Fluency subtest as the criterion and reported larger validity coefficients for m-CBMs focused on computation (0.65-0.73) then concepts and applications (0.59-0.63). Contrary to findings reported for overall mathematics achievement and concepts and applications subtest, Hosp et al. (2014) reported larger coefficients when m-CBMs focused on computation were scored for digits correct (0.73) rather than problems correct (0.65). The same was found for m-

CBMs focused on concepts and applications; larger coefficients were reported for points correct (0.63) than for problems correct (0.59). However, this did not hold true when the Woodcock Johnson-Computation subtest was administered, computation measures yielded similar validity coefficients when correct digits (0.75) or correct problems (0.76) were scored. Concepts and applications measures yielded validity coefficients that were slightly larger for correct problems (0.77) than for points (0.74).

National assessments. National assessments from the United States (i.e., National Assessment of Educational Progress), the Netherlands (i.e., CITO), and Germany (i.e., DEMAT1+ and DEMAT2+) were selected as the criterion by four studies. Validity coefficients ranged across the focus of m-CBMs: computation (0.38-0.84), number sense (0.54-0.71), and basic facts (0.45-0.52). Despite having the largest validity coefficients to standardized measures of overall mathematics achievement, no studies included in this review reported the validity of concepts and applications measures to national assessments of mathematics achievement.

Math readiness. Two studies (Floyd et al., 2006; Polignano & Hojnoski, 2012) selected the Bracken Basic Concept Scale-Quantity subtest as the criterion by which to report validity coefficients of m-CBMs focused on numeracy. Validity coefficients varied across numeracy m-CBMs: counting tasks (0.31-0.42), pattern completion tasks (0.59), shape tasks (0.44-0.63), quantity discrimination tasks (0.46), and number identification tasks (0.29). See Table 3 for validity coefficients by study and measures.

Table 3

M-CBMs, Criterion Measures, and Results

Study	Measures	Results
Allinder (1992)	*CBM*	*Concurrent*
	CBM-Comp	CBM-Comp → (.49 -.93)
	Criterion	
	SAT-Comp	
	SAT-Conc	
Baglici (2010)	*CBM*	*Predictive*
	TEN-MN	TEN-MN → (.23 - .58)
	TEN-NI	TEN-NI → (.21 - .57)
	TEN-OC	TEN-OC → (.05 - .39)
	TEN-QD	TEN-QD → (.02 - .51)
	Criterion	
	AIMSweb-Comp	
	Grades	
	ACES-M	
Betts (2009)	*CBM*	*Predictive*
	MKA	MKA-number sense → (.53)
	Criterion	MKA-pattern → (.48)
	NALT	MKA-spatial → (.37)
Chard (2005)	*CBM*	*Predictive*
	C-20	<u>Kindergarten</u>
	CB2	C-20 → (.38 - .41)
	CB5	CB2 → (.45 - .49)
	CB10	CB5 → (.48 - .53)
	CF3	CB10 → (.50 - .55)
	CF6	CF3 → (.40 - .48)
	MN	CF6 → (.39 - .49)
	NI	MN → (.64 - .69)
	NW	NI → (.58 - .65)
	QD	NW → (.57 - .63)
	Criterion	QD → (.50 - .55)
	NKT	<u>First Grade</u>
		C-20 → (.12 - .17)
		CB2 → (.42 - .43)
		CB5 → (.45 - .48)
		CB10 → (.40)
		CF3 → (.07 - .13)
		CF6 → (.18 - .19)
		MN → (.61)
		NI → (.56 - .58)
		NW → (.46 - .54)
		QD → (.45 - .53)

Table 3 *Continued*

Clarke (2004)	*CBM*	*Concurrent*
	MN	MN → (.68 - .75)
	NI	NI → (.60 - .70)

	OC	OC → (.49 - .70)
	QD	QD → (.71 - .80)
	Criterion	*Predictive*
	CBM-Comp	MN → (.67 - .78)
	NKT	NI → (.58 - .72)
	WJ-AP	OC → (.46 - .72)
		QD → (.71 - .79)
Codding (2015)	*CBM*	*Concurrent*
	AIMSweb-Comp	AIMSweb-Comp → (.35)
	Criterion	*Predictive*
	MCAS-M	AIMSweb-Comp → (.26 - .30)
Codding (2016)	*CBM*	*Predictive*
	iSTEEP BF	<u>Sixth Grade</u>
	iSTEEP C&A	BF → (-.01 - .43)
	iSTEEP Core	C&A → (-.21 - .74)
	Criterion	Core → (.13 - .52)
	MCAS-M	<u>Seventh Grade</u>
		BF → (.34 - .46)
		C&A → (-.18 - .77)
		Core → (.46 - .55)
		<u>Eighth Grade</u>
		BF → (-.08 - .45)
		C&A → (-.18 - .89)
		Core → (.34 - .71)
Daly (1997)	*CBM*	*Concurrent*
	NC	NC → WJ-Broad (.47)
	NP	NP → WJ-Broad (.17)
	NR	NR → WJ-Broad (.03)
	NS	NS → WJ-Broad (.11)
	SN	SN → WJ-Broad (.09)
	Criterion	*Predictive*
	BF	NC → BF (.39)
	WJ-Broad	NP → BF (.36)
		NR → BF (.07)
		NS → BF (.30)
		SN → BF (.04)
Eckert (2006)	*CBM*	*Concurrent*
	CBM addition of two digit by two digit	CBM addition of two digit by two digit → (10-96%), r = .09
	CBM subtraction combinations to 18	CBM subtraction combinations to 18 → (13-89%), r = .13
	CBM sum to 9	CBM sum to 9 → (82-97%), r = .20
	CBM sum to 18	CBM sum to 18 → (13-58%), r = .32
	Criterion	
	Teacher Rating	

Table 3 *Continued*

Floyd (2006)	*CBM*	*Concurrent*
	NNF	NNF → (.29 - .70)
	OCF	OCF → (.31 - .55)

	OOCCF	OOCCF → (.29 - .64)
	QCF	QCF → (.38 - .58)
	Criterion	
	BBCS-Q	
	TEMA	
	WJ-AP	
Foegen (2000)	*CBM*	*Concurrent*
	BF	BF → (.35 - .62)
	Estimation	Estimation → (.46 - .66)
	Criterion	
	ITBS	
	Grades	
	Teacher Ranking	
	Teacher Ratings	
Foegen (2001)	*CBM*	*Concurrent*
	BET	BET → (.39 - .56)
	BMOT	BMOT → (.44 - .63)
	MET	MET → (.29 - .55)
	Criterion	*Predictive*
	CAT-Comp	BET → (.32 - .46)
	CAT-C&A	BMOT → (.32 - .45)
	GPA	MET → (.20 - .30)
	Teacher Rating	
Foegen (2008)	*CBM*	*Concurrent*
	BF	<u>Sixth Grade</u>
	Complex QD	BF → (.38 - .56)
	Estimation	Complex QD → (.50 - .57)
	MBSP-Comp	Estimation → (.51 - .60)
	MBSP-C&A	MBSP-Comp → (.54 - .65)
	MN	MBSP-C&A → (.60 - .76)
		MN → (.45 - .56)
	Criterion	<u>Seventh Grade</u>
	Grades	Basic facts → (.52 - .60)
	ITBS	Complex QD → (.42 - .60)
	NALT	Estimation → (.26 - .51)
	Teacher Rating	MBSP-Comp → (.33 - .38)
		MBSP-C&A → (.73 - .87)
		MN → (.42 - .67)
		<u>Eighth Grade</u>
		Complex QD → (.41 - .55)
		Estimation → (.40 - .53)
		MN → (.31 - .50)

Table 3 *Continued*

		Predictive
		<u>Sixth Grade</u>
		BF → (.55)

		Complex QD → (.53)
		Estimation → (.55)
		MBSP-Comp → (.64)
		MBSP-C&A → (.76)
		MN → (.48)
		<u>Seventh Grade</u>
		BF → (.59)
		Complex QD → (.58)
		Estimation → (.34)
		MBSP-Comp → (.25)
		MBSP-C&A → (.87)
		MN → (.60)
Fuchs (1994)	*CBM*	*Concurrent*
	CBM-C&A	<u>Second Grade</u>
	Criterion	CBM-C&A → (.74 - .81)
	CTBS-Comp	<u>Third Grade</u>
	CTBS-C&A	CBM-C&A → (.64 - .74)
	CTBS-Total	<u>Fourth Grade</u>
		CBM-C&A → (.74 - .79)
Fuchs (2000)	*CBM*	*Concurrent*
	PAs	PAs → (.48 - .68)
	Criterion	
	CTBS	
	ITBS	
	MOAT	
Fuchs (2003)	*CBM*	*Concurrent*
	Far WP	Far WP → (.67)
	Immediate WP	Immediate WP → (0.58)
	Near WP	Near WP → (0.55)
	Criterion	
	TerraNova	
Fuchs (2007)	*CBM*	*Predictive*
	CBM-Comp	CBM-Comp → (.28 - .35)
	CBM-C&A	CBM-C&A → (.40 - .44)
	Fact Retrieval	Fact retrieval → (.10 - .14)
	NI/C	NI/C → (-.19 - .39)
	Criterion	
	WPs	
	WRAT-Arithmetic	

Table 3 *Continued*

Ginsburg (2016)	*CBM*	*Concurrent*
	BF	<u>Kindergarten</u>
	CBM-C&A	CBM Risk → CI-Addition (.32 - .37), CI-

	CBM-Comp	Counting (.11 - .54), CI-Subtraction (.19 - .29)
	CI-Addition	<u>First Grade</u>
	CI-Counting	CBM Risk → CI-Addition (.21 - .43), CI-Counting (.07 - .42), CI-Subtraction (.29 - .41), CI-Written Number (.04 - .17)
	CI-Multiplication	
	CI-Subtraction	
	CI-Written Number	<u>Second Grade</u>
	Counting	CBM Risk → CI-Addition (.09 - .42), CI-Multiplication (.14 - .33), CI-Subtraction (.13 - .32), CI-Written Number (.06 - .32)
	MN	
	NI	<u>Third Grade</u>
	NN	CBM Risk → CI-Addition (.07 - .29), CI-Multiplication (.20 - .35), CI-Subtraction (.12 - .28), CI-Written Number (.14 - .29)
	QD	
	Criterion	
	WJ-Broad	*Predictive*
		<u>Kindergarten</u>
		Significant predictor to WJ-App → CI-Addition
		Significant predictor to WJ-Broad → CI-Subtraction
		Significant predictors to WJ-Calc → CI-Addition, CI-Subtraction, MN, QD
		Significant predictors to WJ-Math Fluency → CI-Addition, CI-Subtraction, MN
		<u>First Grade</u>
		Significant predictor to WJ-App → CI-Addition
		Significant predictor to WJ-Broad → CI-Written Number
		Significant predictors to WJ-Calc → BF, CI-Addition, CI-Written Number
		Significant predictors to WJ-Math Fluency → BF, CI-Addition, CI-Written Number
		<u>Second Grade</u>
		Significant predictor to WJ-App → None
		Significant predictor to WJ-Broad → BF, CI-Multiplication, CI-Written Number
		Significant predictor to WJ-Calc → None
		Significant predictor to WJ-Fluency → None
		<u>Third Grade</u>
		Significant predictors to WJ-App → CBM-C&A, CI- Multiplication, CI-Subtraction
		Significant predictors to WJ-Broad → CBM-C&A, CI-Multiplication, CI-Subtraction
		Significant predictors to WJ-Calc → CBM-Comp, CBM-C&A, CI-Multiplication, CI-Subtraction
		Significant predictors to WJ-Math Fluency → BF, CBM-Comp, MN, QD

Table 3 *Continued*

Helwig (2002)	*CBM*	*Concurrent*
	CBM-C&A	<u>General Education</u>
	Criterion	CBM-C&A → (.80)

	CAT-MA	Special Education
		CBM-C&A → (.61)
		Predictive
		CBM-C&A + Classification (Sped/Not) → Passing CAT-MA (87.1% correct)
Helwig & Tindal (2002)	*CBM*	*Concurrent*
	GOMs	GOM 3 → (.84)
	Criterion	GOM 4 → (.82)
	State Assessment	*Predictive*
		GOM 1 → (.81)
		GOM 2 → (.87)
Hosp (2014)	*CBM*	*Concurrent*
	AIMSweb-Comp	AIMSweb-Comp-Correct Digits → (.40 - .75)
	AIMSweb-C&A	AIMSweb-Comp-Correct Problems → (.47 - .76)
	Criterion	AIMSweb-C&A Correct Problems → (.59 - .81)
	WJ-App	AIMSweb-C&A Points → (.63 - .80)
	WJ-Broad	
	WJ-Comp	
	WJ-Math	
	Fluency	
Jiban (2007)	*CBM*	*Predictive*
	BF	Third Grade
	Cloze Math Facts	BF problems correct → MCA-Math (.11)
	Criterion	BFcorrect minus incorrect → MCA-Math (.26)
	MCA-Math	Cloze Math Facts problems correct → MCA-Math (.38)
		Cloze Math Facts correct minus incorrect → MCA-Math (.44)
		Fifth Grade
		BF problems correct → MCA-Math (.55)
		BFcorrect minus incorrect → MCA-Math (.57)
		Cloze Math Facts problems correct → MCA-Math (.59)
		Cloze Math Facts correct minus incorrect → MCA-Math (.59)
Jitendra (2005)	*CBM*	*Concurrent*
	CBM-Comp	CBM-Comp → (.45 - .66)
	WPs	WPS → (.38 - .71)
	Criterion	*Predictive*
	TerraNova-Comp	CBM-Comp → (.38 - .69)
	TerraNova-C&A	
	SAT-Procedures	
	SAT-PS	

Table 3 *Continued*

Jitendra (2014)	*CBM*	*Predictive*
	WPs	Time 1: WPs → (.38)
	Criterion	Time 2: WPs → (.37)

	MAP-M	Time 3: WPs → (.45)
Johnson (2012)	*CBM*	*Predictive*
	Basic Skills	Seventh Grade
	Algebra	Basic Skills-Algebra → (.67)
	Criterion	Eighth Grade
	ISAT	Basic Skills-Algebra → (.68)
		Tenth Grade
		Basic Skills-Algebra → (.68)
Kettler (2013)	*CBM*	*Concurrent*
	MBSP-Comp	First Grade
	PSG-M	MBSP-Comp → (.63)
	Criterion	PSG Math → (.74)
	MAP-M	Second Grade
	SAT	MBSP-Comp → (.66)
	WKCE	PSG Math → (66)
		Third Grade
		MBSP-Comp → (.61)
		PSG Math → (.68)
		Predictive
		First Grade
		MBSP-Comp → (.44)
		PSG Math → SAT (.38)
		Second Grade
		MBSP-Comp → (.14 - .29)
		PSG Math → (.41 - .44)
		Third Grade
		MBSP-Comp → (.25 - .37)
		PSG Math → (.27 - .48)
Klinkenberg (2011)	*CBM*	*Concurrent*
	Maths Garden	Maths Garden: Addition → (.83)
	Criterion	Maths Garden: Division → (.78)
	CITO	Maths Garden: Multiplication → (.80)
		Maths Garden: Subtraction → (.84)
Laracy (2016)	*CBM*	*Predictive*
	NN	NN → (.72 - .76)
	OC	OC → (.70 - .77)
	OOCC	OOCC → (.65 - .71)
	QC	QC → (.72 - .82)
	Criterion	
	TEN-QD	

Table 3 *Continued*

Lee (2012)	*CBM*	*Concurrent*
	BF	Kindergarten
	MN	MN → (.62)

	NI	NI → (.68)	
	NN	OC → (.53)	
	OC	QD → (.64)	
	QD		First Grade
	Criterion	BF → (.50)	
	TEMA	MN → (.56)	
		NI → (.68)	
		NN → (.59)	
		OC → (.40)	
		QD → (.48)	
		Predictive	
			Kindergarten
		MN → (.59)	
		NI → (.45)	
		OC → (.54)	
		QD → (.51)	
			First Grade
		BF → (.37)	
		MN → (.37)	
		NI → (.53)	
		NN → (.55)	
		OC → (.25)	
		QD → (.40)	
Lee (2016)	*CBM*	*Concurrent*	
	BF		Kindergarten
	CBM-Comp	MN → (.37)	
	CBM-C&A	NI → (.32)	
	MN	OC → (.36)	
	NI	QD → (.44)	
	NN		First Grade
	OC	BF → (.52)	
	QD	MN → (.35)	
	Criterion	NI → (.46)	
	WJ-Broad	NN → (.51)	
		OC → (.14)	
		QD → (.51)	
			Second Grade
		BF → (.58)	
		CBM-Comp → (.32)	
		CBM-C&A → (.30)	
		MN → (.48)	
		QD → (.25)	

Table 3 *Continued*

			Third Grade
		BF → (.49)	
		CBM-Comp → (.46)	

		CBM-C&A → (.55)
		MN → (.54)
		QD → (.35)
Methe (2008)	*CBM*	*Concurrent*
	COF	COF → (.50 - .70)
	MQF	MQF → (.20 - .70)
	NRF	NRF → (.64 - .89)
	OPF	OPF → (.60 - .81)
	Criterion	*Predictive*
	Teacher Rating	COF → (.46 - .70)
	TEMA	MQF → (.41 - .72)
		NRF → (.66 - .88)
		OPF → (.58 - .79)
Polignano (2012)	*CBM*	*Concurrent*
	CAR	CAR → (.41 - .72)
	PC	PC → (.22 - .72)
	SC	SC → (.26 - .53)
	SNF	SNF → (.23 - .56)
	SSF	SSF → (.28 - .64)
	Criterion	
	BBCS-Q	
	PNI-NNF	
	PNI-OCF	
	PNI-OOCCF	
	PNI-QCF	
	TEMA	
Salaschek (2013)	*CBM*	*Concurrent*
	Addition	CBMs → (.40-.50)
	Equation	*Predictive*
	ND	CBMs → (.29 - .71)
	NI	
	NL	
	NSeq1	
	NSeq2	
	Subtraction	
	SQD	
	Criterion	
	DEMAT 1+	
	DEMAT 2+	
	OTZ	
	Teacher rating	

Table 3 *Continued*

Salaschek (2014)	*CBM*	*Concurrent*
	Number Sense	CBM-Comp → (.49 - .56)
	CBM-Comp	Number Sense → (.54 - .62)

	Criterion	CBM-Comp + Number Sense → (.57 - .63)
	DEMAT1+	*Predictive*
	DEMAT2+	CBM-Comp → (.64 - .72)
	Teacher Rating	Number Sense → (.66 - .69)
		CBM-Comp + Number Sense → (.66 - .77)
Seethaler (2011)	*CBM*	*Concurrent*
	CBM-Comp	CBM-Comp → (.69)
	Criterion	*Predictive*
	TEMA	CBM-Comp → (.49 - .61)
Shapiro (2015)	*CBM*	*Concurrent*
	AIMSweb-Comp	Third Grade
	AIMSweb-C&A	AIMSweb-Comp → (.61)
	STAR-Math	AIMSweb-C&A → (.61)
	Criterion	STAR-Math → (.82)
	PSSA	Fourth Grade
		STAR-Math → (.88)
		AIMSweb-Comp → (.75)
		AIMSweb-C&A → (.24)
		Fifth Grade
		AIMSweb-Comp → (.74)
		AIMSweb-C&A → (.49)
		STAR-Math → (.70)
		Predictive
		Third Grade
		AIMSweb-Comp + AIMSweb-Comp (slope) + AIMSweb-C&A + AIMSweb-C&A (slope) + STAR-Math + STAR-Math (slope) → (72%)
		Fourth Grade
		AIMSweb-Comp + AIMSweb-Comp (slope) + AIMSweb-C&A + AIMSweb-C&A (slope) + STAR-Math + STAR-Math (slope) → (82%)
		Fifth Grade
		AIMSweb-Comp + AIMSweb-Comp (slope) + AIMSweb-C&A + AIMSweb-C&A (slope) + STAR-Math + STAR-Math (slope) → (71%)
Thurber (2002)	*CBM*	*Concurrent*
	BF	BF → (.45 - .67)
	CBM-Comp	CBM-Comp → (.36 - .63)
	Criterion	
	CAT-Comp	
	CAT-C&A	
	NAEP	
	SDMT-App	
	SDMT-Comp	

Table 3 *Continued*

VanDerHeyden (2001)	*CBM*	*Concurrent*
	Circle number	Circle number → (.30 - .61)
	Discrimination	Discrimination → (.56)

	Draw circles	Draw circles → (.45 - .64)
	Write number	Write number → (.43 - .44)
	Criterion	
	CIBS-Q	
	Teacher Ranking	
VanDerHeyden (2004)	*CBM*	*Concurrent*
	Choose number	Choose number → (.43 - .81)
	Choose shape	Choose shape → (.06 - .57)
	Count objects	Count objects → (.43 - .56)
	Discrimination	Discrimination → (.50 - .85)
	Free count	Free count → (.19 - .91)
	Number naming	Number naming → (.39 - .91)
	Criterion	
	CIBS-Q	
	Teacher Rating	
	TEMA	

Note. See Appendix B for full validity results and Appendix C for nomenclature regarding assessment abbreviations

Discussion

This study reviewed the current empirical literature to examine: (a) the characteristics of the literature reporting criterion validity of m-CBMs, (b) the criterion validity evidence related to m-CBMs, (c) the administration variables and type of m-CBM that may moderate the criterion validity of m-CBMs, and (d) the criterion measures with the largest and weakest criterion validity evidence. The discussion was structured to highlight how the current review expands upon findings from Christ et al. (2008) and Foegen et al.'s (2007) previous reviews. The primary findings related to evidence of criterion validity, potential moderators to criterion validity, and the selection of criterion measures are discussed.

The current literature review included 39 peer-reviewed articles that administered m-CBMs and reported evidence of criterion validity. Two prior reviews, conducted nearly a decade ago (Christ et al., 2008; Foegen et al., 2007), examined the literature base related to m-CBMs finding 9 and 17 studies that reported criterion validity evidence related to m-CBMs. Results of this 2017 work yield twice as many studies related to m-CBMs focused on computation (Christ et al., 2008) and more than twice as many studies that administered m-CBMs and reported criterion validity (Foegen et al., 2007). Additionally, the current study extends previous literature reviews in two areas: (a) the analysis of administration protocols that may moderate the criterion validity of measures and (b) the analysis of how score validity was moderated by the type of criterion measure selected.

Current Literature Base

A primary finding related to the characteristics of the literature base was the increased number of studies that included students in secondary settings (e.g., middle, high) or pre-kindergarten and early-elementary (K-2) settings. The increased emphasis on early childhood programs (e.g., Head Start) and early intervention in an MTSS framework necessitate the need to identify m-CBMs with validity evidence for these populations of students. This systematic review suggests scholars have begun examining this field; however, additional research is needed. Furthermore, the increased literature base on middle school populations is promising. Recent trends from NAEP (2016) data suggest that as students age from elementary to middle school, the percentage of students reaching grade level expectations decreases.

Criterion Validity Evidence

A primary finding related to the criterion validity is the strong predictive and concurrent validity coefficients for concepts and applications m-CBMs administered with upper elementary and middle school students. This finding is not surprising when considering the distinction Fuchs (2004) made toward CBMs and single-skill measurement:

> CBM tasks are multidimensional, requiring students to simultaneously integrate the various skills required for competent year-end performance…some recent measures, which have been labeled CBM, appear to represent single-skill measurement…this seems unfortunate, especially because few studies document that single-skill measures can be used to model global learning over time. (p. 191).

However, concepts and applications measures are not all identical, Shapiro et al. (2015) reported significantly stronger validity for STAR-M than AIMSweb-Concepts and Applications to the end-of-year state assessment. Contrarily, Codding et al. (2016) reported smaller validity coefficients for an m-CBM aligned to the grade level standards (i.e., iSTEEP-Common Core) than a more general concepts and applications measure (i.e., iSTEEP-Concepts and Applications) when selecting the end-of-year state assessment as the criterion. The selection of a concepts and applications measure should be made judiciously by considering the criterion assessment, the administration protocol, and the criterion validity evidence reported.

A final finding, albeit based on data from one study (Helwig et al., 2002), was the validity coefficients for general education students were larger than for students identified with disabilities when the criterion was set as a computer administered assessments of mathematics achievement. The generalization of this finding would be problematic for schools using m-CBM data to make low- and high-stakes decisions for students identified with disabilities (Gersten et al., 2009). Special education teachers use data (e.g., obtained from m-CBMs) to determine whether students are making progress toward their annual educational goals listed in their Individualized Education Program (IEP); however, if the data lack validity, then these decisions regarding the effectiveness of the student's program are inaccurate. Furthermore, as more states continue to adopt computer administered end-of-year examinations, the effects on score validity for students with disabilities should be considered (Thompson, Thurlow, & Moore, 2003).

Another promising finding was the strong validity reported for numeracy indicators with early elementary students. Numeracy m-CBMs that administered tasks consisting of counting, number identification, quantity discrimination, and missing number yielded similar validity coefficients with kindergarteners. However, as students aged to first and second grades, missing number tasks yielded the strongest criterion validity for a majority of studies. Furthermore, missing number measures yielded stronger validity coefficients than m-CBMs focused on basic facts, computation, and concepts and applications for third grade students (Lee & Lembke, 2016). Two numeracy m-CBMs, complex discrimination and missing numbers, were administered to middle school students; however, validity coefficients were smaller than m-CBMs focused on concepts and applications (Foegen, 2008).

A final point regarding the types of m-CBMs is both computation and basic fact measures have criterion validity evidence to discrete mathematical skills (e.g., computation, arithmetic). A majority of studies that administered m-CBMs with varying foci reported the strongest validity coefficients for basic facts or computation measures to a criterion related to computation or mathematical fluency. The data were inconclusive regarding whether basic fact or computation m-CBMs yielded larger validity coefficients (Foegen, 2008; Fuchs et al., 2007; Lee & Lembke, 2016; Thurber et al., 2002).

Potential Moderators

Initially, four potential variables related to administration protocol were considered for moderator analyses: administration format, type of proctor, grouping used for administration, and timing. Timing was excluded because of the lack of data experimentally comparing validity coefficients for similar measures under timed and untimed conditions. Results for administration grouping and proctor used did not have evidence suggesting these would moderate score validity. It is worth noting a majority of studies used researchers as proctors; studies using teachers as proctors appear to have varying levels of training associated with standardized administration format and scoring. Furthermore, only eight studies reported data regarding the fidelity of administration, which could potentially affect score validity (Lievens & Patterson, 2011).

Administering m-CBMs via paper/pencil compared to computer is an area that may potentially affect criterion validity of scores; a portion of the variance in students' scores may be attributed to their ability and prior experiences navigating the computer software (Noyes & Garland, 2008; Wang, Jiao, Young, Brookes, & Olson, 2007; 2008). Given the current data, results are inconclusive; studies did not administer similar computer and paper/pencil m-CBMs to report criterion validity to a paper/pencil and computer administered criterion of mathematics achievement. An area worth further investigation is identifying whether a mismatch between administration format of the m-CBM and criterion affects the criterion validity (e.g., m-CBM via computers, criterion via paper/pencil). In addition, a majority of the m-CBMs focused on numeracy required performance-based tasks that proctors scored real time. Fidelity of adherence to the administration and scoring protocol likely will dictate score validity (Lee, Reynolds, & Willson, 2003); thus, teachers must receive adequate training on administering and scoring measures.

Limitations

Several limitations should be considered when interpreting findings. First, during the literature search the decision was made to only accept peer-reviewed articles, which may lead to publication bias. However, this decision was made to increase the likelihood of retaining the most rigorous scientific research (Higgins & Green, 2011). Second, administration fidelity was not reported by 32 (82.0%) of the studies, which could affect score validity. Lastly, results should be interpreted with caution; more robust claims can be made when: (a) sample sizes, (b) the quality of validity coefficients, and (c) study designs are considered meta-analytically.

Considerations for Future Research

A significant finding from this review was the amount of unknown questions regarding how the criterion validity of an m-CBM may be moderated across differing administration protocols along with different populations of students. First, administration protocols for m-CBMs and the criterion measures may contribute to score validity. Using meta-analytic techniques to analyze how differences in administration protocols moderate score validity would provide claims that are more robust. Specifically, empirical research that isolates the effect of using computers to administer assessments and how this affects score validity, for example, the mismatch between administration format of the m-CBM and the criterion measure (e.g., m-CBM via paper/pencil, criterion via computer). A portion of the variance in students' scores may be attributed to their ability and prior experiences navigating the computer software. Additionally, examining the applicability of m-CBMs to varying student populations is needed. This study echoes claims by Foegen (2008) that additional research on m-CBMs with secondary students is needed. Furthermore, as pre-kindergarten programs grow nationwide, additional research on m-CBMs used for screening and progress monitoring purposes is needed.

Suggestions for Practitioners

Findings from this review provide some suggestions stakeholders. Currently, there is a dearth of empirical evidence to guide practitioners working in pre-kindergarten or high school settings. Educators in upper elementary and middle school settings are encouraged to select m-CBMs focused on concepts and applications for universal screening and progress monitoring. However, it is vital that educators ensure the m-CBMs align to the state mathematics standards; administering an m-CBM created to represent the *Common Core State Standards for Mathematics* (CCSS-M) may not align closely for educators in states that did not adopt the CCSS-M (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). In early elementary settings, using m-CBMs focused on numeracy (i.e., counting, number identification, quantity discrimination, missing number) will aid in the identification of students who are likely to demonstrate mathematical deficits. Worth noting is teachers in first and second grade should place more emphasis on students' performance on quantity discrimination and missing number tasks. A last consideration for practitioners is m-

CBMs focused on basic facts and computations provide information on a student's performance in discrete areas of mathematics (i.e., computation, arithmetic). These skills are vital to overall mathematics achievement; thus, using data from m-CBMs focused on basic facts and computation can benefit instructional decision-making.

CRITERION VALIDITY OF A COMPUTER ADAPTIVE CURRICULUM-BASED MEASURE TO AN END OF YEAR STATE MATHEMATICS ASSESSMENT

Introduction

The mathematical performance of US students has been examined and deliberated by stakeholders in the educational field. The cause for concern is evidenced by data reported from the National Assessment of Educational Progress (NAEP; US Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2016), which found 60% of fourth-graders and 67% of eighth-graders failed to reach mathematical proficiency. Stakeholders may postulate, “The NAEP standards are too rigorous, thus, the lack of student achievement.” Data from the OECD Programme for International Student Assessment (PISA; OECD, 2016) contradicts this hypothesis. PISA is an international assessment that was administered to students in 72 participating nations. Three findings from PISA (2015) are worth noting: (a) US students scored below the mean, (b) US students scored worse in 2015 than they scored three years ago, and (c) the US ranked 38th out of 72 participating nations.

The passing of Every Student Succeeds Act (ESSA, 2015) responded to these trends by providing state departments of education flexibility in their testing procedures. However, ESSA (2015) retained requirements about reporting results on specific subgroups of students (i.e., ethnicity, socioeconomic status, disability) to identify potential gaps in achievement. Thus, schools are tasked with ensuring systems are in place to identify students failing to reach the expected criteria and remediate these deficits.

Curriculum-Based Measures

Deno (1985) published an influential article on the use of curriculum-based measures (CBMs) to monitor student progress over time and gage effectiveness of instruction. The ingenuity of this idea was the conceptualization of CBMs as *simple indicators* that provide current data on performance and predict future student achievement. *Simple* implies the measure is quick and easy to administer and multiple alternative forms can be generated; *indicator* implies CBMs' ability to predict student performance on achievement tests. Deno (1985) suggested empirical research must be conducted across three strands: (a) identifying measures with technical adequacy, (b) testing efficiency of teachers implementing CBMs, and (c) testing the efficacy of designing interventions to improve teachers' decision-making. The focus of this research was to analyze the technical adequacy, specifically validity, of the i-Ready Diagnostic assessment.

Recently, two literature reviews analyzed the technical adequacy of mathematics CBMs (m-CBMs; Christ et al., 2008; Foegen et al., 2007). Christ et al. (2008) identified nine empirical studies (9 elementary, 3 middle, 1 high school) that reported information regarding the validity of m-CBMs. Results indicated that m-CBMs focused on computation had sufficiency evidence for screening-type decision-making; however, the authors emphasized that m-CBMs targeting computation do not measure the entirety of the latent construct mathematical achievement. Foegen et al. (2007) identified 17 studies (7 included early elementary students, 9 included upper elementary students, 6 included middle school students, and 0 included high school students) that reported evidence of criterion validity. The authors reported that m-CBMs targetting concepts and applications had a stronger correlation with overall mathematics achievement than m-CBMs targeting computation. In addition, the authors noted m-CBMs targeting mathematical problem solving were moderate to strongly correlated to overall mathematics achievement measures. Both Christ et al. (2008) and Foegen et al. (2007) reported the need for future empirical research on (a) the application of m-CBMs for middle school students, (b) the validity of m-CBMs to end of the year state assessments, and (c) potential administration procedures that may moderate the technical adequacy. This research addressed the application of i-Ready Diagnostic to elementary and middle school students and the analysis of potential variables that may moderate the technical adequacy.

Lastly, the literature review conducted included an updated literature search and insights into future research questions. Strong concurrent and predictive validity coefficients were found for m-CBMs focused on concepts and applications when the end-of-year state assessment was selected as the criterion. One study (Shapiro et al., 2015) administered a computer adaptive m-CBM and a paper/pencil m-CBM both focused on concepts and applications and reported stronger concurrent and predictive validity coefficients for the computer adaptive measure to the end-of-year state mathematics assessment. However, no study had administered an m-CBM focused on concepts and applications and examined the concurrent and predictive validity to the end-of-year state assessment. Additionally, the analysis of how student demographic variables moderate the criterion validity of the m-CBM was considered in few studies.

One study (Helwig et al., 2002) segregated the criterion validity coefficients for general education and special education students; students receiving special education services had much lower validity coefficients for m-CBMs focused on concepts and applications (Helwig et al., 2002). The current study examined additional demographic variables (i.e., gender, race, free and reduced meals, Section 504) in addition to special education status to determine if these may moderate the criterion validity of an m-CBM. Furthermore, the current study examined the criterion validity of an m-CBM focused on concepts and applications for third through eighth grade students, which extends the literature on the criterion validity of concepts and applications measures to end-of-year state assesssments for middle school students.

i-Ready Diagnostic

Curriculum Associates LLC developed the i-Ready program to provide data and instructional supports for teachers and administers. The i-Ready Diagnostic serves as a screener, providing skill level deficits and prediction to end of the year achievement. The i-Ready Diagnostic is an adaptive computer-based assessment that samples from the *Common Core State Standards for Mathematics* (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). The Curriculum Associates LLC funded research conducted by the Educational Research Institute of America (ERIA), a private educational research company. ERIA reported moderate to strong concurrent validity for i-Ready Diagnostic to: (a) the Partnership for Assessment of Readiness for College and Careers (PARCC; 0.77 – 0.84), (b) SmartBalanced (0.82 – 0.85), and (c) New York State Common Core Assessments (0.80 – 0.86). Furthermore, ERIA reported strong validity when administering the i-Ready Diagnostic during the fall and winter to predictive student performance on the end of the year state mathematics assessment. However, there is a need for independent empirical research to be conducted with considerations paid to potential moderators of the technical adequacy of the i-Ready Diagnostic.

Current Study

The current study extends previous research by analyzing the technical adequacy of an m-CBM to elementary and middle school students. In addition, the predictive and concurrent validity of an m-CBM to an end of year state assessment was analyzed. Lastly, variables that may moderate the technical adequacy of an m-CBM were analyzed. The following research questions were addressed:

1. What are the predictive and concurrent validity of a computer adaptive m-CBM (i.e., i-Ready Diagnostic) to an end-of-year state assessment (i.e., LEAP 2025)?
2. How do the predictive and concurrent validity of the computer adaptive m-CBMs vary across time (fall, winter) and across domains of the end-of-year state assessment (i.e., major content, expressing mathematical reasoning, modeling and application, additional and supporting content)?
3. Is the predictive or concurrent validity moderated by demographic variables of participants: (a) gender, (b) race, (c) free and reduced meals (FARMS), (d) limited English proficiency (LEP), (e) special education, and (f) Section 504?

Method

Participants and Setting

The number of participants varied across time (i.e., fall, winter, spring) due to the high mobility of the school district and missing data. Sample sizes ranged from: 121 to 143 for third grade (84.6% complete cases across measures), 124 to 141 for fourth grade (87.9% complete cases across measures), 164 to 192 for fifth grade (85.4% complete cases across measures), 157 to 180 for sixth grade (87.2% comple cases across measures), 187 to 285 for seventh grade (65.6% comple cases across measures), and 188 to 254 for eighth grade (74.0% complete cases across measures). Students were enrolled in one elementary school (3-4), one intermediate school (5-6), and two junior high schools (7-8) located in a rural area within Louisiana during the 2016-2017 school year. All students, including students with disabilities, receiving instruction in the general education classroom were included. Across schools, instructional minutes for mathematics varied from 60-75 minutes per day. All students received instruction aligned to the Louisiana Student Standards for Mathematics. All schools in the district implemented *Eureka Math* (https://greatminds.org/math), which is considered a research-based core curriculum by the Louisiana Board of Elementary and Secondary Education (See Tables 4-5 for complete sample demographics).

Table 4

Student Demographic Information for Third- through Sixth-Graders

	Elementary		Intermediate	
	Grade 3 (n = 143)	Grade 4 (n = 141)	Grade 5 (n = 192)	Grade 6 (n = 180)
Gender				
Male	72	72	103	98
Female	71	69	89	82
Race				
White	79	80	85	86
African-American	33	33	77	63
Hispanic	23	18	13	13
Native Hawaiian or Pacific Islander	6	2	3	1
Two or more	4	4	5	14
Asian	1	4	6	3
American Indian or Alaska Native	0	0	1	0
Special education	15	11	22	24
Section 504	8	11	21	20
Free and reduced meals	27	22	74	95
LEP	6	5	2	4

Note. LEP = Limited English Proficiency

Table 5

Student Demographic Information for Seventh- and Eighth-Graders

	Jr. High 1		Jr. High 2	
	Grade 7 (n = 195)	Grade 8 (n = 164)	Grade 7 (n = 90)	Grade 8 (n = 90)
Gender				
Male	99	81	47	47
Female	96	83	43	43
Race				
White	83	82	53	47
African-American	62	59	17	18
Hispanic	34	13	7	11
Native Hawaiian or Pacific Islander	2	2	2	2
Two or more	6	5	8	9
Asian	3	3	3	3
American Indian or Alaska Native	5	0	0	0
Special education	26	15	11	10
Section 504	13	8	7	4
Free and reduced meals	89	65	46	42
LEP	3	2	0	2

Note. LEP = Limited English Proficiency

Experimental Measure

i-Ready Diagnostic. A computer-adaptive assessment of mathematical skills designed for K-12 was used. The i-Ready Diagnostic assesses skills across four domains encompassed by the *Common Core State Standards for Mathematics*: Number and Operations, Measurement and Data, Geometry, and Algebra (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). Items are selected dynamically, meaning specific items were selected based on student responses. The assessment has no time constraint. The measure was administered classwide in a computer lab; each student accessed the assessment via a desktop computer. Students who transferred into the school district were administered the measure within two weeks of enrolling in school. The i-Ready Diagnostic was administered individually, the student accessed the measure via a desktop computer.

Curriculum Associates, LLC. reported that i-Reading Diagnostic has evidence of strong reliability and validity; however, no specific reliability coefficients were reported. Evidence of concurrent validity has been reported in the moderate range (SmartBalanced [0.86], PARCC [0.78], New York State Common Core Assessment [0.84]); also reported is strong predictive validity (AUC > 0.89) across end of year state assessments (Curriculum Associates, LLC., 2015; 2015; 2016). Scaled scores are derived by a linear transformation of Rasch ability estimates on the logit scale to make all scores positive integers. Rasch ability estimates accounts for the students "estimated ability" and the question difficulty to control for guessing. The i-Ready Diagnostic scaled scores range from 0 to 800 and were used for all analyses in this study.

Criterion Measure

Louisiana Education Assessment Program 2025. The Louisiana Education Assessment Program 2025 (LEAP 2025) is a standards-based, criterion-referenced measure that measures students' performance on the Louisiana Student Standards for Mathematics**.** All students in third through eighth grade were assessed during the spring semester. The assessment was administered paper-pencil for third and fourth grades and computer-based for fifth through eighth grades. Students completed the assessment across three sessions that ranged from 70-90 minutes in duration. Items were classified into one of three types: conceptual/fluency, arguments/justifications, modeling/application. Reliability coefficients were not reported by the state department of education. Individual student scores were reported as scaled scores, ranging from 650 to 850, that were linear transformations of the Rasch ability estimates. Each student's mathematics achievement performance was classified as "Unsatisfactory", "Approaching Basic", "Basic", "Mastery", or "Advanced." The classifications were based on cutpoints established by the scaled scores; the cutpoints varied across grades. For the data analysis purposes, students' mathematics achievement was coded as 0 through 4. Furthermore, student performance across the following LEAP domains: major content, expressing mathematical reasoning, modeling & application, and additional & supporting content, were reported as "Weak", "Moderate", or "Strong." For data analysis purposes, student performance in each LEAP domain was coded as 0 through 3. Scaled scores and proficiency in the LEAP domains were used for the analyses in this study.

Procedures

All participants were administered the i-Ready Diagnostic measure at three different time periods (i.e., fall, winter, spring) during the 2016-2017 school year. The assessment was administered classwide in the computer lab; each student accessed the assessment via a desktop computer. New students were administered the i-Ready Diagnostic within two weeks upon enrolling in the school; scores obtained from newly enrolled students were categorized with data from the most recent administration date. The LEAP 2025 was administered during the state dictated assessment window (i.e., April 3-May 5). The third and fourth grade assessment was administered via paper-pencil by a trained proctor who followed standardized assessment protocol. Fifth-through eighth-graders completed a computer-based assessment that was administered in a computer lab by a trained proctor who followed standardized assessment protocol.

Data Analysis

All analyses were conducted using STATA, version 14. To partially address Research Questions #1 and #2, correlational analyses were used to examine the relationship between the i-Ready Diagnostic scores from a given time frame (i.e., fall, winter, spring) to the LEAP 2025 scaled scores and LEAP 2025 domain proficiencies. Multiple linear regression analyses were used to address Research Question #3 to determine if demographic variables moderated the concurrent and/or predictive validity of the i-Ready Diagnostic to the LEAP 2025.

To analyze the predictive validity of the i-Ready Diagnostic, correlational analyses were used comparing data from the i-Ready-fall and i-Ready-winter administrations to the LEAP 2025 scaled scores and LEAP 2025 domains (i.e., major content, expressing mathematical reasoning, modeling & application, additional & supporting content). Concurrent validity of the i-Ready Diagnostic was analyzed by conducting correlational analyses comparing data from the i-Ready-spring administration to the LEAP 2025 scaled scores and LEAP 2025 domains (i.e., major content, expressing mathematical reasoning, modeling & application, additional & supporting content).

To examine whether demographic variables moderated the predictive and/or concurrent validity of the i-Ready Diagnostic multiple linear regression analyses were used. First, multiple linear regression was used to analyze the predictive validity of the fall i-Ready Diagnostic to the LEAP 2025 by using the current model:

$$\text{LEAP 2025 Scaled Score}_i = \beta_0 + \beta_1\text{i-Ready} + \beta_2\text{Gender} + \beta_3\text{Race} + \beta_4\text{SPED} + \beta_5\text{FARMS} + \beta_6\text{LEP} + \beta_7\text{Section 504} + \beta_8\text{i-Ready*Gender} + \beta_9\text{i-Ready*Race} + \beta_{10}\text{i-Ready*SPED} + \beta_{11}\text{i-Ready*FARMS} + \beta_{12}\text{i-Ready*LEP} + \beta_{13}\text{i-Ready*Section 504} + e_i.$$

Next, the same model was used to examine the predictive validity of the winter i-Ready Diagnostic to the LEAP 2025 scaled scores. Lastly, the same model was used to examine the concurrent validity of the spring i-Ready Diagnostic to the LEAP 2025 scaled scores. The demographic variables special education (SPED), free and reduced meals (FARMS), limited English proficiency (LEP), and Section 504 were categorical; Students were coded as 1 if they were identified with the label and 0 if not. Gender was reported as male (i.e., coded as 1) or female (i.e., coded as 0). The variable race was coded as white (i.e., coded as 1) or non-white (i.e., coded as 0). This decision was made based on the nature of the data; there were limited samples of data for a variety of races (e.g., Asian, Pacific Islander).

Results

Descriptive statistics for the experimental and criterion measures are displayed in Tables 6-7. Mean scores on the i-Ready Diagnostic show an increase from fall to winter for all grades except eighth grade. The same holds true when comparing mean scores on the i-Ready Diagnostic from winter to spring. Scaled scores on the LEAP 2025 show students in third grade had the highest level of performance; student performance worsened as grades increased.

Table 6

Means and Standard Deviations for Experimental Measures

		i-Ready Diagnostic		
Grade	*n*	Fall	Winter	Spring
Third	121-143	431.98 (24.47)	444.34 (25.48)	446.28 (35.56)
Fourth	124-141	450.91 (23.85)	458.63 (27.28)	475.59 (26.55)
Fifth	164-192	465.95 (27.89)	470.92 (31.29)	480.29 (31.79)
Sixth	157-180	479.93 (28.58)	482.79 (32.38)	489.20 (34.28)
Seventh	187-285	482.53 (32.53)	485.15 (31.61)	481.69 (42.43)
Eighth	188-254	496.60 (28.36)	495.33 (28.59)	502.20 (32.17)

Table 7

Means and Standard Deviations for Criterion Measure

		LEAP 2025					
Grade	n	SS	Achieve[a]	Content[b]	Reasoning[b]	Modeling[b]	SC[b]
Third	143	752.56 (32.80)	2.38 (1.06)	1.33 (0.85)	1.27 (0.86)	1.31 (0.87)	1.27 (0.84)
Fourth	141	746.11 (29.64)	2.23 (0.98)	1.02 (0.81)	1.18 (0.86)	1.24 (0.84)	1.28 (0.86)
Fifth	192	740.78 (27.93)	2.06 (1.00)	0.95 (0.82)	0.87 (0.90)	1.03 (0.86)	1.31 (0.82)
Sixth	180	740.57 (28.67)	2.04 (1.00)	1.06 (0.87)	0.99 (0.84)	0.99 (0.86)	1.29 (0.85)
Seventh	285	732.28 (24.90)	1.79 (0.95)	0.78 (0.80)	0.80 (0.83)	0.84 (0.79)	0.94 (0.88)
Eighth	254	737.75 (31.35)	1.89 (1.05)	1.02 (0.82)	0.95 (0.91)	0.99 (0.89)	0.96 (0.83)

Note. SS = Scaled Score, Achieve = Mathematics Achievement, Content = Major Content, Reasoning = Expressing Mathematical Reasoning, Modeling = Modeling & Application, SC = Additional & Supporting Content.

[a] The LEAP 2025 Mathematical Achievement score range was 0-4

[b] The LEAP 2025 subdomains score range was 0-2

Concurrent Validity of i-Ready Diagnostic

To examine the concurrent validity of the i-Ready Diagnostic, validity coefficients were calculated between the scores on the i-Ready Diagnostic administered in the spring and scores on the LEAP 2025. When analyzing the concurrent validity to the LEAP scaled scores, coefficients ranged from 0.70 (i.e., seventh grade) to 0.84 (i.e., fourth grade) across grades. When considering the concurrent validity of the i-Ready

Diagnostic to the LEAP 2025 domains, coefficients were the largest for the major content domain across grades with a range of 0.63 (i.e., eighth grade) to 0.74 (i.e., fifth grade). Validity coefficients for additional and supporting content (SC) were the smallest for all but eighth grade with a range of 0.43 (sixth grade) to 0.56 (fourth and fifth grades). No discernable pattern across grades was identified when comparing validity coefficients for the domains expressing mathematical reasoning and modeling and application. Table 8 contains complete list of concurrent validity coefficients

Table 8

Concurrent Validity for Spring i-Ready for Third through Eighth Grades

	Spring i-Ready					
	Third	Fourth	Fifth	Sixth	Seventh	Eighth
LEAP 2025						
Scaled Score[a]	0.73*	0.84*	0.82*	0.82*	0.70*	0.75*
Achieve[b]	0.71*	0.79*	0.78*	0.81*	0.64*	0.69*
Content[b]	0.64*	0.71*	0.74*	0.74*	0.65*	0.63*
Reasoning[b]	0.56*	0.68*	0.58*	0.65*	0.55*	0.55*
Modeling[b]	0.58*	0.55*	0.70*	0.64*	0.58*	0.59*
SC[b]	0.50*	0.56*	0.56*	0.43*	0.55*	0.57*

Note. *$p < 0.05$, Achieve = Mathematics achievement, Content = Major content, Reasoning = Expressing mathematical reasoning, Modeling = Modeling & application, SC = Additional & supporting content.
[a]Correlation coefficients are Pearson *r*s.
[b]Correlation coefficients are Spearman ρs

Predictive Validity of i-Ready Diagnostic

To examine the predictive validity of the i-Ready Diagnostic, validity coefficients were calculated between scores obtained from the fall and winter administrations on the i-Ready Diagnostic and the LEAP 2025. Predictive validity coefficients for the fall administration to the scaled scores on the LEAP 2025 ranged from 0.73 (third grade) to 0.81 (seventh and eighth grades). Similarly, the predictive validity coefficients for the winter administration were in the same range, 0.74 (fourth and eighth grades) to 0.80 (sixth grade). Scores on the fall administration yielded larger predictive validity coefficients for seventh and eighth grades. Scores on the winter administration yielded larger predictive validity coefficients for third grade. The predictive validity coefficients on the fall and winter administrations of the i-Ready Diagnostic for fourth through sixth grades were similar.

When examining the predictive validity of the i-Ready Diagnostic to the LEAP 2025 domains, the strongest validity coefficients were found for major content across all grades and administration dates (except third grade fall i-Ready Diagnostic scores). Additionally, the smallest predictive validity coefficients were found for additional and supporting content across all but fourth grade and administration dates. No discernible pattern across grades or administration dates was evident when comparing predictive validity coefficients of expressing mathematical reasoning and modeling and application. Table 9 and 10 display the complete predictive validity coefficients for fall and winter i-Ready Diagnostic, respectively.

Table 9

Predictive Validity for Fall i-Ready for Third through Eighth Grades

	Fall i-Ready					
	Third	Fourth	Fifth	Sixth	Seventh	Eighth
LEAP 2025						
Scaled Score[a]	0.73*	0.74*	0.76*	0.78*	0.81*	0.81*
Achieve[b]	0.70*	0.70*	0.72*	0.76*	0.77*	0.77*
Content[b]	0.56*	0.62*	0.69*	0.74*	0.70*	0.70*
Reasoning[b]	0.60*	0.55*	0.52*	0.61*	0.66*	0.65*
Modeling[b]	0.59*	0.53*	0.63*	0.64*	0.68*	0.65*
SC[b]	0.55*	0.57*	0.46*	0.32*	0.55*	0.66*

Note. *$p < 0.05$, Achieve = Mathematics achievement, Content = Major content, Reasoning = Expressing mathematical reasoning, Modeling = Modeling & application, SC = Additional & supporting content.
[a]Correlation coefficients are Pearson *r*s.
[b]Correlation coefficients are Spearman ρs

Table 10

Predictive Validity for Winter i-Ready for Third through Eighth Grades

	Winter i-Ready					
	Third	Fourth	Fifth	Sixth	Seventh	Eighth
LEAP 2025						
Scaled Score[a]	0.78*	0.74*	0.75*	0.80*	0.75*	0.74*
Achieve[b]	0.72*	0.71*	0.72*	0.77*	0.69*	0.69*
Content[b]	0.63*	0.59*	0.66*	0.75*	0.69*	0.62*
Reasoning[b]	0.56*	0.59*	0.51*	0.65*	0.58*	0.57*
Modeling[b]	0.65*	0.53*	0.63*	0.70*	0.60*	0.61*
SC[b]	0.53*	0.54*	0.50*	0.37*	0.53*	0.55*

Note. *$p < 0.05$, Achieve = Mathematics achievement, Content = Major content, Reasoning = Expressing mathematical reasoning, Modeling = Modeling & application, SC = Additional & supporting content.
[a]Correlation coefficients are Pearson *r*s.
[b]Correlation coefficients are Spearman ρs

Moderator Analyses of the Concurrent Validity of the i-Ready Diagnostic

Multiple linear regression was used to identify student demographic variables (i.e., gender, race, SPED, FARMS, LEP, and Section 504) that may moderate the concurrent validity of the spring i-Ready Diagnostic to the LEAP 2025. Full results are reported in Table 11 and Appendix E contains graphical displays of statistically significant moderators.

Third grade. For third grade, the model predicted a statistically significant amount of variance, F (13, 106) = 11.11, $p < 0.01$, with an adjusted R^2 of 0.525. This can be interpretted as the model explained 52.5% of the variance in LEAP 2025 scaled scores. When analyzing the main effects, the β weight for the i-Ready Diagnostic was 0.89, $p < 0.01$ which can be interpreted as one standardized unit of change on the i-Ready Diagnostic being 0.89 standardized unit of change on the LEAP 2025 scaled scores. To examine whether the concurrent validity was moderated by demographic variables, interaction effects were analyzed between the i-Ready Diagnostic and each demographic variable. None of the interaction terms were statistically significant predictors.

Fourth grade. For fourth grade, the model predicted a statistically significant amount of variance, F (13, 110) = 25.80, $p < 0.01$, with an adjusted R^2 of 0.724. This can be interpreted as the model explained 72.4% of the variance in LEAP 2025 scaled scores. When analyzing the main effects, the β weight for the i-Ready Diagnostic was 0.97, $p < 0.01$ which can be interpreted as one standardized unit of change on the i-Ready Diagnostic being 0.97 standardized unit of change on the LEAP 2025 scaled scores. The interaction term i-Ready Diagnostic*SPED was a statistically significant predictor with a $\beta = -2.57$, $p < 0.01$, which suggests the concurrent validity was moderated based on disability.

Fifth grade. For fifth grade, the model predicted a statistically significant amount of variance, $F(13, 150) = 36.59$, $p < 0.01$, with an adjusted R^2 of 0.739. This can be interpreted as the model explained 73.9% of the variance in LEAP 2025 scaled scores. When analyzing the main effects, the β weight for the i-Ready Diagnostic was 0.98, $p < 0.01$ which can be interpreted as one standardized unit of change on the i-Ready Diagnostic being 0.98 standardized unit of change on the LEAP 2025 scaled scores. Both interaction terms, i-Ready Diagnostic*SPED and i-Ready Diagnostic*Section 504, were statistically significant predictors: i-Ready Diagnostic*SPED $\beta = -3.05$, $p < 0.01$ and for i-Ready Diagnostic*Section 504, $\beta = -1.80$, $p < 0.01$.

Sixth grade. For sixth grade, the model predicted a statistically significant amount of variance, $F(13, 143) = 29.41$, $p < 0.01$, with an adjusted R^2 of 0.703. This can be interpreted as the model explained 70.3% of the variance in LEAP 2025 scaled scores. When analyzing the main effects, the β weight for the i-Ready Diagnostic was 1.20, $p < 0.01$ which can be interpreted as one standardized unit of change on the i-Ready Diagnostic being 1.17 standardized unit of change on the LEAP 2025 scaled scores. The interaction terms, i-Ready Diagnostic*Gender, i-Ready Diagnostic*SPED, and i-Ready Diagnostic*FARMS, were statistically significant predictor: i-Ready

Diagnostic*Gender β = -2.28, $p < 0.01$, i-Ready Diagnostic*SPED β = -1.84, $p = 0.01$, and i-Ready Diagnostic*FARMS, β = -1.51, $p = 0.03$.

Seventh grade. For seventh grade, the model predicted a statistically significant amount of variance, F (13, 173) = 19.84, $p < 0.01$, with an adjusted R^2 of 0.568. This can be interpreted as the model explained 56.8% of the variance in LEAP 2025 scaled scores. When analyzing the main effects, the β weight for the i-Ready Diagnostic was 0.57, $p < 0.01$ which can be interpreted as one standardized unit of change on the i-Ready Diagnostic being 0.57 standardized unit of change on the LEAP 2025 scaled scores. Both interaction terms, i-Ready Diagnostic*FARMS and i-Ready Diagnostic*LEP, were statistically significant predictor: i-Ready Diagnostic*FARMS β = -1.26, $p = 0.04$ and for i-Ready Diagnostic*LEP, β = 3.70, $p < 0.01$.

Eighth grade. For eighth grade, the model predicted a statistically significant amount of variance, F (13, 174) = 23.85, $p < 0.01$, with an adjusted R^2 of 0.614. This can be interpreted as the model explained 61.4% of the variance in LEAP 2025 scaled scores. When analyzing the main effects, the β weight for the i-Ready Diagnostic was 1.04, $p < 0.01$ which can be interpreted as one standardized unit of change on the i-Ready Diagnostic being 1.04 standardized unit of change on the LEAP 2025 scaled scores. The interaction terms, i-Ready Diagnostic*FARMS, was a statistically significant predictor with β = -2.86, $p < 0.01$.

Table 11

Moderator Analyses of Concurrent Validity for Spring-CBM

Grade	R^2	F	Predictors	β (p)	r_s
Third	0.525	11.11**	i-Ready	0.89 (<0.01)**	1.38
			Gender	1.32 (0.17)	-0.13
			Race	0.31 (0.73)	0.37
			SPED	0.84 (0.33)	-0.39
			FARMS	-1.15 (0.31)	-0.49
			LEP	0.72 (0.49)	-0.19
			Section 504	1.24 (0.16)	-0.13
			i-Ready*Gender	-1.33 (0.16)	
			i-Ready*Race	-0.24 (0.80)	
			i-Ready*SPED	-0.89 (0.29)	
			i-Ready*FARMS	1.06 (0.34)	
			i-Ready*LEP	-0.67 (0.52)	
			i-Ready*Section 504	-1.15 (0.19)	
Fourth	0.724	25.80**	i-Ready	0.97 (<0.01)**	1.17
			Gender	0.80 (0.38)	-0.06
			Race	-0.11 (0.90)	0.16
			SPED	2.59 (<0.01)**	-0.30
			FARMS	0.89 (0.53)	-0.10
			LEP	-0.22 (0.88)	0.07
			Section 504	2.40 (0.07)	-0.17
			i-Ready*Gender	-0.81 (0.37)	
			i-Ready*Race	0.17 (0.85)	
			i-Ready*SPED	-2.57 (<0.01)**	
			i-Ready*FARMS	-0.84 (0.55)	
			i-Ready*LEP	0.17 (0.91)	
			i-Ready*Section 504	-2.41 (0.07)	
Fifth	0.739	36.59**	i-Ready	0.98 (<0.01)**	1.11
			Gender	-0.46 (0.51)	-0.14
			Race	-0.12 (0.86)	0.25
			SPED	3.17 (<0.01)**	-0.40
			FARMS	0.99 (0.16)	-0.42
			LEP	-2.16 (0.18)	0.09
			Section 504	1.80 (0.01)*	-0.29
			i-Ready*Gender	0.47 (0.50)	
			i-Ready*Race	0.15 (0.82)	
			i-Ready*SPED	-3.05 (<0.01)**	
			i-Ready*FARMS	-1.05 (0.13)	
			i-Ready*LEP	2.18 (0.18)	
			i-Ready*Section 504	-1.80 (<0.01)**	

Table 11 *Continued*

Sixth	0.703	29.41**	i-Ready	1.20 (<0.01)**	1.16
			Gender	2.51 (<0.01)**	-0.08
			Race	0.56 (0.40)	0.09
			SPED	1.89 (<0.01)**	-0.56
			FARMS	1.53 (0.03)*	-0.35
			LEP	0.11 (0.95)	-0.12
			Section 504	-0.31 (0.70)	-0.32
			i-Ready*Gender	-2.51 (<0.01)**	
			i-Ready*Race	-0.54 (0.42)	
			i-Ready*SPED	-1.84 (0.01)*	
			i-Ready*FARMS	-1.51 (0.03)*	
			i-Ready*LEP	-0.16 (0.93)	
			i-Ready*Section 504	0.22 (0.78)	
Seventh	0.568	19.84**	i-Ready	0.57 (<0.01)**	1.22
			Gender	-0.85 (0.20)	-0.06
			Race	-0.92 (0.13)	0.04
			SPED	-0.27 (0.70)	-0.68
			FARMS	1.12 (0.08)	-0.50
			LEP	-3.79 (<0.01)**	-0.15
			Section 504	0.38 (0.39)	-0.07
			i-Ready*Gender	0.87 (0.19)	
			i-Ready*Race	0.94 (0.12)	
			i-Ready*SPED	0.10 (0.89)	
			i-Ready*FARMS	-1.26 (0.04)*	
			i-Ready*LEP	3.70 (0.01)*	
			i-Ready*Section 504	-0.40 (0.36)	
Eighth	0.613	23.85**	i-Ready	1.04 (<0.01)**	1.23
			Gender	0.60 (0.47)	-0.07
			Race	0.87 (0.26)	-0.10
			SPED	1.02 (0.17)	-0.51
			FARMS	2.82 (<0.01)*	-0.37
			LEP	0.05 (0.88)	-0.16
			Section 504	0.79 (0.56)	-0.18
			i-Ready*Gender	-0. 59 (0.47)	
			i-Ready*Race	-0.88 (0.25)	
			i-Ready*SPED	-1.11 (0.13)	
			i-Ready*FARMS	-2.86 (<0.01)**	
			i-Ready*LEP	-0.07 (0.84)	
			i-Ready*Section 504	-0.86 (0.53)	

Note. The reported R^2 was adjusted. FARMS = Free and reduced meals, LEP = limited English proficiency, r_s = structure coefficients
*p<0.05, **p<0.01

Moderator Analyses of the Predictive Validity of the i-Ready Diagnostic

To examine if the predictive validity of the i-Ready Diagnostic to the LEAP 2025 was moderated by demographic variables, multiple linear regression was used. Two separate regressions were run to examine the predictive validity of the fall and winter administration of the i-Ready Diagnostic. Full results are reported for the fall (i.e., Table 12) and winter (i.e., Table 13) administration. Appendix E contains graphical disaplys of statistically significant moderators.

Third grade. For third grade, the model predicted a statistically significant amount of variance in the fall (F [13, 121] = 13.91, $p < 0.01$, adjusted $R^2 = 0.556$) and winter (F [13, 120] = 17.29, $p < 0.01$, adjusted $R^2 = 0.614$). When examining the main effects, the β weight for the fall i-Ready Diagnostic was 0.76, $p < 0.01$ and 1.04, $p < 0.01$ for the winter. No interaction terms were significant for the winter administration; the interaction term i-Ready Diagnostic*FARMS was a statistically significant predictor for the fall administration with a β = -2.49, $p = 0.05$.

Fourth grade. For fourth grade, the model predicted a statistically significant amount of variance in the fall (F [13, 117] = 14.17, $p < 0.01$, adjusted $R^2 = 0.568$) and winter (F [13, 117] = 12.63, $p < 0.01$, adjusted $R^2 = 0.638$). When examining the main effects, the β weight for the fall i-Ready Diagnostic was 0.95, $p < 0.01$ and 1.01, $p < 0.01$ for the winter. None of the interaction terms were statistically significant predictors for the fall administration; the interaction term i-Ready Diagnostic*Gender was a statistically significant predictor for the winter administration with a β = -2.47, $p = 0.03$.

Fifth grade. For fifth grade, the model predicted a statistically significant amount of variance in the fall (F [13, 172] = 26.78, $p < 0.01$, adjusted $R^2 = 0.644$) and winter (F [13, 171] = 26.20, $p < 0.01$, adjusted $R^2 = 0.640$). When examining the main effects, the β weight for the fall i-Ready Diagnostic was 0.80, $p < 0.01$ and 0.89, $p < 0.01$ for the winter. Both, i-Ready Diagnostic*SPED (β = -2.42, $p < 0.01$) and i-Ready Diagnostic*LEP (β = -11.01, $p = 0.02$), were statistically significant predictors for the fall administration. The following interaction terms were statistically significant predictors for the winter administration: i-Ready Diagnostic*SPED (β = -2.11, $p < 0.01$), i-Ready Diagnostic*FARMS (β = -1.78, $p = 0.02$), i-Ready Diagnostic*Section 504 (β = -1.69, $p = 0.02$).

Sixth grade. For sixth grade, the model predicted a statistically significant amount of variance in the fall (F [13, 163] = 26.35, $p < 0.01$, adjusted $R^2 = 0.652$) and winter (F [13, 161] = 27.77, $p < 0.01$, adjusted $R^2 = 0.667$). When examining the main effects, the β weight for the fall i-Ready Diagnostic was 1.17, $p < 0.01$ and 1.00, $p < 0.01$ for the winter. Both, i-Ready Diagnostic*Gender (β = -2.28, $p < 0.01$) and i-Ready Diagnostic*FARMS (β = -3.27, $p < 0.01$), were statistically significant predictors for the fall administration. The interaction term i-Ready Diagnostic*Gender (β = -1.94, $p < 0.01$) was a statistically significant predictor for the winter administration.

Seventh grade. For seventh grade, the model predicted a statistically significant amount of variance in the fall (F [13, 265] = 44.40, $p < 0.01$, adjusted $R^2 = 0.670$) and winter (F [13, 252] = 29.27, $p < 0.01$, adjusted $R^2 = 0.581$). When examining the main effects, the β weight for the fall i-Ready Diagnostic was 0.99, $p < 0.01$ and 0.83, $p <$

0.01 for the winter. Both, i-Ready Diagnostic*FARMS (Fall: $\beta = -1.32$, $p = 0.02$; winter: $\beta = -1.42$, $p = 0.03$) and i-Ready Diagnostic*LEP (Fall: $\beta = 1.36$, $p < 0.01$; winter: $\beta = 4.83$, $p < 0.01$), were statistically significant predictors for the fall and winter administrations.

Eighth grade. For eighth grade, the model predicted a statistically significant amount of variance in the fall (F [13, 236] = 41.31, $p < 0.01$, adjusted $R^2 = 0.678$) and winter (F [13, 210] = 24.64, $p < 0.01$, adjusted $R^2 = 0.580$). When examining the main effects, the β weight for the fall i-Ready Diagnostic was 0.79, $p < 0.01$ and 0.83, $p < 0.01$ for the winter. Both, i-Ready Diagnostic*Gender ($\beta = 1.51$, $p = 0.02$) and i-Ready Diagnostic*SPED ($\beta = -2.10$, $p < 0.01$) were statistically significant predictors for the fall administration. For the winter administration the interaction term i-Ready Diagnostic*Section 504 was a statistically significant predictor with $\beta = -1.91$, $p = 0.03$.

Table 12

Moderator Analyses of Predictive Validity for Fall-CBM

Grade	R^2	F	Predictors	β (p)	r_s
Third	0.556	13.91**	i-Ready	0.76 (<0.01)**	1.31
			Gender	-0.23 (0.83)	-0.13
			Race	-0.23 (0.84)	0.35
			SPED	0.78 (0.38)	-0.37
			FARMS	2.35 (0.07)	-0.46
			LEP	-5.38 (0.06)	-0.18
			Section 504	0.46 (0.74)	-0.13
			i-Ready*Gender	0.23 (0.83)	
			i-Ready*Race	0.28 (0.80)	
			i-Ready*SPED	-0.71 (0.41)	
			i-Ready*FARMS	-2.49 (0.05)*	
			i-Ready*LEP	5.36 (0.06)	
			i-Ready*Section 504	-0.47 (0.74)	
Fourth	0.568	14.17**	i-Ready	0.95 (<0.01)**	1.33
			Gender	0.92 (0.49)	-0.08
			Race	1.14 (0.35)	0.21
			SPED	2.01 (0.15)	-0.38
			FARMS	2.80 (0.07)	-0.13
			LEP	-1.61 (0.36)	0.09
			Section 504	1.53 (0.35)	-0.21
			i-Ready*Gender	-0.95 (0.47)	
			i-Ready*Race	-1.09 (0.38)	
			i-Ready*SPED	-2.04 (0.15)	
			i-Ready*FARMS	-2.79 (0.07)	
			i-Ready*LEP	1.69 (0.33)	
			i-Ready*Section 504	-1.53 (0.35)	
Fifth	0.644	26.78**	i-Ready	0.80 (<0.01)**	1.20
			Gender	-1.10 (0.19)	-0.16
			Race	-0.04 (0.96)	0.29
			SPED	2.48 (<0.01)**	-0.46
			FARMS	0.76 (0.35)	-0.49
			LEP	11.08 (0.02)*	0.11
			Section 504	1.33 (0.13)	-0.34
			i-Ready*Gender	1.09 (0.19)	
			i-Ready*Race	0.14 (0.86)	
			i-Ready*SPED	-2.42 (<0.01)**	
			i-Ready*FARMS	-0.89 (0.27)	
			i-Ready*LEP	-11.01 (0.02)*	
			i-Ready*Section 504	-1.38 (0.11)	

Table 12 *Continued*

Sixth	0.652	26.35**	i-Ready	1.17 (<0.01)**	1.19
			Gender	2.27 (<0.01)**	-0.09
			Race	0.37 (0.65)	0.10
			SPED	1.41 (0.21)	-0.60
			FARMS	3.31 (<0.01)	-0.38
			LEP	-0.03 (0.97)	-0.13
			Section 504	0.58 (0.57)	-0.35
			i-Ready*Gender	-2.28 (<0.01)**	
			i-Ready*Race	-0.29 (0.72)	
			i-Ready*SPED	-1.38 (0.21)	
			i-Ready*FARMS	-3.27 (<0.01)**	
			i-Ready*LEP	-0.01 (0.99)	
			i-Ready*Section 504	-0.68 (0.50)	
Seventh	0.670	44.40**	i-Ready	0.99 (<0.01)**	1.21
			Gender	0.73 (0.20)	-0.05
			Race	0.33 (0.54)	0.03
			SPED	1.01 (0.06)	-0.58
			FARMS	1.28 (0.02)*	-0.43
			LEP	-1.35 (<0.01)**	-0.13
			Section 504	-0.11 (0.82)	-0.06
			i-Ready*Gender	-0.73 (0.20)	
			i-Ready*Race	-0.31 (0.58)	
			i-Ready*SPED	-1.00 (0.06)	
			i-Ready*FARMS	-1.32 (0.02)*	
			i-Ready*LEP	1.36 (<0.01)**	
			i-Ready*Section 504	0.09 (0.86)	
Eighth	0.678	41.31**	i-Ready	0.79 (<0.01)**	1.16
			Gender	-1.51 (0.02)*	-0.07
			Race	0.79 (0.23)	-0.09
			SPED	2.06 (<0.01)**	-0.46
			FARMS	-0.01 (0.99)	-0.33
			LEP	0.46 (0.24)	-0.14
			Section 504	1.73 (0.09)	-0.16
			i-Ready*Gender	1.51 (0.02)*	
			i-Ready*Race	-0.81 (0.22)	
			i-Ready*SPED	-2.10 (<0.01)**	
			i-Ready*FARMS	-0.02 (0.98)	
			i-Ready*LEP	-0.45 (0.24)	
			i-Ready*Section 504	-1.81 (0.08)	

Note. The reported R^2 was adjusted. FARMS = Free and reduced meals, LEP = limited English proficiency, r_s = structure coefficients
*p<0.05, **p<0.01

Table 13

Moderator Analyses of Predictive Validity for Winter-CBM

Grade	R^2	F	Predictors	β (p)	r_s
Third	0.614	17.29**	i-Ready	1.04 (<0.01)**	1.26
			Gender	1.50 (0.13)	-0.11
			Race	0.73 (0.49)	0.32
			SPED	1.12 (0.15)	-0.33
			FARMS	2.17 (0.07)	-0.42
			LEP	-0.14 (0.89)	-0.16
			Section 504	3.49 (0.12)	-0.11
			i-Ready*Gender	-1.50 (0.13)	
			i-Ready*Race	-0.74 (0.49)	
			i-Ready*SPED	-1.06 (0.16)	
			i-Ready*FARMS	-2.23 (0.06)	
			i-Ready*LEP	0.18 (0.85)	
			i-Ready*Section 504	-3.46 (0.12)	
Fourth	0.538	12.63**	i-Ready	1.01 (<0.01)**	1.37
			Gender	2.49 (0.03)*	-0.08
			Race	0.99 (0.36)	0.22
			SPED	1.31 (0.35)	-0.40
			FARMS	1.54 (0.34)	-0.14
			LEP	-0.93 (0.59)	0.09
			Section 504	-0.07 (0.97)	-0.23
			i-Ready*Gender	-2.47 (0.03)*	
			i-Ready*Race	-0.92 (0.40)	
			i-Ready*SPED	-1.38 (0.33)	
			i-Ready*FARMS	-1.54 (0.34)	
			i-Ready*LEP	1.00 (0.57)	
			i-Ready*Section 504	0.09 (0.97)	
Fifth	0.640	26.20**	i-Ready	0.89 (<0.01)**	1.19
			Gender	-0.69 (0.37)	-0.16
			Race	-0.49 (0.48)	0.29
			SPED	2.18 (<0.01)**	-0.47
			FARMS	1.68 (0.03)*	-0.49
			LEP	-7.58 (0.08)	0.11
			Section 504	1.70 (0.03)*	-0.34
			i-Ready*Gender	0.68 (0.38)	
			i-Ready*Race	0.60 (0.40)	
			i-Ready*SPED	-2.11 (<0.01)**	
			i-Ready*FARMS	-1.78 (0.02)*	
			i-Ready*LEP	7.64 (0.08)	
			i-Ready*Section 504	-1.69 (0.02)*	

Table 13 *Continued*

Sixth	0.667	27.77**	i-Ready	1.00 (<0.01)**	1.19
			Gender	1.94 (<0.01)**	-0.09
			Race	-0.28 (0.68)	0.09
			SPED	0.64 (0.45)	-0.59
			FARMS	1.29 (0.07)	-0.37
			LEP	-2.27 (0.11)	-0.12
			Section 504	0.56 (0.50)	-0.34
			i-Ready*Gender	-1.94 (<0.01)**	
			i-Ready*Race	0.33 (0.63)	
			i-Ready*SPED	-0.65 (0.45)	
			i-Ready*FARMS	-1.29 (0.07)	
			i-Ready*LEP	2.20 (0.13)	
			i-Ready*Section 504	-0.70 (0.40)	
Seventh	0.581	49.11**	i-Ready	0.83 (<0.01)**	1.29
			Gender	0.41 (0.53)	-0.06
			Race	0.13 (0.84)	0.04
			SPED	0.24 (0.74)	-0.66
			FARMS	1.36 (0.05)*	-0.49
			LEP	-4.86 (<0.01)**	-0.15
			Section 504	0.21 (0.68)	-0.07
			i-Ready*Gender	-0.40 (0.55)	
			i-Ready*Race	-0.10 (0.88)	
			i-Ready*SPED	-0.35 (0.62)	
			i-Ready*FARMS	-1.42 (0.03)*	
			i-Ready*LEP	4.83 (<0.01)**	
			i-Ready*Section 504	-0.24 (0.64)	
Eighth	0.579	24.64**	i-Ready	0.83 (<0.01)**	1.26
			Gender	-0.29 (0.73)	-0.08
			Race	1.07 (0.19)	-0.11
			SPED	0.24 (0.74)	-0.54
			FARMS	1.22 (0.14)	-0.39
			LEP	0.27 (0.50)	-0.17
			Section 504	1.84 (0.03)*	-0.19
			i-Ready*Gender	0.36 (0.66)	
			i-Ready*Race	-1.06 (0.19)	
			i-Ready*SPED	-0.90 (0.32)	
			i-Ready*FARMS	-1.31 (0.11)	
			i-Ready*LEP	-0.31 (0.43)	
			i-Ready*Section 504	-1.91 (0.03)*	

Note. The reported R^2 was adjusted. FARMS = Free and reduced meals, LEP = limited English proficiency, r_s = structure coefficients
*p<0.05, **p<0.01

Discussion

The current study extended previous literature on the criterion validity of m-CBMs by including data for middle school students and the analyses of student demographic variables that may moderate the validity of an m-CBM to the end-of-year state assessment. Overall, criterion validity coefficients for third through eighth grades were in the same range as those reported by the Educational Research Institute of America who compared the criterion validity of the i-Ready Diagnostic to a criterion measure aligned to the Common Core State Standards for Mathematics (e.g., NY State Assessment, and the developers of the i-Ready Diagnostic who reported criterion validity coefficients to two criterion measures aligned to the Common Core State Standards for Mathematics (e.g., PARCC, Smarter Balanced). In addition, all concurrent and predictive validity coefficients were above the threshold (i.e., 0.70) suggested by the National Center for Response to Intervention (Gersten et al., 2009; Nunnally, 1978).

Another finding for the overall effects were the lower criterion validity coefficients related to subdomain additional and supporting content. For a majority of grades the validity coefficients for additional and supporting content were well below the other LEAP domains. Additional measures may need to be used that have stronger criterion validity to items reflected additional and supporting content.

Lastly, the mismatch between administration format (i.e., computer versus paper/pencil) of the i-Ready Diagnostic and end-of-year assessment did not appear to affect overall validity coefficients. Third- and fourth-graders were administered a paper/pencil end-of-year criterion assessment and validity coefficients were in the same range as fifth- through eighth-graders who completed a computer administered end-of-year assessment.

A primary finding was the concurrent validity of the i-Ready Diagnostic was moderated by the following demographic variables: gender (4th, 6th), SPED (4th, 5th), Section 504 (5th), FARMS (6th, 7th, 8th), and LEP (7th). Furthermore, the predictive validity of the i-Ready Diagnostic was moderated by the following demographic variables: gender (6th, 8th), SPED (5th, 8th), FARMS (3rd, 6th, 7th), and LEP (5th, 7th). However, it is worth noting that the moderation found for students identified as LEP was based off a limited sampling of data due to the demographics of the school district; this result should be examined in future research before drawing definitive conclusions. Students identified with disabilities (i.e., receiving services under IDEA or Section 504), families who have low income, and students with limited English proficiency are all subgroups that local education agencies must report for accountability purposes under the ESSA (2015). In order for schools to meet the needs of these students, it is imperative universal screeners are providing reliable and valid data related to the students' mathematical ability and not capturing irrelevant variance related to their demographic information. Variance related to their demographic information is problematic because this is not malleable, whereas mathematical performance can be improved. The moderation of the predictive validity is concerning because students matching these demographic variables have a higher likelihood of not reaching end-of-year mathematics expectations (Office of Special Education and Rehabilitation Services, U.S. Department of Education, 2016; U.S. Department of Education, Institute of

Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress, 2016), and invalid data may lead to invalid decisions within an MTSS framework. Lastly, the end-of-year mathematics assessment required student to use interactive tools that required a minimum level of expertise, which the i-Ready Diagnostic did not capture.

Limitations

Several limitations should be considered when interpreting the findings of the study. First, calculating the intra-class correlation and/or using multi-level modeling were not feasible given the missing classroom identifier per student. Ignoring the nestedness of the data can lead to inflated effect sizes (Burstein, 1980; Cronbach, 1976; Luo & Kwok, 2009); however, given the limited number of clusters per grade level in the analyses makes the likelihood less severe. Second, missing data across administrations of the m-CBM or the end-of-year assessment may be problematic. Lastly, fidelity of administration of the m-CBM and the end-of-year state assessment was not collected. Failing to adhere to the standardized protocol could affect score validity.

Implications for Future Research

Future research should consider the impact of demographic variables as moderators to the validity of m-CBMs. Replication in social science research has received increased attention and the field has addressed the need to increase replication studies (Makel & Plucker, 2014; Makel, Plucker, Freeman, Lombardi, Simonsen, & Coyne, 2016) to examine the generalization of findings. Furthermore, the increased use of computer administered assessments raises validity concerns. Empirical research identifying whether the mismatch between computer and paper/pencil administration of mathematics assessments moderates validity is needed. Lastly, the provision of accommodations may affect the validity of assessments; future research should investigate the administration of m-CBMs and criterion measures under similar and dissimilar accommodation environments.

DISCUSSION AND CONCLUSIONS

This dissertation aimed to synthesize evidence of criterion validity related to m-CBMs, to evaluate the criterion validity of a computer adaptive m-CBM to an end-of-year state assessment, and to examine the effect of demographic variables on the criterion validity of the m-CBM. To address these aims, the current dissertation included a systematic review of the literature and a correlational research study. Results from the systematic review of the literature indicated that the criterion validity of m-CBMs varied on a variety of factors. Measures focused on concepts and applications demonstrate the strongest concurrent and predictive validity to criterion measures focused on overall mathematics achievement for students in upper elementary and middle school. A majority of the research with early elementary has focused on numeracy measures; concurrent and predictive validity for these measures were moderate to strong to criterion measures of mathematics achievement. Results from the correlational study suggested that scores on the computer adaptive m-CBM had strong concurrent and predictive validity to the end-of-year mathematics assessment for third through eighth grades. The analysis of the domains of the assessment revealed that the concurrent and predictive validity coefficients for the m-CBM were the strongest for the major contain domain and the weakest for the additional and supporting content domain. Lastly, multiple linear regression analyses indicated that the following demographic variables: FARMS, SPED, Gender, Section 504, and LEP (although this was likely due to limited sample size), moderated the criterion validity of scores. Results from these two studies indicate that m-CBMs focused on concepts and applications will yield scores that have strong criterion validity to measures of overall mathematics achievement; however, the need for additional research is also warranted.

Despite the promising findings in regards to criterion validity for m-CBMs, additional research questions should be addressed. First, the administration protocol of m-CBMs differs across studies, which raises questions as to how these may affect score reliability and validity. Secondly, analyzing how administering m-CBMs and standardized assessment of mathematical achievement via computer platforms affects score validity is warranted. Lastly, additional research analyzing how student demographics affect score validity is needed.

Given the current body of literature, practitioners are recommended to select m-CBMs with careful consideration. Measures focused on concepts and applications yield stronger criterion validity to end-of-year expectations; thus, interpreting scores obtained from these measures should be used for similar purposes. Measures focused on basic facts and computations provide additional information regarding students' performance in more discrete areas of mathematics (e.g., fluency and arithmetic); thus, these measures can serve a supplementary role to provide additional information to inform instructional planning and data-based decision-making. Lastly, practitioners should ensure they are adequately prepared to administer and score m-CBMs following a standardized protocol to increase the likelihood of obtaining valid scores.

REFERENCES

Allinder, R. M., Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1992). Effects of summer break on math and spelling performance as a function of grade level. *The Elementary School Journal*, *92*, 451-460.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational and Psychological Research Association.

Baglici, S. P., Codding, R., & Tryon, G. (2010). Extending the research on the tests of early numeracy: Longitudinal analyses over two school years. *Assessment for Effective Intervention*, *35*, 89-102.

Ball, C. R., & Christ, T. J. (2012). Supporting valid decision making: Uses and misuses of assessment data within the context of RTI. *Psychology in the Schools*, *49*, 231-244.

Benz, M. R., Lindstrom, L., & Yovanoff, P. (2000). Improving graduation and employment outcomes of students with disabilities: Predictive factors and student perspectives. *Exceptional Children*, *66*, 509-529.

Berch, D. B., & Mazzocco, M. M. (2007). *Why is math so hard for some children? The nature and origins of mathematical learning difficulties and disabilities*. Baltimore, MD: Brookes Publishing Company.

Betts, J., Pickart, M., & Heistad, D. (2009). Construct and predictive validity evidence for curriculum-based measures of early literacy and numeracy skills in kindergarten. *Journal of Psychoeducational Assessment*, *27*, 83-95.

Blankenship, C. S. (1985). Using curriculum-based assessment data to make instructional decisions. *Exceptional Children*, *52*, 233-238.

Brualdi, A. (2002). Traditional and modern concepts of validity. In L. Rudner & W. Schafer (Eds.), *What teachers need to know about assessment*. Washington, DC: National Education Association.

Burstein, L. (1980). Chapter 4: The analysis of multilevel data in educational research and evaluation. *Review of Research in Education*, *8*, 158-233.

Bynner, J., & Parsons, S. (1997). *Does numeracy matter? Evidence from the national child development study on the impact of poor numeracy on adult life*. London: The Basic Skills Agency.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*, 270-295.

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*, 3-14.

Chard, D. J., Harn, B. A., Sugai, G., Horner, R. H., Simmons, D. C., & Kame'enui, E. J. (2008). Core features of multi-tiered systems of reading and behavioral support. In Greenwood, C. R., Kratochwill, T., & Clements, M. (Eds.), *School-*

wide prevention models: Lessons learned in elementary schools (pp. 31-58). New York, NY: Guilford Press.

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*, 198-205.

Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review*, *33*, 234-248.

Codding, R. S., Mercer, S., Connell, J., Fiorello, C., & Kleinert, W. (2016). Mapping the relationships among basic facts, concepts and application, and common core curriculum-based mathematics measures. *School Psychology Review, 45*, 19-38.

Codding, R. S., Petscher, Y., & Truckenmiller, A. (2015). CBM reading, mathematics, and written expression at the secondary level: Examining latent composite relations among indices and unique predictions with a state achievement test. *Journal of Educational Psychology*, *107*, 437-450.

Cooper, H. (2016). *Research synthesis and meta-analysis: A step-by-step approach* (Vol. 2). Thousand Oaks, CA: Sage Publications.

Council for Exceptional Children. (2009). *What every special educator must know: Ethics, standards, and guidelines* (6th ed.). Arlington, VA: Author.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*, 475-494.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design, and analysis* (Unpublished report). Stanford, CA: Stanford University, Stanford Evaluation Consortium.

Curriculum Associates LLC. (2015*). i-Ready and the Smarter Balanced assessments: Findings from independent research linking the i-Ready Diagnostic and Smarter Balanced assessments*. Retrieved at http://www.casamples.com/downloads/i-Ready-and-the-Smarter-Balanced-Assessments.pdf.

Curriculum Associates LLC. (2015). *i-Ready and PARCC: Findings from independent research linking the i-Ready Diagnostic and 2015 PARCC assessments*. Retrieved at http://www.casamples.com/downloads/i-Ready-and-PARCC.pdf.

Curriculum Associates LLC. (2016, April). *i-Ready diagnostic New York state validity study:* Report Number 518. Retrieved at http://www.casamples.com/downloads/i-Ready-Diagnostic-New-York-State-Validity-Study.pdf.

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

Datnow, A., & Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: A literature review of international research. *Journal of Educational Change*, *17*, 7-28.

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

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, *37*, 184-192.

Deno, S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual*. Reston, VA: Council for Exceptional Children.

Derosa, A. P., & Patalano, F. (1991). Effects of familiar proctor on fifth and sixth grade students' test anxiety. *Psychological reports*, *68*, 103-113.

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT‐Based cognitive diagnostic models and related methods. *Journal of Educational Measurement*, *44*, 285-291.

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools*, *43*, 247-265.

Elliott, S. N., & Roach, A. T. (2007). Alternate assessments of students with significant disabilities: Alternative approaches, common technical challenges. *Applied Measurement in Education*, *20*, 301-333.

Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., & Rooij, M. (2017). Data-based decision-making: Developing a method for capturing teachers' understanding of CBM graphs. *Learning Disabilities Research & Practice*, *32*, 8-21.

ESSA (2015). Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).

Fletcher, J. M., & Vaughn, S. (2009). Response to intervention: Preventing and remediating academic difficulties. *Child Development Perspectives*, *3*, 30-37.

Floyd, R. G., Hojnoski, R., & Key, J. (2006). Preliminary evidence of the technical adequacy of the preschool numeracy indicators. *School Psychology Review*, *35*, 627-644.

Foegen, A. (2000). Technical adequacy of general outcome measures for middle school mathematics. *Diagnostique*, *25*, 175-203.

Foegen, A., & Deno, S. L. (2001). Identifying growth indicators for low-achieving students in middle school mathematics. *The Journal of Special Education*, *35*, 4-16.

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education*, *41*, 121-139.

Foegen, A., Olson, J. R., & Impecoven-Lind, L. (2008). Developing progress monitoring measures for secondary mathematics: An illustration in algebra. *Assessment for Effective Intervention*, *33*, 240-249.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review*, *33*, 188-193.

Fuchs, L. S., & Deno, S. L. (1991). Effects of curriculum within curriculum-based measurement. *Exceptional Children*, *58*, 232-243.

Fuchs, L. S., Deno, S. L., & Marston, D. (1983). Improving the reliability of curriculum-based measures of academic skills for psychoeducational decision making. *Diagnostique*, *8*, 135-149.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it?. *Reading Research Quarterly*, *41*, 93-99.

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*, 311-330.

Fuchs, L. S., Fuchs, D., & Courey, S. J. (2005). Curriculum-based measurement of mathematics competence: From computation to concepts and applications to real-life problem solving. *Assessment for Effective Intervention*, *30*, 33-46.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlet, C., & Karns, K. M. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data sources. *School Psychology Review*, *29*, 65-85.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Thompson, A., Roberts, P. H., Kubek, P., & Stecker, P. M. (1994). Technical features of a mathematics concepts and applications curriculum-based measurement system. *Diagnostique*, *19*, 23-49.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., Dutka, S., & Katzaroff, M. (2000). The importance of providing information on the structure and scoring of performance assessments. *Applied Measurement in Education, 13*, 1–34

Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., et al. (2003). Explicitly teaching for transfer: Effects on third-grade students' mathematical problem solving. *Journal of Educational Psychology, 95*, 293–305.

Geary, D. C. (2011). Consequences, characteristics, and causes of mathematical learning disabilities and persistent low achievement in mathematics. *Journal of Developmental and Behavioral Pediatrics, 32*, 250-263.

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). *Assisting students struggling with mathematics: Response to Intervention (RtI) for elementary and middle schools* (Institute of Education Sciences Practice Guide). Washington, DC: U.S. Department of Education.

Ginsburg, H. P., Lee, Y. S., & Pappas, S. (2016). Using the clinical interview and curriculum based measurement to examine risk levels. *ZDM*, *48*, 1031-1048.

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

Grandy, J. (1987). *Characteristics of examinees who leave questions unanswered on the GRE general test under rights-only scoring*. ETS Research Report 87-38. Princeton, NJ: Educational Testing Service.

Great Minds (2015). *Eureka math curriculum.* Washington, DC: Great Minds.

Gulliksen, H., & Wilks, S. S. (1950). Regression tests for several samples. *Psychometrika*, *15*, 91-114.

Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (3rd ed.). Gaithersburg, MD.: Advanced Analytics, LLC.

Helwig, R., Anderson, L., & Tindal, G. (2002). Using a concept-grounded, curriculum-based measure in mathematics to predict statewide test scores for middle school students with LD. *The Journal of Special Education*, *36*, 102-112.

Helwig, R., & Tindal, G. (2002). Using general outcome measures in mathematics to measure adequate yearly progress as mandated by Title I. *Assessment for Effective Intervention*, *28*, 9-18.

Higgins J. P. T. , & Green, S. (Eds.). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

Hollenbeck, A. F. (2007). From IDEA to implementation: A discussion of foundational and future responsiveness-to-intervention research. *Learning Disabilities Research & Practice*, *22*, 137-146.

Hosp, J. L., Hensley, K., Huddle, S. M., & Ford, J. W. (2014). Using curriculum-based measures with postsecondary students with intellectual and developmental disabilities. *Remedial and Special Education*, *35*, 247-257.

Hughes, C. A., & Dexter, D. D. (2011). Response to intervention: A research-based summary. *Theory into Practice*, *50*, 4-11.

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004).

Jiban, C. L., & Deno, S. L. (2007). Using math and reading curriculum-based measurements to predict state mathematics test performance: Are simple one-minute measures technically adequate?. *Assessment for Effective Intervention*, *32*, 78-89.

Jitendra, A. K., Dupuis, D. N., & Zaslofsky, A. F. (2014). Curriculum-based measurement and standards-based mathematics: Monitoring the arithmetic word problem-solving performance of third-grade students at risk for mathematics difficulties. *Learning Disability Quarterly*, *37*, 241-251.

Jitendra, A. K., Sczesniak, E., & Deatline-Buchman, A. (2005). An exploratory validation of curriculum-based mathematical word problem-solving tasks as indicators of mathematics proficiency for third graders. *School Psychology Review*, *34*, 358-371.

Johnson, E. S., Galow, P. A., & Allenger, R. (2013). Application of algebra curriculum-based measurements for decision making in middle and high school. *Assessment for Effective Intervention*, *39*, 3-11.

Kane, M. T. (1992) An argument-based approach to validity. *Psychological Bulletin, 112*, 527 – 535.

Karkee, T., Kim, D. I., & Fatica, K. (2010, April). Comparability study of online and paper and pencil tests using modified internally and externally matched criteria. In *annual meeting of the American Educational Research Association, Denver, CO*.

Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review*, *37*, 374-390.

Kettler, R. J., & Albers, C. A. (2013). Predictive validity of curriculum-based measurement and teacher ratings of academic achievement. *Journal of School Psychology*, *51*, 499-515.

Ketterlin-Geller, L. R., Alonzo, J., Braun-Monegan, J., & Tindal, G. (2007). Recommendations for accommodations implications of (in) consistency. *Remedial and Special Education*, *28*, 194-206.

Keuning, T., Geel, M., & Visscher, A. (2017). Why a data-based decision-making intervention works in some schools and not in others. *Learning Disabilities Research & Practice*, *32*, 32-45.

Kim, D. H., & Huynh, H. (2007). Comparability of Computer and Paper-and-Pencil Versions of Algebra and Biology Assessments. Journal of Technology, Learning, and Assessment, 6. Retrieved May, 1, 2017 from http://www.jtla.org.

Kingston, N. M. (2008). Comparability of computer-and paper-administered multiple-choice tests for K–12 populations: A synthesis. *Applied Measurement in Education*, *22*, 22-37.

Klingner, J. K., & Edwards, P. A. (2006). Cultural considerations with response to intervention models. *Reading Research Quarterly, 41*, 108–117.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*, 1813-1824.

Koedel, C., & Tyhurst, E. (2012). Math skills and labor-market outcomes: Evidence from a resume-based field experiment. *Economics of Education Review*, *31*, 131-140.

Laracy, S. D., Hojnoski, R. L., & Dever, B. V. (2016). Assessing the classification accuracy of early numeracy curriculum-based measures using receiver operating characteristic curve analysis. *Assessment for Effective Intervention*, *41*, 172-183.

Lee, Y. S., & Lembke, E. (2016). Developing and evaluating a kindergarten to third grade CBM mathematics assessment. *ZDM*, *48*, 1019-1030.

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*, 107-117.

Lee, D., Reynolds, C. R., & Willson, V. L. (2003). Standardized test administration: Why bother? *Journal of Forensic Neuropsychology, 3*, 55-81.

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology*, *96*, 927-940.

Luo, W., & Kwok, O. M. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivariate Behavioral Research*, *44*, 182-212.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*, 304-316.

Makel, M. C., Plucker, J. A., Freeman, J., Lombardi, A., Simonsen, B., & Coyne, M. (2016). Replication of special education research: Necessary but far too rare. *Remedial and Special Education*, *37*, 205-212.

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, *47*, 71-85.

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education: Evidence from recent RAND research*. Santa Monica, CA: Rand.

Marston, D., Mirkin, P., & Deno, S. (1984). Curriculum-based measurement: An alternative to traditional screening, referral, and identification. *The Journal of Special Education*, *18*, 109-117.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*, 449-458.

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

Messick, S. (1993). *Foundation of validity: Meaning and consequences in psychological assessment*. Princeton, NJ: Educational Testing Service.

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

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13,* 241-256.

Methe, S. A., Hintze, J. M., & Floyd, R. G. (2008). Validation and decision accuracy of early numeracy skill indicators. *School Psychology Review*, *37*, 359-373.

Murphy, M. M., Mazzocco, M. M., Hanich, L. B., & Early, M. C. (2007). Cognitive characteristics of children with mathematics learning disability (MLD) vary as a function of the cutoff criterion used to define MLD. *Journal of Learning Disabilities*, *40*, 458-478.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core state standards mathematics.* Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.

Noam, G. G., & Hermann, C. A. (2002). Where education and mental health meet: Developmental prevention and early intervention in schools. *Development and Psychopathology*, *14*, 861-875.

No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002).

Noyes, J. M., & Garland, K. J. (2008). Computer vs. paper-based tasks: Are they equivalent?. *Ergonomics*, *51*, 1352-1375.

Nunnally, J. C. (1978). *Psychometric theory.* New York, NY: McGraw-Hill.

Organization for Economic Cooperation and Development [OECD] (2016). *PISA 2015 Results (Volume I): Excellent and equity in education, PISA*. Paris, France: OECD Publishing. doi.org/10.1787/9789264266490-en.

Peabody, D. (1962). Two components in bipolar scales: Direction and extremeness. *Psychological Review*, *69*, 65-73.

Pearce, N. (1996). Traditional epidemiology, modern epidemiology, and public health. *American Journal of Public Health*, *86*, 678-683.

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education*, *7*, 93-120.

Polignano, J. C., & Hojnoski, R. L. (2012). Preliminary evidence of the technical adequacy of additional curriculum-based measures for preschool mathematics. *Assessment for Effective Intervention*, *37*, 70-83.

Reynolds, A. M., Elksnin, N., & Brown, F. R. (1996). Specific reading disabilities: Early identification and long-term outcome. *Developmental Disabilities Research Reviews*, *2*, 21-27.

Rivera-Batiz, F. L. (1992). Quantitative literacy and the likelihood of employment among young adults in the United States. *Journal of Human Resources, 27*, 313–328.

Salaschek, M., & Souvignier, E. (2013). Web-based progress monitoring in first grade mathematics. *Frontline Learning Research*, *1*, 53-69.

Salaschek, M., & Souvignier, E. (2014). Web-based mathematics progress monitoring in second grade. *Journal of Psychoeducational Assessment*, *32*, 710-724.

Schildkamp, K., Ehren, M., & Lai, M. K. (2012). Editorial article for the special issue on data-based decision making around the world: From policy to practice to results. *School Effectiveness and School Improvement*, *23,* 123–131.

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement*, *41*, 1101-1114.

Seethaler, P. M., & Fuchs, L. S. (2011). Using curriculum-based measurement to monitor kindergarteners' mathematics development. *Assessment for Effective Intervention, 36,* 219-229.

Shapiro, E. S., Dennis, M. S., & Fu, Q. (2015). Comparing computer adaptive and curriculum-based measures of math in progress monitoring. *School Psychology Quarterly*, *30*, 470-487.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*, 5-24.

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

Shinn, M. R. (Ed.). (1998). *Advanced applications of curriculum-based measurement*. New York, NY: Guilford Press.

Snow, C. E., Bums, S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children: Report of the committee on the prevention of reading difficulties in young children*. Washington, DC: National Academy Press.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*, 795-819.

Sugai, G., & Horner, R. H. (2009). Responsiveness-to-intervention and school-wide positive behavior supports: Integration of multi-tiered system approaches. *Exceptionality*, *17*, 223-237.

Test, D. W., Mazzotti, V. L., Mustian, A. L., Fowler, C. H., Kortering, L., & Kohler, P. (2009). Evidence-based secondary transition predictors for improving postschool outcomes for students with disabilities. *Career Development for Exceptional Individuals*, *32*, 160-181.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837–847.

Thompson, S., Thurlow, M., & Moore, M. (2003). *Using computer-based tests with students with disabilities* (Policy Directions No. 15). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

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

Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review*, *31*, 498-513.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress. (2016). *2016 Mathematics assessment*. Retrieved from http://nces.ed.gov/nationsreportcard.

U.S. Department of Education, Office of Special Education and Rehabilitative Services, Office of Special Education Programs, *38th Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act*, *2016*, Washington, D.C. 2016.

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, *62*, 562-569.

Van den Bosch, R. M., Espin, C. A., Chung, S., & Saab, N. (2017). Data-based decision-making: Teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learning Disabilities Research & Practice, 32,* 46-60.

Vanderheyden, A. M., Broussard, C., Fabre, M., Stanley, J., Legendre, J., & Creppell, R. (2004). Development and validation of curriculum-based measures of math performance for preschool children. *Journal of Early Intervention*, *27*, 27-41.

VanDerHeyden, A. M., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. *School Psychology Review*, *30*, 363-382.

Wagner, D. L., Hammerschmidt-Snidarich, S. M., Espin, C. A., Steifert, K., & McMaster, K. L. (2017). Pre-service teachers' interpretation of CBM progress monitoring data. *Learning Disabilities Research & Practice, 32,* 22-31.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, *67*, 219-238.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, *68*, 5-24.

Winter, P. C., Kopriva, R. J., Chen, C. S., & Emick, J. E. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences*, *16*, 267-276.

Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment*, *13*, 170-192.

Ysseldyke, J. E., Thurlow, M., Graden, J., Wesson, C., Algozzine, B., & Deno, S. (1983). Generalizations from five years of research on assessment and decision making: The University of Minnesota Institute. *Exceptional Education Quarterly*, *4*, 75-93.

Zirkel, P. A., & Thomas, L. B. (2010). State laws and guidelines for implementing RTI. *Teaching Exceptional Children, 43*, 60–73.

APPENDIX A

DESCRIPTION OF CODED VARIABLES

Label	Description
Publication Date	• Identify the date of publication
Journal	• Identify the name of journal
Authors	• Identify the last name of all contributing authors
Development of m-CBM	• Identify the method used to develop the m-CBM • Curriculum sampling: measures are created by randomly sampling skills across the curriculum for that grade level. • Robust indicators: skills that are strongly correlated with mathematics success. Looking at discrete skills in isolation.
Focus of m-CBM	• Identify the mathematical concept/skill that is the focus of the m-CBM (e.g., computation, basic facts, concepts & applications, algebra, numeracy)
Format of Administration	• Identify the process for administering • Paper/pencil • Computer or tablet administered • Performance tasks consisted of the proctor providing a direction and observing student perform a skill
Proctor	• Identify who served as the proctor • Teacher • Researcher (i.e., Graduate students part of the research team are included) • Trained assessors

Grouping	• Identify the method used for administration • Whole class • Small group • Individual
Timing	• Identify whether a time limit was part of protocol • For performance tasks, if students were provided a set time to respond for proctor moved on than this was coded as a time limit
m-CBM	• Identify the name of the m-CBM(s) administered to students
Criterion Measures	• Identify the measures administered to students and selected as the criterion by which to compare m-CBM scores
Criterion Validity	• Identify whether the correlation coefficients reported are reporting the concurrent or predictive validity • Only report the correlation coefficients between the m-CBM(s) and the criterion measure(s)
Sample Size	• Identify the number of students who completed the m-CBM and the criterion measure • If authors included participants across grades report sample size by grade
Grade/Age of Participants	• Identify the grade and age range of participants

APPENDIX B

CRITERION VALIDITY OF M-CBMS

Study	Measures	Results
Allinder (1992)	*CBM* CBM-Comp *Criterion* SAT-Comp SAT-Conc	*Concurrent* CBM-Comp → SAT-Comp (.55-.93) CBM-Comp → SAT-Conc (.49-.88)
Baglici (2010)	*CBM* TEN-MN TEN-NI TEN-OC TEN-QD *Criterion* AIMSweb-Comp Grades ACES-M	*Predictive* TEN-MN (w/s) → ACES-M (.44/.58), AIMSweb-Comp (w) (.42/.52), AIMSweb-Comp (s) (.29/.47), Grades (.23/.33) TEN-NI (w/s) → ACES-M (.46/.57), AIMSweb-Comp (w) (.41/.37), AIMSweb-Comp (s) (.41/.26), Grades (.21/.33) TEN-OC (w/s) → ACES-M (.39/.36), AIMSweb-Comp (w) (.36/.31), AIMSweb-Comp (s) (.35/.33), Grades (.18/.05) TEN-QD (w/s) → ACES-M (.51/.40), AIMSweb-Comp (w) (.23/.03), AIMSweb-Comp (s) (.22/.02), Grades (.35/.24)
Betts (2009)	*CBM* MKA *Criterion* NALT	*Predictive* MKA-number sense → NALT (.53) MKA-pattern → NALT (.48) MKA-spatial → NALT (.37)
Chard (2005)	*CBM* C-20 CB2 CB5 CB10 CF3 CF6 MN NI NW QD *Criterion* NKT	*Predictive* Kindergarten C-20 (f/s) → NKT (.41/.38) CB2 (f/s) → NKT (.45/.49) CB5 (f/s) → NKT (.48/.53) CB10 (f/s) → NKT (.50/.55) CF3 (f/s) → NKT (.48/.40) CF6 (f/s) → NKT (.49/.39) MN (f/s) → NKT (.69/.64) NI (f/s) → NKT (.65/.58) NW (f/s) → NKT (.63/.57) QD (f/s) → NKT (.55/.50) First Grade C-20 (f/s) → NKT (.12/.17) CB2 (f/s) → NKT (.42/.43) CB5 (f/s) → NKT (.48/.45) CB10 (f/s) → NKT (.40/.40) CF3 (f/s) → NKT (.07/.13) CF6 (f/s) → NKT (.18/.19) MN (f/s) → NKT (.61/.61) NI (f/s) → NKT (.56/.58)

		NW (f/s) → NKT (.46/.54) QD (f/s) → NKT (.45/.53)
Clarke (2004)	*CBM* MN NI OC QD *Criterion* CBM-Comp NKT WJ-AP	*Concurrent* MN (f/w/s) → CBM-Comp (-/.75/.74), NKT (.74/-/-), WJ-AP (.68/-/.69) NI (f/w/s) → CBM-Comp (-/.66/.60), NKT (.70/-/-), WJ-AP (.65/-/.63) OC (f/w/s) → CBM-Comp (-/.49/.50), NKT (.70/-/-), WJ-AP (.64/-/.60) QD (f/w/s) → CBM-Comp (-/.71/.75), NKT (.80/-/-), WJ-AP (.71/-/.79) *Predictive* F-MN (w/s) → CBM-Comp (.78/.67), WJ-AP (-/.72) F-NI (w/s) → CBM-Comp (.68/.60), WJ-AP (-/.72) F-OC (w/s) → CBM-Comp (.56/.56), WJ-AP (-/.72) F- QD (w/s) → CBM-Comp (.76/.70), WJ-AP (-/.79) W-MN (s) → CBM-Comp (.72), WJ-AP (.71) W-NI (s) → CBM-Comp (.58), WJ-AP (.68) W-OC (s) → CBM-Comp (.46), WJ-AP (.68) W-QD (s) → CBM-Comp (.71), WJ-AP (.79)
Codding (2015)	*CBM* AIMSweb-Comp *Criterion* MCAS-M	*Concurrent* AIMSweb-Comp → MCAS-M (.35) *Predictive* AIMSweb-Comp (F/W) → MCAS-M (.26/.30)
Codding (2016)	*CBM* iSTEEP BF iSTEEP C&A iSTEEP Core *Criterion* MCAS-M	*Predictive* Sixth Grade BF (intercept/slope1/slope2) → MCAS-M (.43/.24/-.01) C&A (intercept/slope1/slope2) → MCAS-M (.74/.09/-.21) Core (intercept/slope1/slope2) → MCAS-M (.52/.13/.41) Seventh Grade BF (intercept/slop1) → MCAS-M (.46/.34) C&A (intercept/slope1) → MCAS-M (.77/-.18) Core (intercept/slope1) → MCAS-M (.46/.55) Eighth Grade BF (intercept/slop1) → MCAS-M (.45/-.08) C&A (intercept/slope1/slope2) → MCAS-M (.89/-.18/.20) Core (intercept/slope1/slope2)→ MCAS-M (.71/.34/.40)
Daly (1997)	*CBM* NC NP NR NS	*Concurrent* NC → WJ-Broad (.47) NP → WJ-Broad (.17) NR → WJ-Broad (.03) NS → WJ-Broad (.11)

	SN *Criterion* BF WJ-Broad	SN → WJ-Broad (.09) *Predictive* NC → BF (.39) NP → BF (.36) NR → BF (.07) NS → BF (.30) SN → BF (.04)
Eckert (2006)	*CBM* CBM addition of two digit by two digit CBM subtraction combinations to 18 CBM sum to 9 CBM sum to 18 *Criterion* Teacher Rating	*Concurrent* CBM addition of two digit by two digit → Teacher Rating (10-96%), r = .09 CBM subtraction combinations to 18 → Teacher Rating (13-89%) , r = .13 CBM sum to 9 → Teacher Rating (82-97%), r = .20 CBM sum to 18 → Teacher Rating (13-58%), r = .32
Floyd (2006)	*CBM* NNF OCF OOCCF QCF *Criterion* BBCS-Q TEMA WJ-AP	*Concurrent* NNF → BBCS-Q (.29), TEMA (.70), WJ-AP (.36) OCF → BBCS-Q (.31), TEMA (.55), WJ-AP (.33) OOCCF → BBCS-Q (.31), TEMA (.64), WJ-AP (.29) QCF → BBCS-Q (.46), TEMA (.58), WJ-AP (.38)
Foegen (2000)	*CBM* BF Estimation *Criterion* ITBS Grades Teacher Ranking Teacher Ratings	*Concurrent* BF → ITBS-Concepts & Estimation (.47-.54), ITBS-Problems & Data (.35-.36), ITBS-Total (45-.47), Grades (.52-.54), Teacher Ranking (.56-.62), Teacher Rating (.56-.60) Estimation → ITBS-Concepts & Estimation (.58-.64), ITBS-Problems & Data (.53-.59), ITBS-Total (.59-.66), Grades (.46-.52), Teacher Ranking (.51-.64), Teacher Rating (.49-.60)
Foegen (2001)	*CBM* BET BMOT MET *Criterion* CAT-Comp CAT-C&A GPA Teacher Rating	*Concurrent* BET → CAT-Comp (.56), CAT-Concept (.45), GPA (.39), TR-problem solving (.54), TR-proficiency (.49), TR-reasoning (.43) BMOT → CAT-Comp (.63), CAT-C&A (.44), GPA (.44), TR-problem solving (.54), TR-proficiency (.52), TR-reasoning (.49) MET → CAT-Comp (.47-.55), CAT-C&A (.29-.55), GPA (.22-.30), TR-problem solving (.42/.50), TR-proficiency (.39-.51), TR-reasoning (.40/.44) *Predictive* BET → CAT-Comp (.35), CAT-C&A (.32),

			TR-problem solving (.46), TR-proficiency (.36), TR-reasonsing (.36) BMOT → CAT-Comp (.44), CAT-C&A (.32), TR-problem solving (.45), TR-proficiency (.42), TR-reasoning (.43) MET → CAT-Comp (.30), CAT-C&A (.30), TR-problem solving (.25), TR-proficiency (.26), TR-reasonsing (.20)
Foegen (2008)	*CBM* BF Complex QD Estimation MBSP-Comp MBSP-C&A MN *Criterion* Grades ITBS NALT Teacher Rating	*Concurrent*	<u>Sixth Grade</u> BF → ITBS-Concepts & Estimation (.52), ITBS-Problems & Data (.38), ITBS-Total (.49), NALT (.50), TR (.56), Complex QD → ITBS-Concepts & Estimation (.52), ITBS-Problems & Data (.50), ITBS-Total (.53), NALT (.52), TR (.57) , , Estimation → ITBS-Concepts & Estimation (.51), ITBS-Problems & Data (.50), ITBS-Total (.53), NALT (.57), TR (.60) MBSP-Comp → ITBS-Concepts & Estimation (.60), ITBS-Problems & Data (.54), ITBS-Total (.59), NALT (.64), TR (.65) MBSP-C&A → ITBS-Concepts & Estimation (.72), ITBS-Problems & Data (.60), ITBS-Total (.71), NALT (.76), TR (.75) MN → ITBS-Concepts & Estimation (.45), ITBS-Problems & Data (.45), ITBS-Total (.46), NALT (.47), TR (.56) <u>Seventh Grade</u> Basic facts → ITBS-Concepts & Estimation (.55), ITBS-Problems & Data (.52), ITBS-Total (.55), NALT (.60), TR (.54) Complex QD → ITBS-Concepts & Estimation (.57), ITBS-Problems & Data (.57), ITBS-Total (.60), NALT (.57), TR (.42) Estimation → ITBS-Concepts & Estimation (.46), ITBS-Problems & Data (.47), ITBS-Total (.51), NALT (.45), TR (.26) MBSP-Comp → ITBS-Concepts & Estimation (.34), ITBS-Problems & Data (.33), ITBS-Total (.38), NALT (.38), TR (.35) MBSP-C&A → ITBS-Concepts & Estimation (.87), ITBS-Problems & Data (.80), ITBS-Total (.87), NALT (.86), TR (.73), MN → ITBS-Concepts & Estimation (.53), ITBS-Problems & Data (.51), ITBS-Total (.54), NALT (.67), TR (.42) <u>Eighth Grade</u> Complex QD → ITBS-Concepts & Estimation (.41), ITBS-Problems & Data (.52), ITBS-Total (.55), TR (.51) Estimation → ITBS-Concepts & Estimation

		(.43), ITBS-Problems & Data (.52), ITBS-Total (.53), TR (.40),
		MN → ITBS-Concepts & Estimation (.31), ITBS-Problems & Data (.41), ITBS-Total (.47), TR (.50)
		Predictive
		Sixth Grade
		BF → NALT (.55)
		Complex QD → NALT (.53)
		Estimation → NALT (.55)
		MBSP-Comp → NALT (.64)
		MBSP-C&A → NALT (.76)
		MN → NALT (.48)
		Seventh Grade
		BF → NALT (.59)
		Complex QD → NALT (.58)
		Estimation → NALT (.34)
		MBSP-Comp → NALT (.25)
		MBSP-C&A → NALT (.87)
		MN → NALT (.60)
Fuchs (1989)	*CBM*	*Concurrent*
	BF	General Education
	Criterion	BF → TR (.73)
	Teacher Rating	Special Education
		BF → TR (.86)
		First Grade: BF → TR (.62)
		Second Grade: BF → TR (.69)
		Third Grade: BF → TR (.65)
		Fourth Grade: BF → TR (.83)
		Fifth Grade: BF → TR (.71)
		Sixth Grade: BF → TR (.88)
Fuchs (1994)	*CBM*	*Concurrent*
	CBM-C&A	Second Grade
	Criterion	CBM-C&A → CTBS-Comp (.74), CTBS-C&A (.76), CTBS-Total (.81)
	CTBS-Comp	Third Grade
	CTBS-C&A	CBM-C&A → CTBS-Comp (.73), CTBS-C&A (.64), CTBS-Total (.74)
	CTBS-Total	Fourth Grade
		CBM-C&A → CTBS-Comp (.74), CTBS-C&A (.75), CTBS-Total (.79)
Fuchs (2000)	*CBM*	*Concurrent*
	PAs	PAs → CTBS (0.48-0.68)
	Criterion	PAs → ITBS (0.60-0.67)
	CTBS	PAs → MOAT (0.48-0.68)
	ITBS	
	MOAT	
Fuchs (2003)	*CBM*	*Concurrent*
	Far WP	Far WP → TerraNova (0.67)

	Immediate WP Near WP *Criterion* TerraNova	Immediate WP → TerraNova (0.58) Near WP → TerraNova (0.55)
Fuchs (2007)	*CBM* CBM-Comp CBM-C&A Fact Retrieval NI/C *Criterion* WPs WRAT-Arithmetic	*Predictive* CBM-Comp → WPs (.35), WRAT-Arithmetic (.34) CBM-Comp slope → WPs (.28), WRAT-Arithmetic (.34) CBM-C&A → WPs (.44), WRAT-Arithmetic (.40) Fact retrieval → WPs (.10), WRAT-Arithmetic (.14) NI/C → WPs (.39), WRAT-Arithmetic (.34) NI/C slope → WPs (-.19), WRAT-Arithmetic (-.11)
Ginsburg (2016)	*CBM* BF CBM-C&A CBM-Comp CI-Addition CI-Counting CI-Multiplication CI-Subtraction CI-Written Number Counting MN NI NN QD *Criterion* WJ-Broad	*Concurrent* Kindergarten CBM Risk → CI-Addition (.32-.37), CI-Counting (.11-.54), CI-Subtraction (.19-.29) First Grade CBM Risk → CI-Addition (.21-.43), CI-Counting (.07-.42), CI-Subtraction (.29-.41), CI-Written Number (.04-.17) Second Grade CBM Risk → CI-Addition (.09-.42), CI-Multiplication (.14-.33), CI-Subtraction (.13-.32), CI-Written Number (.06-.32) Third Grade CBM Risk → CI-Addition (.07-.29), CI-Multiplication (.20-.35), CI-Subtraction (.12-.28), CI-Written Number (.14-.29) *Predictive* Kindergarten Significant predictor to WJ-App → CI-Addition Significant predictor to WJ-Broad → CI-Subtraction Significant predictors to WJ-Calc → CI-Addition, CI-Subtraction, MN, QD Significant predictors to WJ-Math Fluency → CI-Addition, CI-Subtraction, MN First Grade Significant predictor to WJ-App → CI-Addition Significant predictor to WJ-Broad → CI-Written Number Significant predictors to WJ-Calc → BF, CI-Addition, CI-Written Number Significant predictors to WJ-Math Fluency → BF, CI-Addition, CI-Written Number Second Grade Significant predictor to WJ-App → None

		Significant predictor to WJ-Broad → BF, CI-Multiplication, CI-Written Number Significant predictor to WJ-Calc → None Significant predictor to WJ-Fluency → None <u>Third Grade</u> Significant predictors to WJ-App → CBM-C&A, CI- Multiplication, CI-Subtraction Significant predictors to WJ-Broad → CBM-C&A, CI-Multiplication, CI-Subtraction Significant predictors to WJ-Calc → CBM-Comp, CBM-C&A, CI-Multiplication, CI-Subtraction Significant predictors to WJ-Math Fluency → BF, CBM-Comp, MN, QD
Helwig (2002)	*CBM* CBM-C&A *Criterion* CAT-MA	*Concurrent* <u>General Education</u> CBM-C&A → CAT-MA (.80) <u>Special Education</u> CBM-C&A → CAT-MA (.61) *Predictive* CBM-C&A + Classification (Sped/Not) → Passing CAT-MA (87.1% correct)
Helwig & Tindal (2002)	*CBM* GOMs *Criterion* State Assessment	*Concurrent* GOM 3 → State Assessment (.84) GOM 4 → State Assessment (.82) *Predictive* GOM 1 → State Assessment (.81) GOM 2 → State Assessment (.87)
Hosp (2014)	*CBM* AIMSweb-Comp AIMSweb-C&A *Criterion* WJ-App WJ-Broad WJ-Comp WJ-Math Fluency	*Concurrent* AIMSweb-Comp-Correct Digits → WJ-App (.40), WJ-Broad (.67), WJ-Comp (.75), WJ-Math Fluency (.73) AIMSweb-Comp-Correct Problems → WJ-App (.47), WJ-Broad (.70), WJ-Comp (.76), WJ-Math Fluency (.65) AIMSweb-C&A Correct Problems → WJ-App (.71), WJ-Broad (.81), WJ-Comp (.77), WJ-Math Fluency (.59) AIMSweb-C&A Points → WJ-App (.70), WJ-Broad (.80), WJ-Comp (.74), WJ-Math Fluency (.63)
Jiban (2007)	*CBM* BF Cloze Math Facts *Criterion* MCA-Math	*Predictive* <u>Third Grade</u> BF problems correct → MCA-Math (.11) BFcorrect minus incorrect → MCA-Math (.26) Cloze Math Facts problems correct → MCA-Math (.38) Cloze Math Facts correct minus incorrect → MCA-Math (.44) <u>Fifth Grade</u>

		BF problems correct → MCA-Math (.55) BFcorrect minus incorrect → MCA-Math (.57) Cloze Math Facts problems correct → MCA-Math (.59) Cloze Math Facts correct minus incorrect → MCA-Math (.59)
Jitendra (2005)	*CBM* CBM-Comp WPs *Criterion* TerraNova-Comp TerraNova-C&A SAT-Procedures SAT-PS	*Concurrent* CBM-Comp (w) → SAT-Procedures (.64), SAT-PS (.49) CBM-Comp (s) → SAT-Procedures (.66), SAT-PS (.50), TerraNova Comp (.51), TerraNova C&A (.45), WPS (w) → SAT-Procedures (.58), SAT-PS (.71) WPS (s) → SAT-Procedures (.38), Stanford-PS (.54), TerraNova Comp (.48), TerraNova C&A (.58) *Predictive* CBM-Comp (w) → TerraNova-Comp (.59), TerraNova C&A (.38) WPS (w) → TerraNova Comp (.62), TerraNova C&A (.69)
Jitendra (2014)	*CBM* WPs *Criterion* MAP-M	*Predictive* Time 1: WPs → MAP-M (.38) Time 2: WPs → MAP-M (.37) Time 3: WPs → MAP-M (.45)
Johnson (2012)	*CBM* Basic Skills Algebra *Criterion* ISAT	*Predictive* <u>Seventh Grade</u> Basic Skills-Algebra → ISAT (.67) <u>Eighth Grade</u> Basic Skills-Algebra → ISAT (.68) <u>Tenth Grade</u> Basic Skills-Algebra → ISAT (.68)
Kettler (2013)	*CBM* MBSP-Comp PSG-M *Criterion* MAP-M SAT WKCE	*Concurrent* <u>First Grade</u> MBSP-Comp → MAP-M (.63) PSG Math → MAP-M (.74) <u>Second Grade</u> MBSP-Comp → MAP-M (.66) PSG Math → MAP-M (66) <u>Third Grade</u> MBSP-Comp → MAP-M (.61) PSG Math → MAP-M (.68) *Predictive* <u>First Grade</u> MBSP-Comp → SAT (46%, .44) PSG Math → SAT (50%, .38) <u>Second Grade</u> MBSP-Comp → SAT (35%, .14), WKCE (42%, .29)

		PSG Math → SAT (51%, .44), WKCE (44%, .41) <u>Third Grade</u> MBSP-Comp → SAT (36%, .25), WKCE (40%, .37) PSG Math → SAT (35%, .27), WKCE (53%, .48)
Klinkenberg (2011)	*CBM* Maths Garden *Criterion* CITO	*Concurrent* Maths Garden: Addition → CITO (.83) Maths Garden: Division → CITO (.78) Maths Garden: Multiplication → CITO (.80) Maths Garden: Subtraction → CITO (.84)
Laracy (2016)	*CBM* NN OC OOCC QC *Criterion* TEN-QD	*Predictive* NN (f/w) → below 40^{th} percentile Ten-QD (.72/.73) NN (w/s) → below 25^{th} percentile Ten-QD (.76/.74) OC (f/w) → below 40^{th} percentile Ten-QD (.70/.73) OC (w/s) → below 25^{th} percentile TEN-QD (.77/.70) OOCC (f/w) → below 40^{th} percentile Ten-QD (.66/.65) OOCC (w/s) → below 25^{th} percentile Ten-QD (.71/.69) QC (f/w) → below 40^{th} percentile Ten-QD (.72/.76) QC (w/s) → below 25^{th} percentile-QD (.77/.82)
Lee (2012)	*CBM* BF MN NI NN OC QD *Criterion* TEMA	*Concurrent* <u>Kindergarten</u> MN → TEMA (.62) NI → TEMA (.68) OC → TEMA (.53) QD → TEMA (.64) <u>First Grade</u> BF → (.50) MN → TEMA (.56) NI → TEMA (.68) NN → TEMA (.59) OC → TEMA (.40) QD → TEMA (.48) *Predictive* <u>Kindergarten</u> MN → TEMA (.59) NI → TEMA (.45) OC → TEMA (.54) QD → TEMA (.51) <u>First Grade</u> BF → TEMA (.37) MN → TEMA (.37)

		NI → TEMA (.53) NN → TEMA (.55) OC → TEMA (.25) QD → TEMA (.40)
Lee (2016)	*CBM*	*Concurrent*
	BF	Kindergarten
	CBM-Comp	MN → WJ-Broad (.37)
	CBM-C&A	NI → WJ-Broad (.32)
	MN	OC → WJ-Broad (.36)
	NI	QD → WJ-Broad (.44)
	NN	First Grade
	OC	BF → WJ-Broad (.52)
	QD	MN → WJ-Broad (.35)
	Criterion	NI → WJ-Broad (.46)
	WJ-Broad	NN → WJ-Broad (.51)
		OC → WJ-Broad (.14)
		QD → WJ-Broad (.51)
		Second Grade
		BF → WJ-Broad (.58)
		CBM-Comp → WJ-Broad (.32)
		CBM-C&A → WJ-Broad (.30)
		MN → WJ-Broad (.48)
		QD → WJ-Broad (.25)
		Third Grade
		BF → WJ-Broad (.49)
		CBM-Comp → WJ-Broad (.46)
		CBM-C&A → WJ-Broad (.55)
		MN → WJ-Broad (.54)
		QD → WJ-Broad (.35)
Methe (2008)	*CBM*	*Concurrent*
	COF	COF (f/s) → Teacher Rating (.68/.70), TEMA (.50/.55)
	MQF	MQF (f/s) → Teacher Rating (.70/.66), TEMA (.55/.20)
	NRF	NRF (f/s) → Teacher Rating (.89/.89), TEMA (.72/.64)
	OPF	OPF (f/s) → Teacher Rating (.81/.79), TEMA (.63/.60)
	Criterion	*Predictive*
	Teacher Rating	COF (f/w) → Teacher Rating (.57/.70), TEMA (.46/.62)
	TEMA	MQF (f/w) → Teacher Rating (.72/.61), TEMA (.41/.47)
		NRF (f/w) → Teacher Rating (.87/.88), TEMA (.70/.66)
		OPF (f/w) → Teacher Rating (.79/.77), TEMA (.58/.57)
Polignano (2012)	*CBM*	*Concurrent*
	CAR	CAR → BBCS-Q (.42), PNI-NNF (.65), PNI-OCF (.52), PNI-OOCCF (.61), PNI-QCF (.41),
	PC	

	SC SNF SSF *Criterion* BBCS-Q PNI-NNF PNI-OCF PNI-OOCCF PNI-QCF TEMA	TEMA (.72) PC → BBCS-Q (.59), PNI-NNF (.57), PNI-OCF (.44), PNI-OOCCF (.22), PNI-QCF (.45), TEMA (.72), SC → BBCS-Q (.53), PNI-NNF (.48), PNI-OCF (.33), PNI-OOCCF (.36), PNI-QCF (.26), TEMA (.39), SNF → BBCS-Q (.44), PNI-NNF (.56), PNI-OCF (.51), PNI-OOCCF (.25), PNI-QCF (.23), TEMA (.53) SSF → BBCS-Q (.64), PNI-NNF (.57), PNI-OCF (.44), PNI-OOCCF (.28), PNI-QCF (.48), TEMA (.45)
Salaschek (2013)	*CBM* Addition Equation ND NI NL NSeq1 NSeq2 Subtraction SQD *Criterion* DEMAT 1+ DEMAT 2+ OTZ Teacher rating	*Concurrent* CBMs → OTZ (.40-.50) *Predictive* CBMs → DEMAT 1+ (.64-.71) CBMs → DEMAT 2+ (.61-.68) CBMs → Teacher ratings (beg 1) (.29-.42) CBMs → Teacher ratings (end 1) (.54-.64) CBMs → Teacher ratings (end 2) (.54-.66)
Salaschek (2014)	*CBM* Number Sense CBM-Comp *Criterion* DEMAT1+ DEMAT2+ Teacher Rating	*Concurrent* CBM-Comp → DEMAT 1+ (.49-.56) Number Sense → DEMAT 1+ (.54-.62) CBM-Comp + Number Sense → DEMAT 1+ (.59-.63), Teacher rating (.57-.61) *Predictive* CBM-Comp → DEMAT 2+ (.64-.72) Number Sense → DEMAT 2+ (.66-.69) CBM-Comp + Number Sense → DEMAT 2+ (.72-.77), Teacher rating (.66-.70)
Seethaler (2011)	*CBM* CBM-Comp *Criterion* TEMA	*Concurrent* CBM-Comp (intercept) → TEMA (.69) *Predictive* CBM-Comp (intercept/slope)→ TEMA (.61/.49)
Shapiro (2015)	*CBM* AIMSweb-Comp AIMSweb-C&A STAR-Math *Criterion* PSSA	*Concurrent* Third Grade AIMSweb-Comp → PSSA (.61) AIMSweb-C&A → PSSA (.61) STAR-Math → PSSA (.82) Fourth Grade STAR-Math → PSSA (.88)

		AIMSweb-Comp → PSSA (.75) AIMSweb-C&A → PSSA (.24) <u>Fifth Grade</u> AIMSweb-Comp → PSSA (.74) AIMSweb-C&A → PSSA (.49) STAR-Math → PSSA (.70) *Predictive* <u>Third Grade</u> AIMSweb-Comp + AIMSweb-Comp (slope) + AIMSweb-C&A + AIMSweb-C&A (slope) + STAR-Math + STAR-Math (slope) → PSSA (72%) <u>Fourth Grade</u> AIMSweb-Comp + AIMSweb-Comp (slope) + AIMSweb-C&A + AIMSweb-C&A (slope) + STAR-Math + STAR-Math (slope) → PSSA (82%) <u>Fifth Grade</u> AIMSweb-Comp + AIMSweb-Comp (slope) + AIMSweb-C&A + AIMSweb-C&A (slope) + STAR-Math + STAR-Math (slope) → PSSA (71%)
Thurber (2002)	*CBM* BF CBM-Comp *Criterion* CAT-Comp CAT-C&A NAEP SDMT-App SDMT-Comp	*Concurrent* BF → CAT-Comp (.62-.66), CAT-C&A (.50-.55), NAEP (.45-.52), SDMT-App (.47-.51), SDMT-Comp (.61-.67) CBM-Comp → CAT-Comp (.59-.63), CAT-C&A (.44-.51), NAEP (.38-.44), SDMT-App (.36-.42), SDMT-Comp (.54-.59)
VanDerHeyden (2001)	*CBM* Circle number Discrimination Draw circles Write number *Criterion* CIBS-Q Teacher Ranking	*Concurrent* Circle number → CIBS-Q (.61), Teacher Ranking (.30) Discrimination → CIBS-Q (.56) Draw circles → CIBS-Q (.52), Teacher Ranking (.45-.64) Write number → CIBS-Q (.44), Teacher Ranking (.43)
VanDerHeyden (2004)	*CBM* Choose number Choose shape Count objects Discrimination Free count Number naming *Criterion* CIBS-Q Teacher Rating TEMA	*Concurrent* Choose number → CIBS-Q (.57), Teacher Rating (.43-.81), TEMA (.52) Choose shape → CIBS-Q (.06), Teacher Rating (.51-.57), TEMA (.38) Count objects → CIBS-Q (.44), Teacher Rating (.43-.56), TEMA (.49) Discrimination → CIBS-Q (.55), Teacher Rating (.51-.85), TEMA (.50)

	Free count → CIBS-Q (.56), Teacher Rating (.26-.91), TEMA (.19) Number naming → CIBS-Q (.47), Teacher Rating (.53-.91), TEMA (.39)

APPENDIX C

NOMENCLATURE FOR ASSESSMENTS

ACES-M = Academic Competence Evaluation Scales-Mathematics

AIMSweb-Comp = AIMSweb-Computation

AIMSweb-C&A = AIMSweb-Concepts and Applications

BBCS-Q = Bracken Basic Concepts Scale-Quantitative

BET = Basic Estimation Task

BF = Basic Facts

BMOT = Basic Math Operations Task

CAR = Cardinality

CAT-Comp = California Achievement Test-Computation

CAT-C&A = California Achievement Test-Concepts and Applications

CAT-MA = Computer Adaptive Test of Math Achievement

CBM-Comp = CBM-Computation

CBM-C&A = CBM-Concepts and Applications

CB2 = Count by 2s

CB5 = Count by 5s

CB10 = Count by 10s

CF3 = Count from 3

CF6 = Count from 6

CIBS-Q = Comprehensive Inventory of Basic Skills-Quantitative

CITO = National Institute for Educational Measurements-Math Tests

CI-Counting = Clinical Interview-Counting

CI-Addition = Clinical Interview-Addition

CI-Subtraction = Clinical Interview-Subtraction

CI-Multiplication = Clinical Interview-Multiplication

CI-Written Number = Clinical Interview-Written Number

COF = Count on Fluency

Complex QD = Complex Quantity Discrimination

CTBS-Comp = California Test of Basic Skills-Computation

CTBS-C&A = California Test of Basic Skills-Concepts and Applications

CTBS-Total = California Test of Basic Skills-Total Math

Count to 20 = C-20

GOMs = General Outcome Measures within Mathematics

ISAT = Idaho Standardized Assessment Test

iSTEEP BF = iSTEEP Basic Facts

iSTEEP C&A = iSTEEP Concepts and Applications

iSTEEP Core = iSTEEP Common Core

ITBS = Iowa Test of Basic Skills

MAP-M = Measures of Academic Progress-Mathematics

MBSP-Comp = MBSP Basic Math Computation

MBSP-C&A = MBSP Concepts and Applications

MCA-Math = Minnesota Comprehensive Assessment in Mathematics

MCAS-M = Massachusetts Comprehensive Assessment System-Mathematics

MET = Modified Estimation Task

MKA = Minneapolis Kindergarten Assessment

MN = Missing Number

MOAT = Mathematics Operations and Applications Test

MQF = Match Quantity Fluency

NAEP = National Assessment of Educational Progress-Mathematics

NALT = Northwest Achievement Levels Test

NC = Number Counting

ND = Number Discrimination

NI = Number Identification

NI/C = Number Identification/Counting

NKT = Number Knowledge Test

NL = Number Line

NN = Number Naming

NNF = Number Naming Fluency

NP = Number Production

NR = Number Reading

NRF = Number Recognition Fluency

NS = Number Selection

NSeq = Number Sequence

NW = Number Writing

OC = Oral Counting

OCF = Oral Counting Fluency

OOCC= One-to-One Correspondence Counting

OOCCF = One-to-One Correspondence Counting Fluency

OPF = Ordinal Position Fluency

OTZ = Osnabrück Test of Number Concept Development

PAs = Performance Assessments

PC = Pattern Completion

PNI-NNF = Preschool Numeracy Indicators-Number Naming Fluency

PNI-OCF = Preschool Numeracy Indicators-Oral Counting Fluency

PNI-OOCCF = Preschool Numeracy Indicators-One-to-One Correspondence Counting Fluency

PNI-QCF = Preschool Numeracy Indicators-Quantity Comparison Fluency

PSG-M = Performance Screening Guides-Mathematics

PSSA = Pennsylvania System of School Assessment

SAT-Conc = Stanford Achievement Test-Concepts of Number

SAT-Comp = Stanford Achievement Test-Computation

SAT PS = Stanford Achievement Test-Problem Solving

SAT Proc = Stanford Achievement Test-Procedures

SC = Shape Composition

SDMT-App = Stanford Diagnostic Mathematics Test-Applications

SDMT-Comp = Stanford Diagnostic Mathematics Test-Computation

SN = Shape Naming

SNF = Shape Naming Fluency

SQD = Symbol Quantity Discrimination

SSF = Shape Selection Fluency

TEMA = Test of Early Mathematics Ability

TEN-MN = Tests of Early Numeracy-Missing Number

TEN-NI = Tests of Early Numeracy-Number Identification

TEN-OC = Tests of Early Numeracy-Oral Counting

TEN-QD = Tests of Early Numeracy-Quantity Discrimination

TerraNova-Comp = TerraNova-Computation

TerraNova-C&A = TerraNova-Concepts and Applications

TRF = Teacher Rating Form

QC = Quantity Comparison

QCF = Quantity Comparison Fluency

QD = Quantity Discrimination

WJ-App = Woodcock-Johnson-Applied Problems

WJ-Broad = Woodcock-Johnson-Broad Math

WJ-Comp = Woodcock-Johnson-Computation

WJ-Math Fluency = Woodcock-Johnson-Math Fluency

WKCE = Wisconsin Knowledge and Concepts Examination

WPs = Word Problems

WRAT-Arithmetic = Wide Ranging Achievement Test- Arithmetic

APPENDIX D

STATISTICALLY SIGNIFICANT MODERATING VARIABLES

	Demographic Variables					
	Gender	Race	SPED	FARMS	LEP	Section 504
3				F		
4	W		S			
5			F W S	W	F	W S
6	F W S		S	F S		
7				F W S	F W S	
8	F		F	S		W

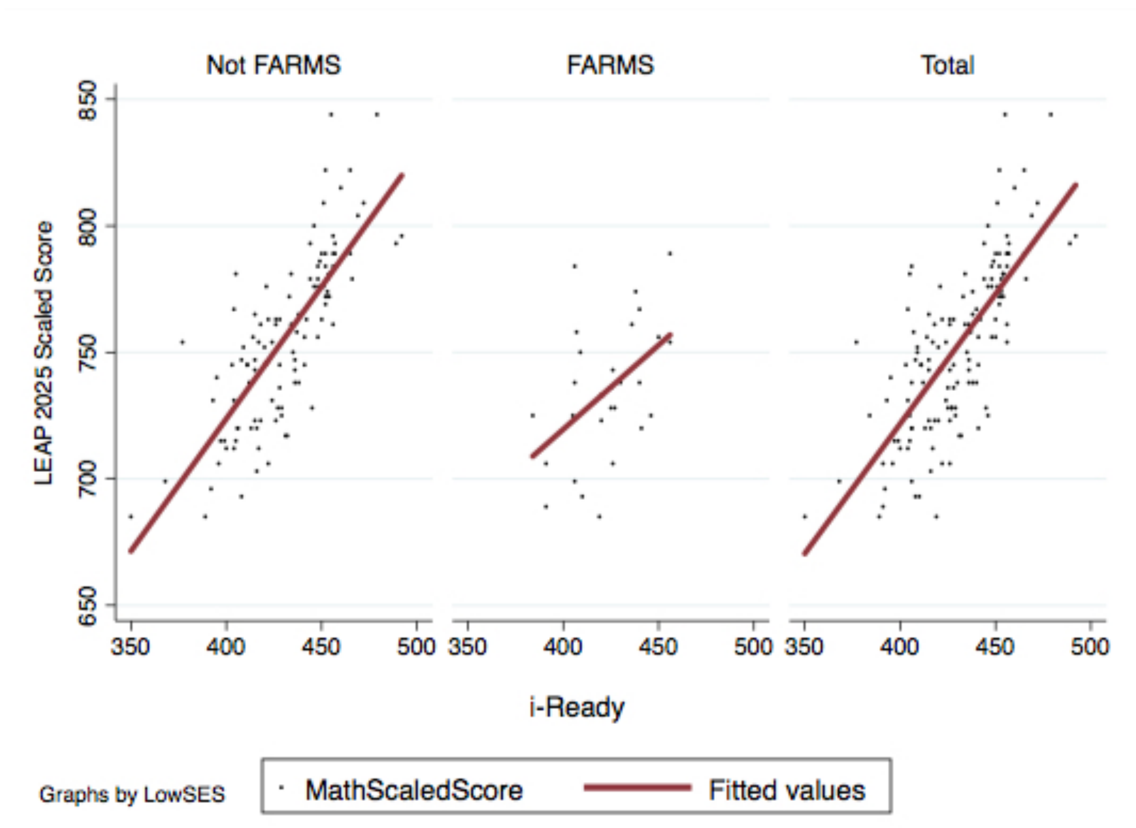
Note. F = Fall i-Ready, W = Winter i-Ready, S = Spring i-Ready

APPENDIX E

GRAPHS OF INTERACTION EFFECTS FOR STATISTICALLY SIGNIFICANT MODERATING VARIABLES

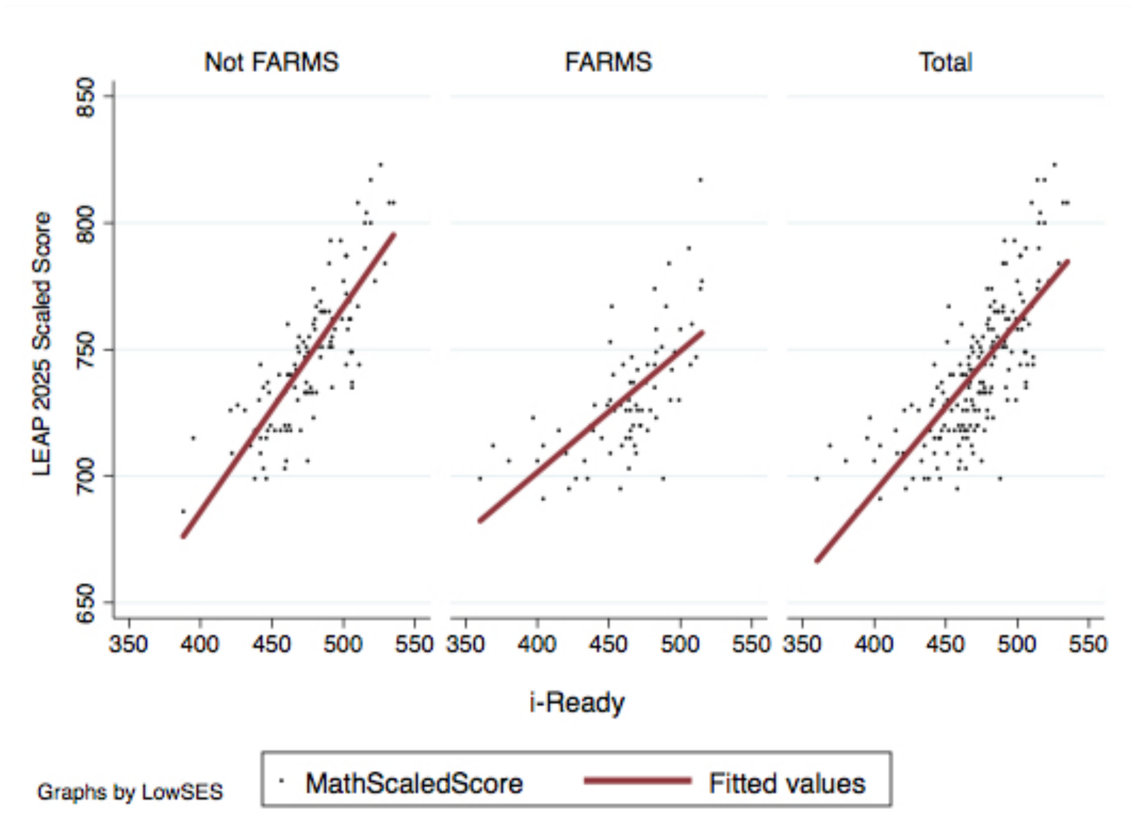
Third Grade

Predictive Validity Fall i-Ready



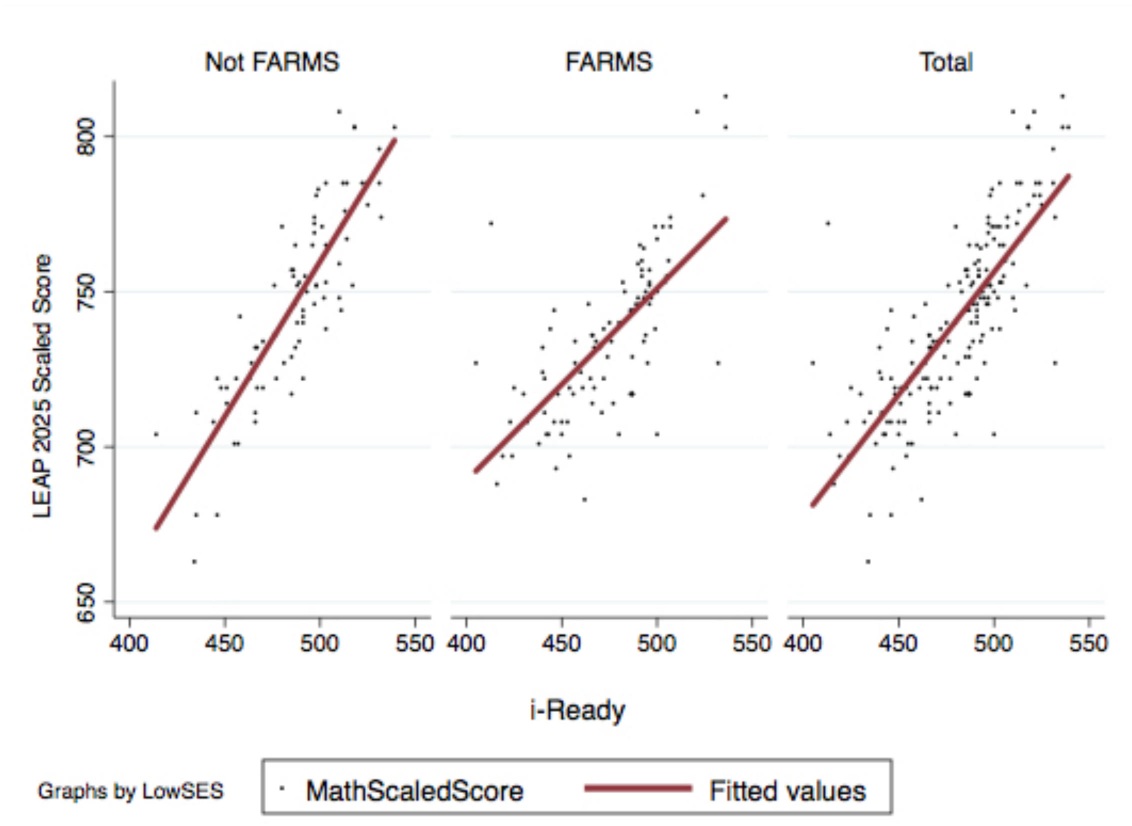
Fifth Grade

Predictive Validity of Winter i-Ready



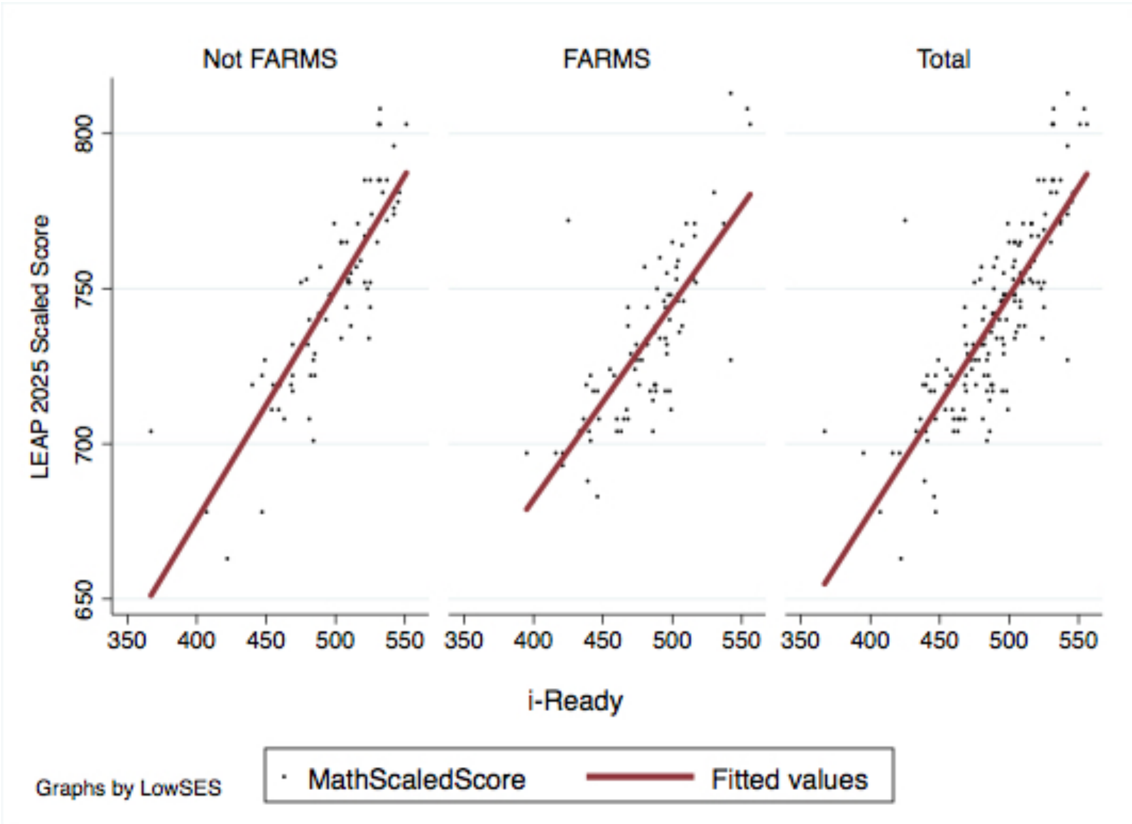
Sixth Grade

Predictive Validity of Fall i-Ready



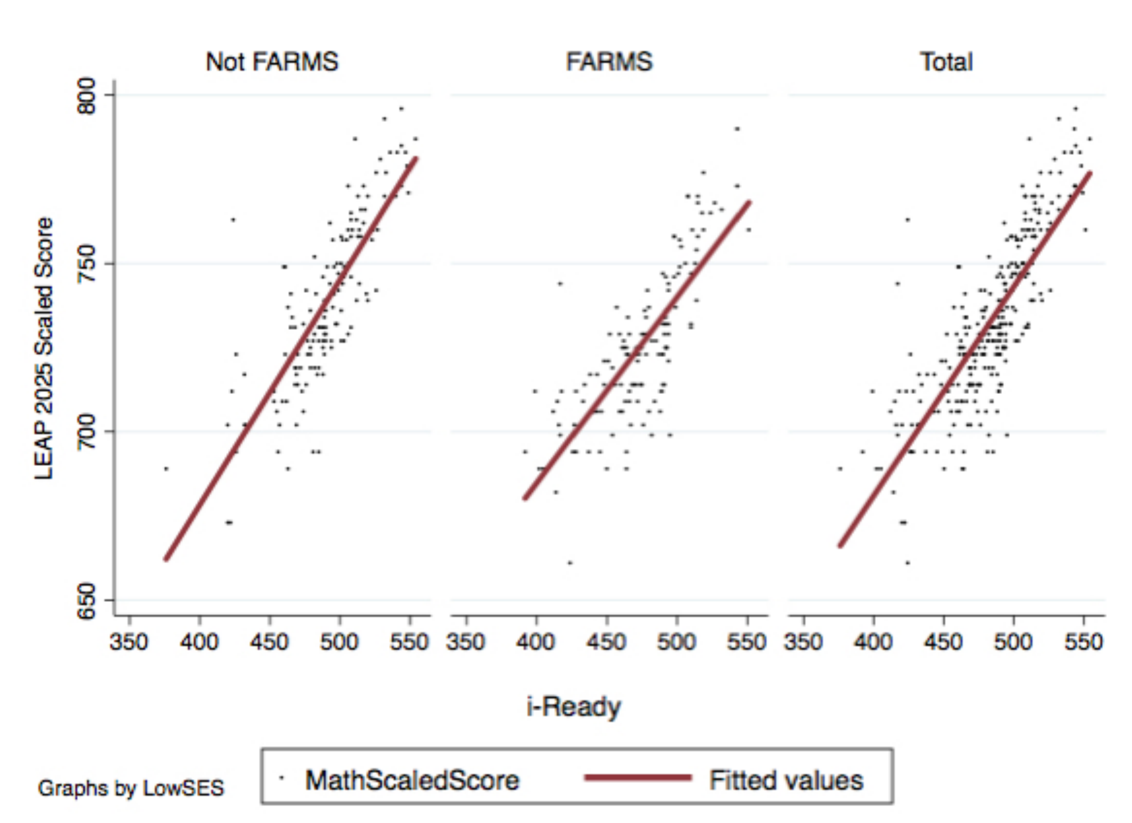
Sixth Grade

Concurrent Validity of Spring i-Ready



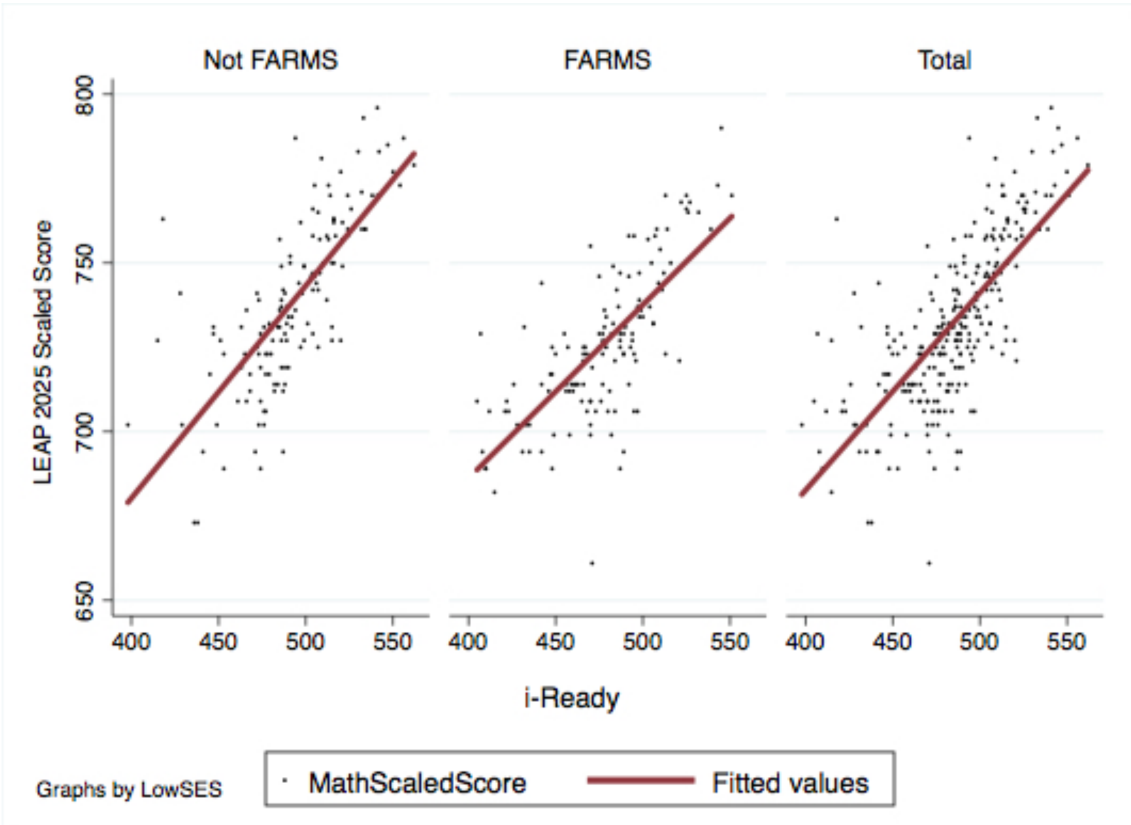
Seventh Grade

Predictive Validity of Fall i-Ready



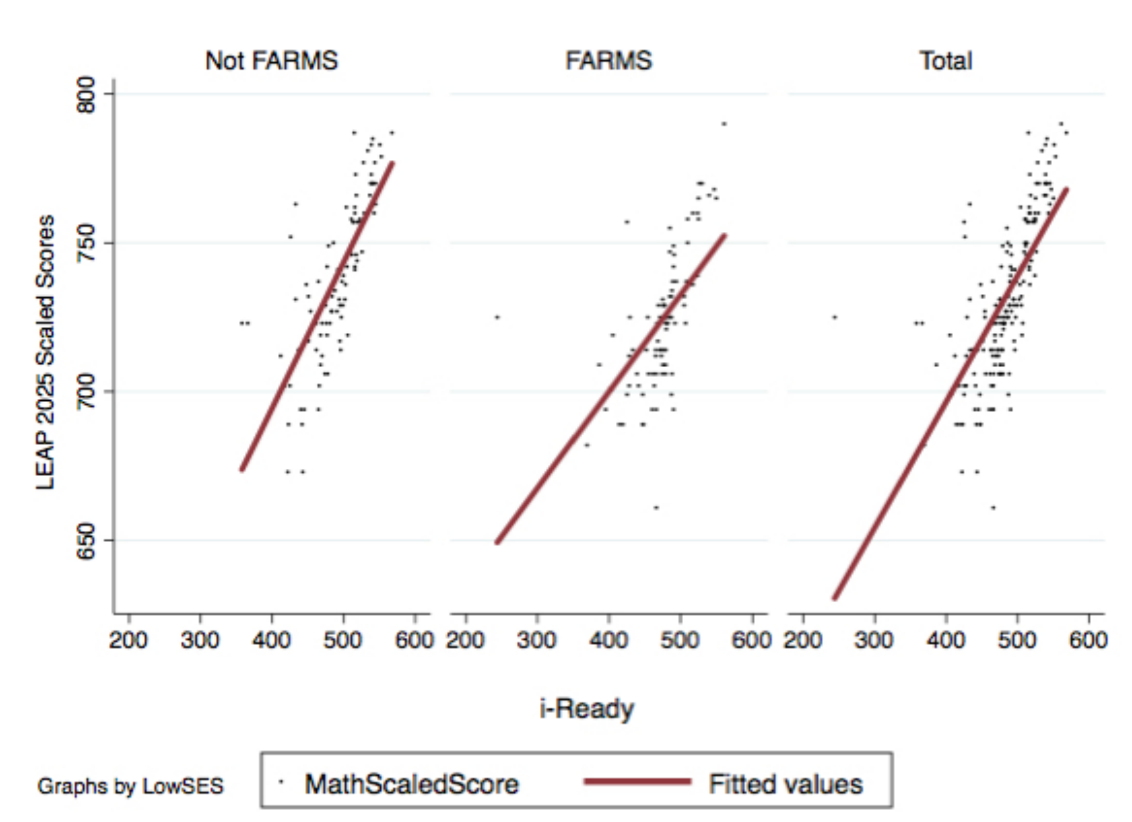
Seventh Grade

Predictive Validity of Winter i-Ready



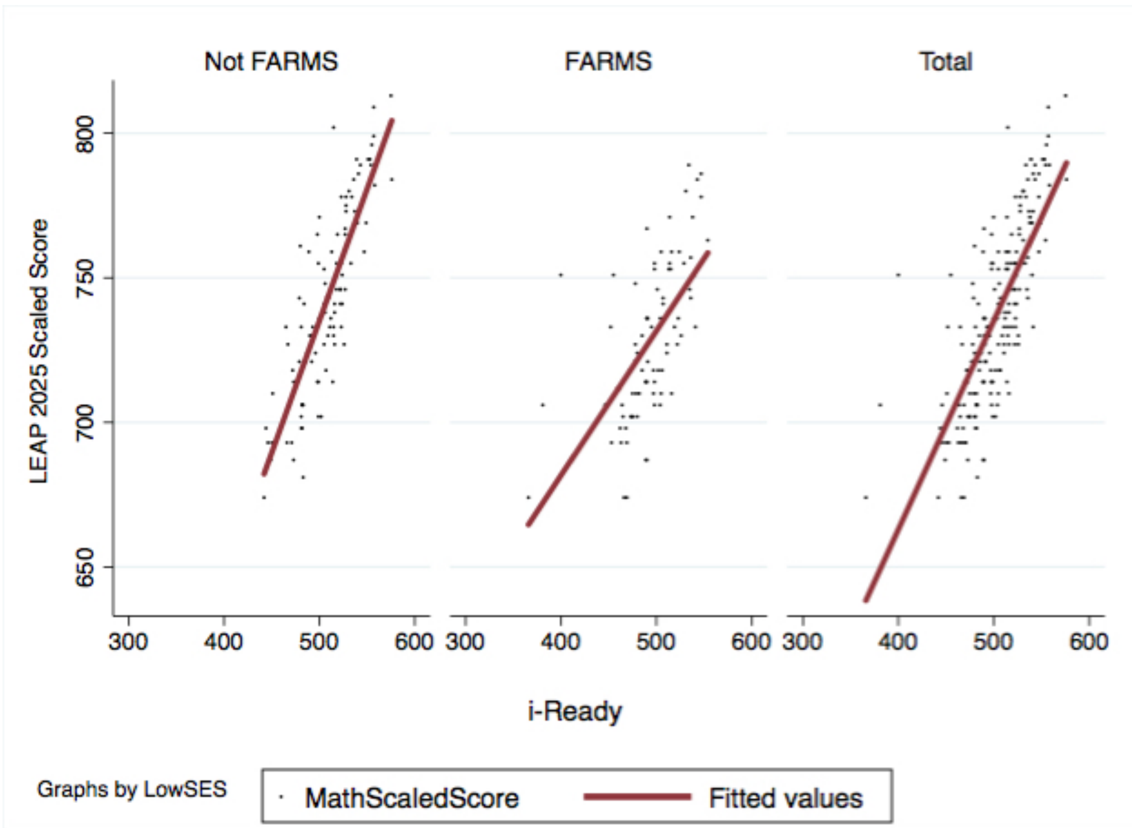
Seventh Grade

Concurrent Validity of Spring i-Ready



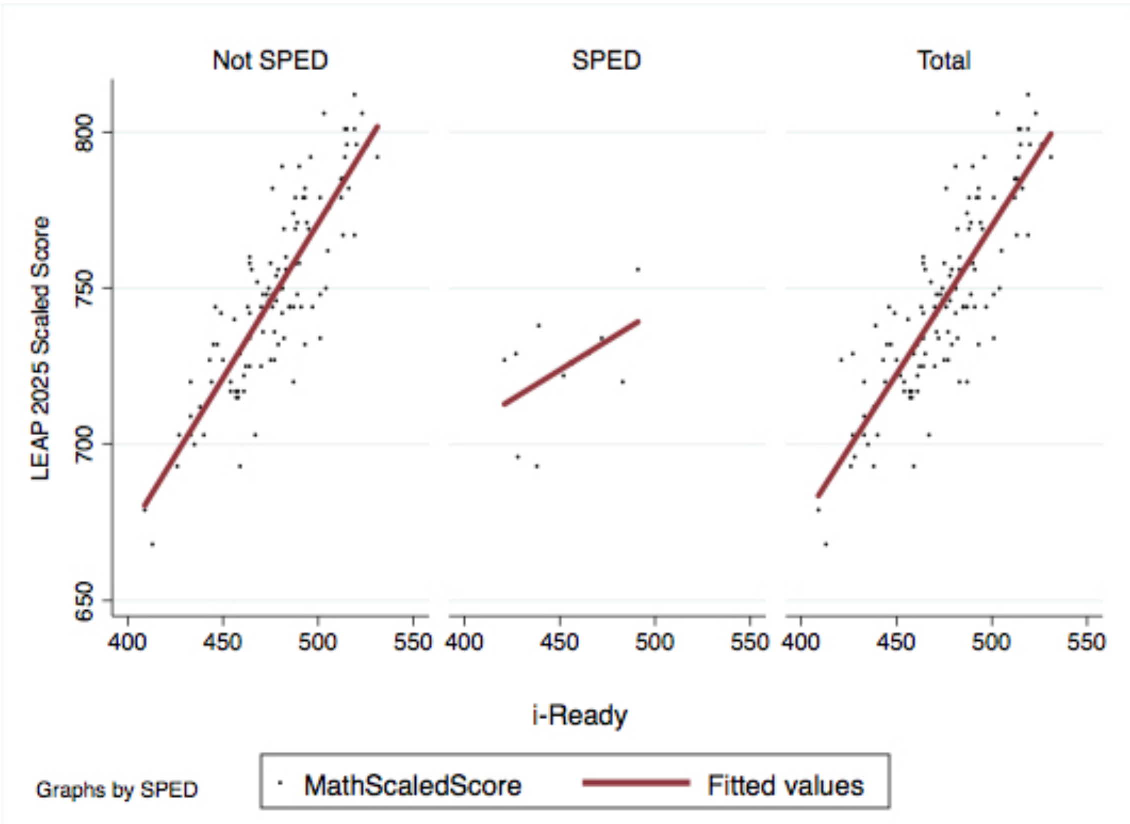
Eighth Grade

Concurrent Validity of Spring i-Ready



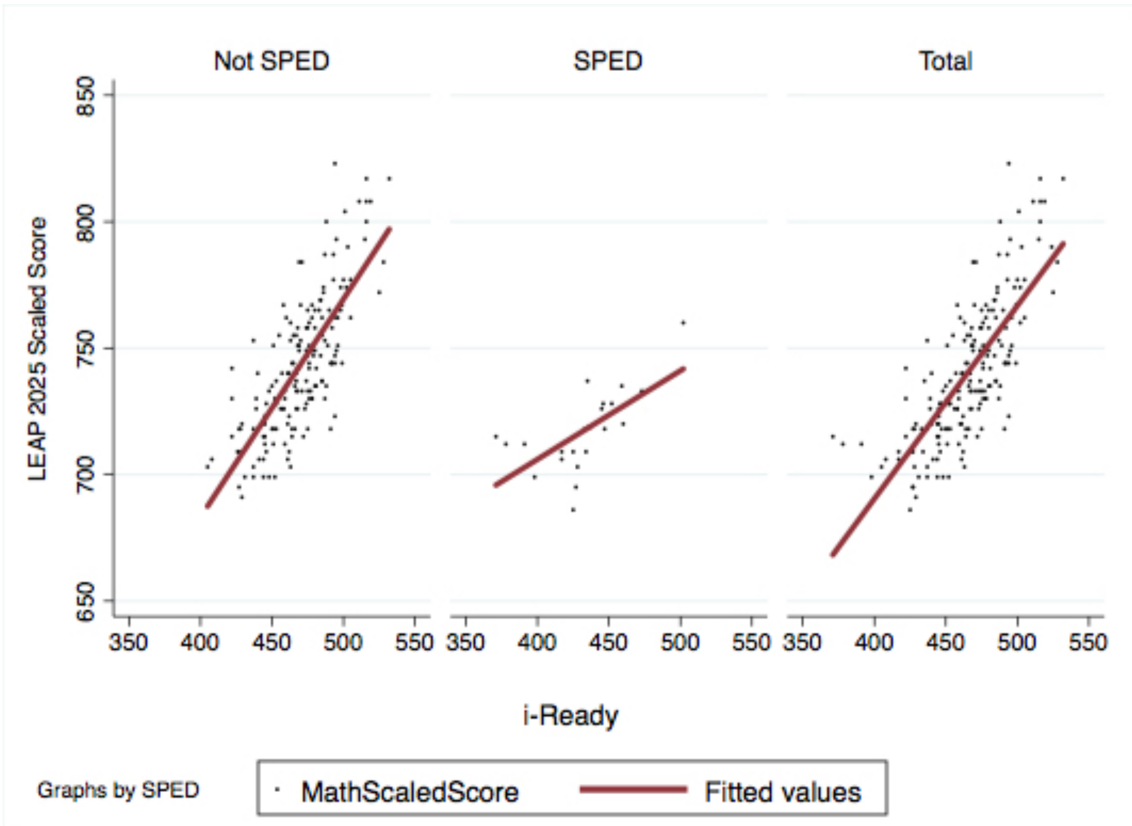
Fourth Grade

Concurrent Validity of Spring i-Ready



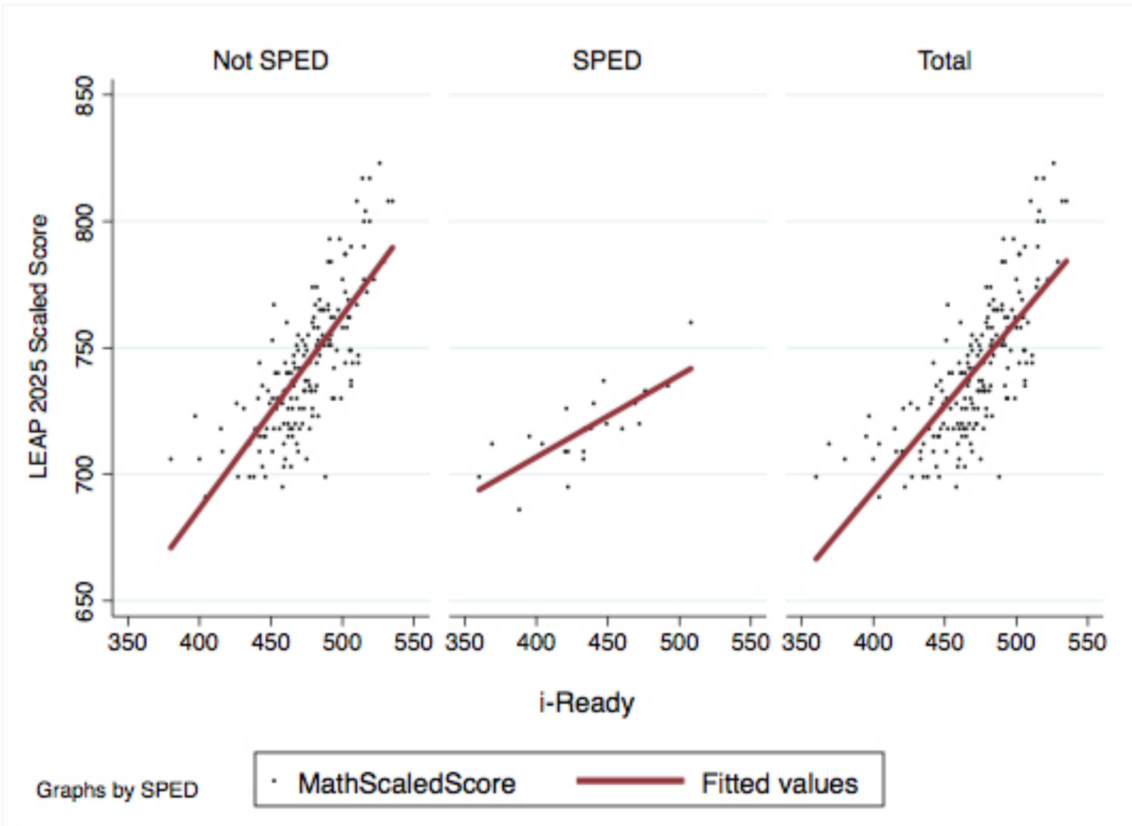
Fifth Grade

Predictive Validity of Fall i-Ready



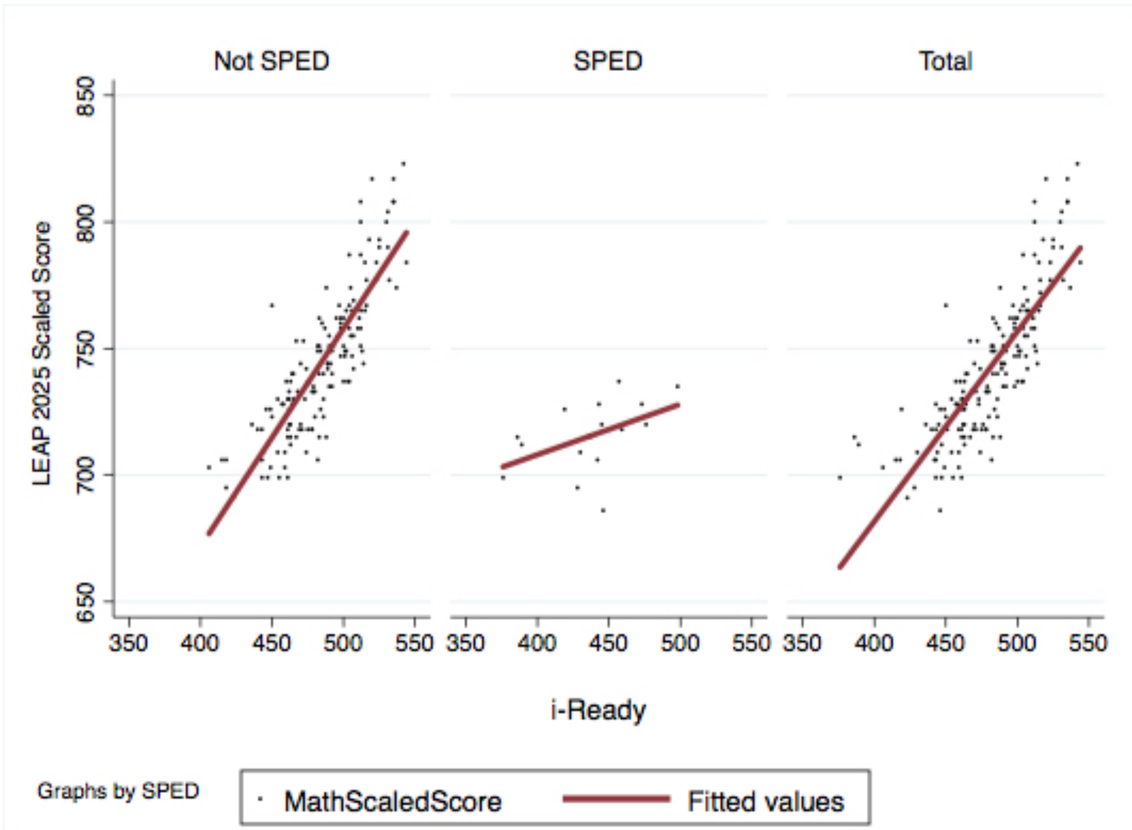
Fifth Grade

Predictive Validity of Winter i-Ready



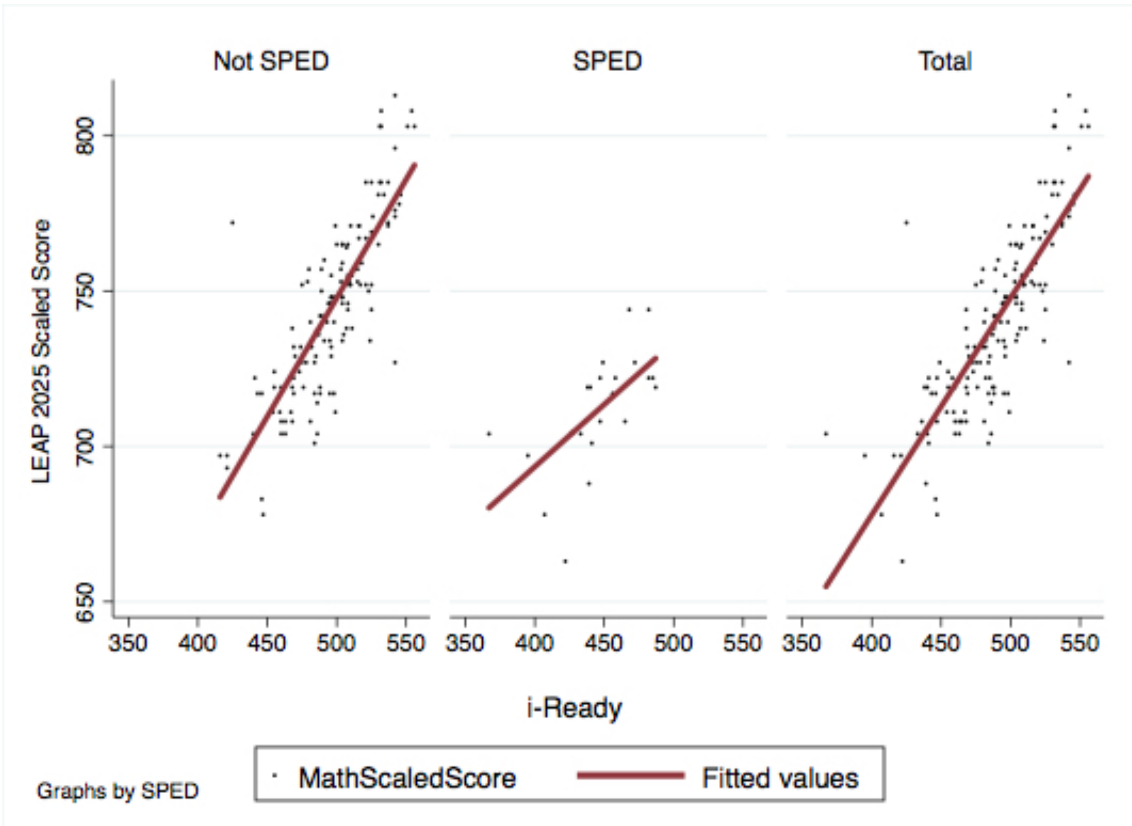
Fifth Grade

Concurrent Validity of Spring i-Ready



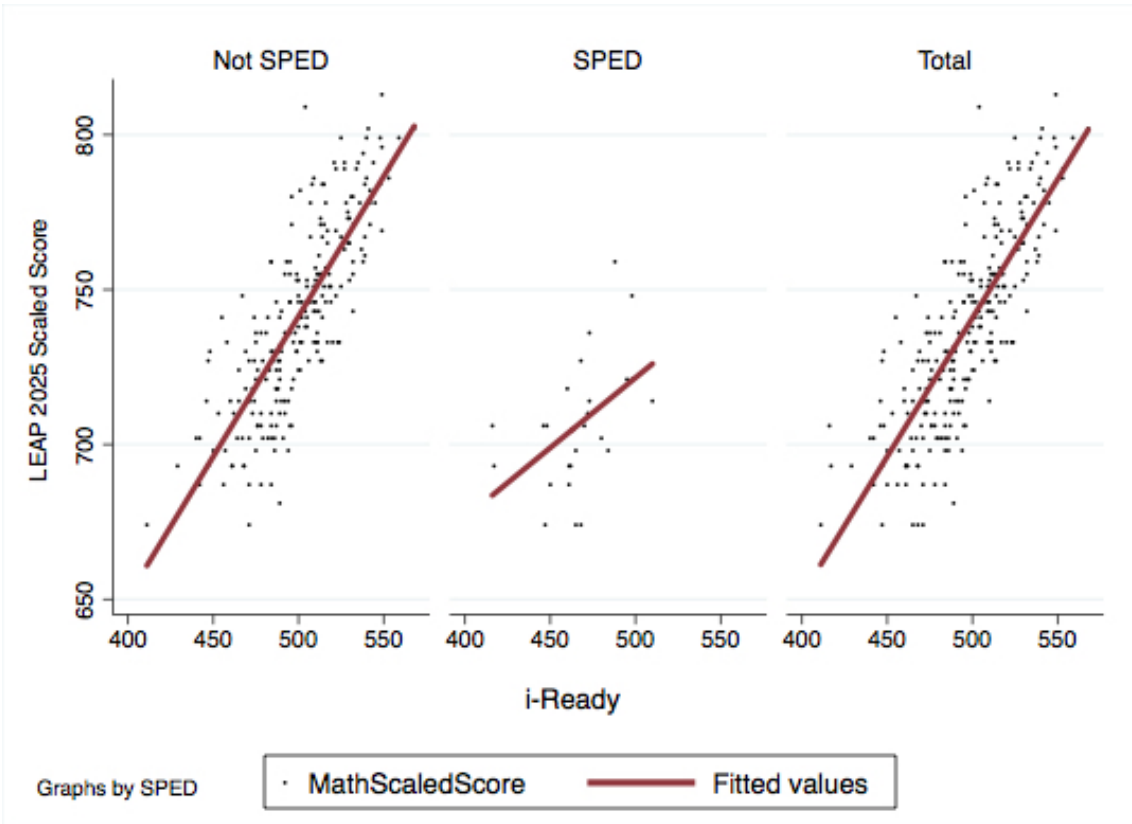
Sixth Grade

Concurrent Validity of Spring i-Ready



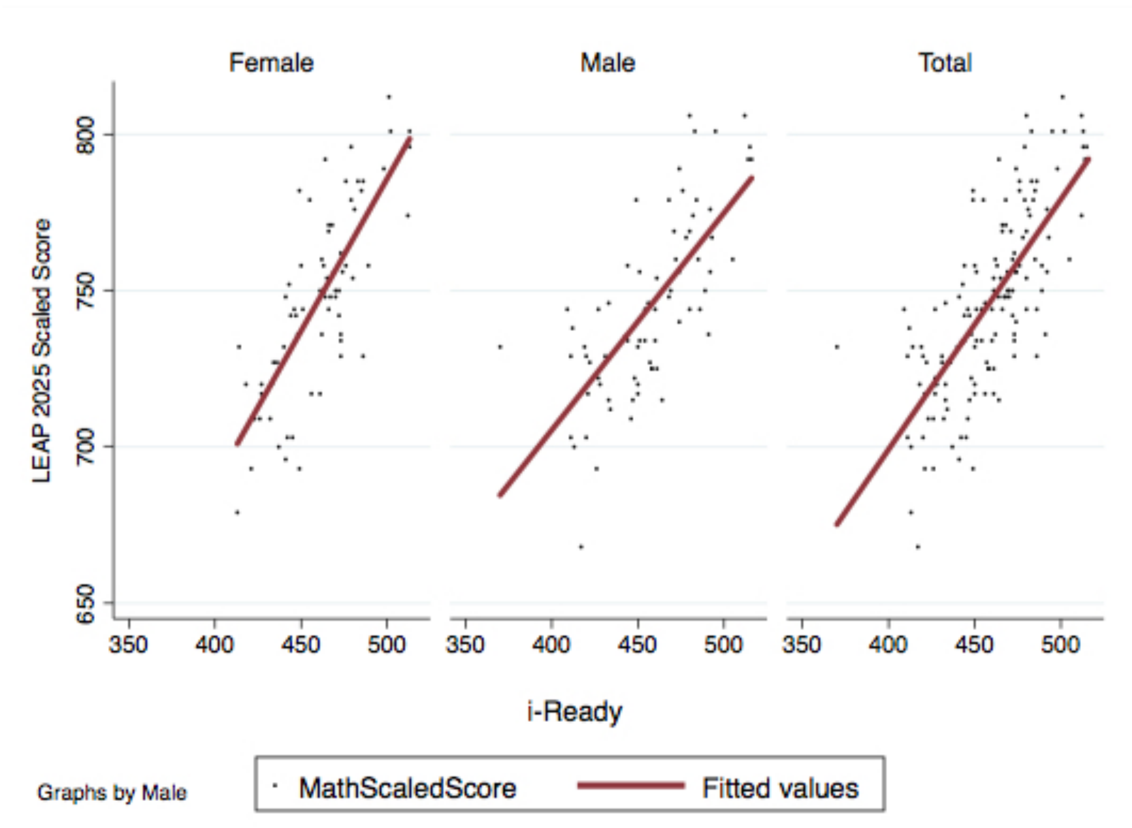
Eighth Grade

Predictive Validity of Fall i-Ready



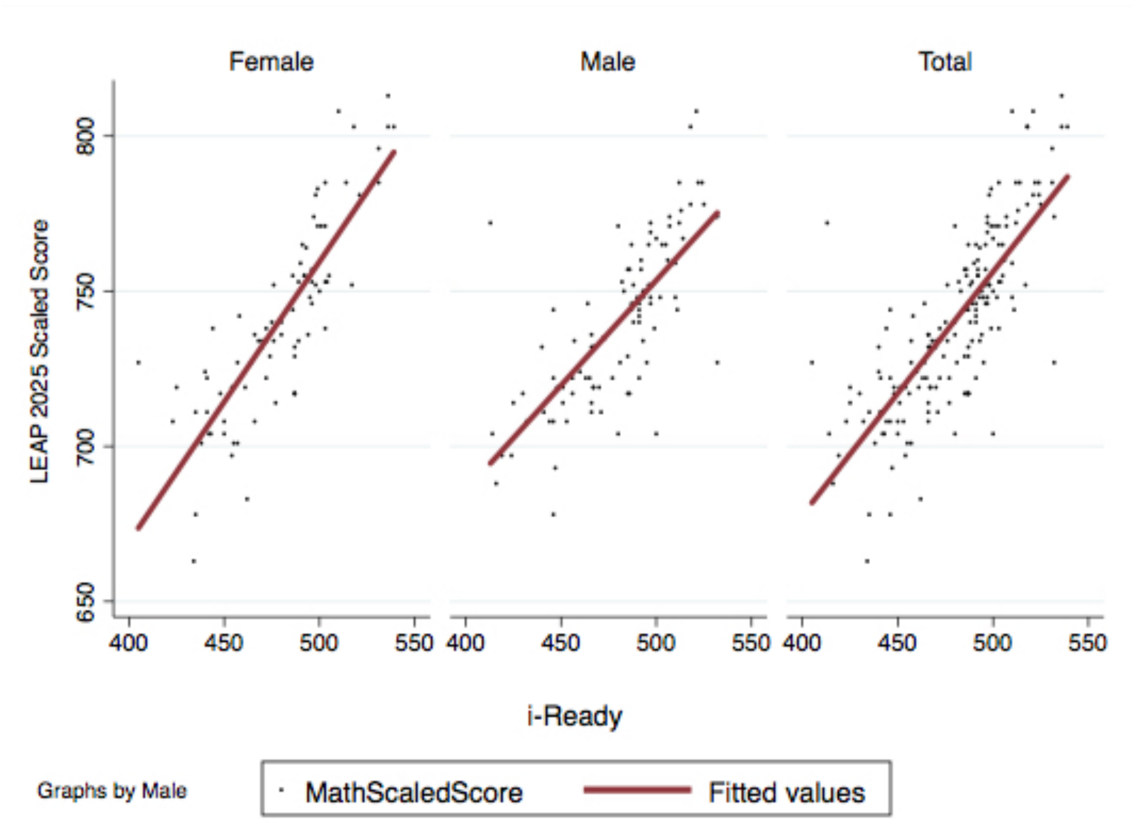
Fourth Grade

Predictive Validity of Winter i-Ready



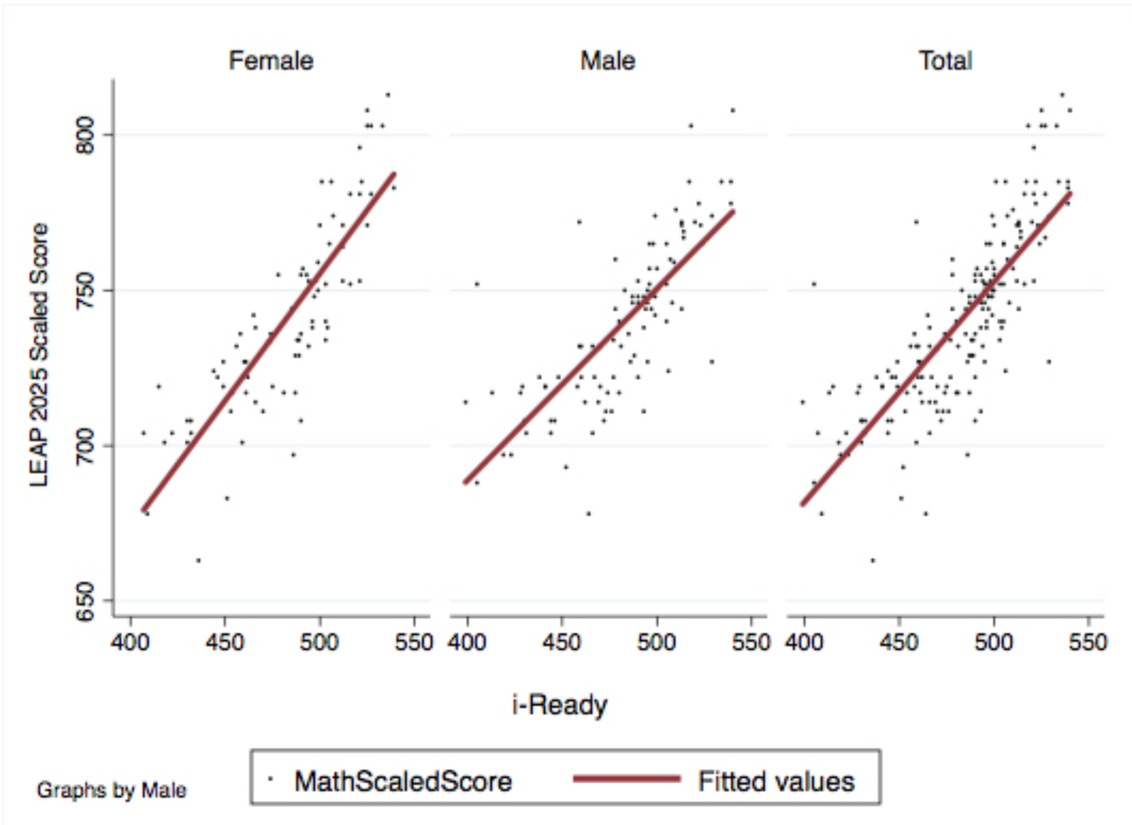
Sixth Grade

Predictive Validity of Fall i-Ready



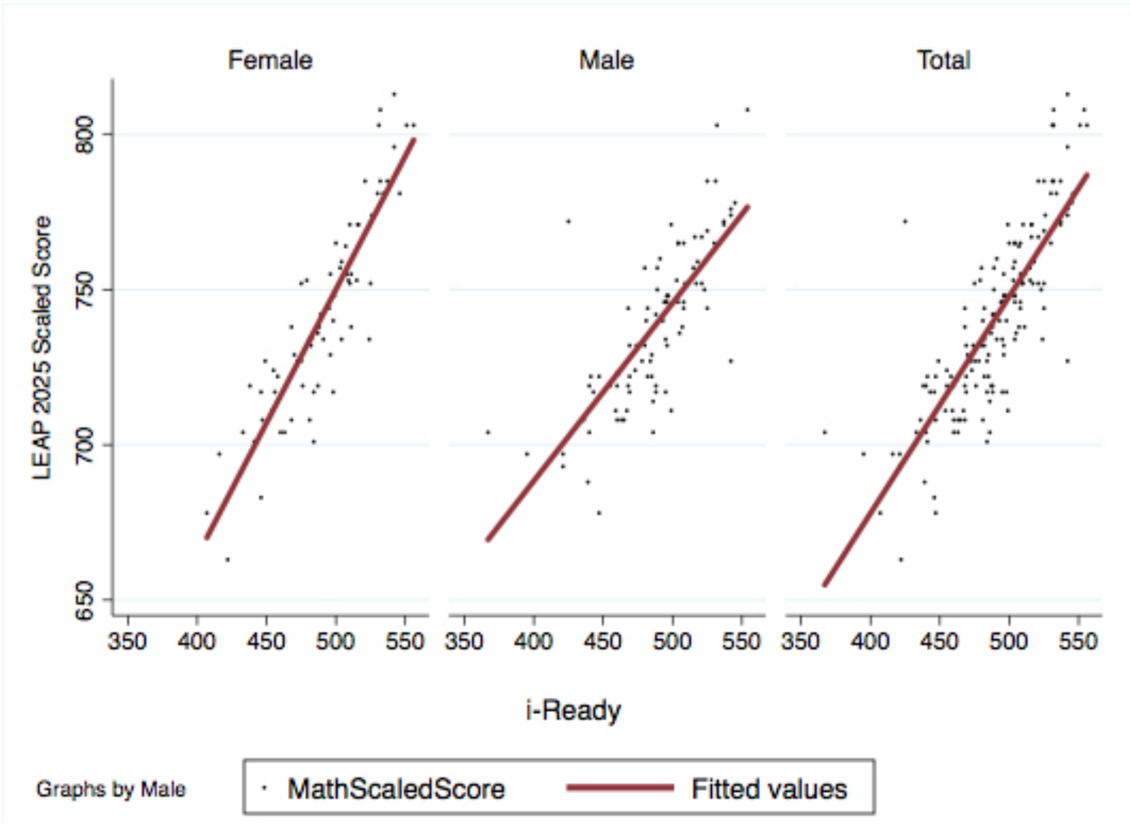
Sixth Grade

Predictive Validity of Winter i-Ready



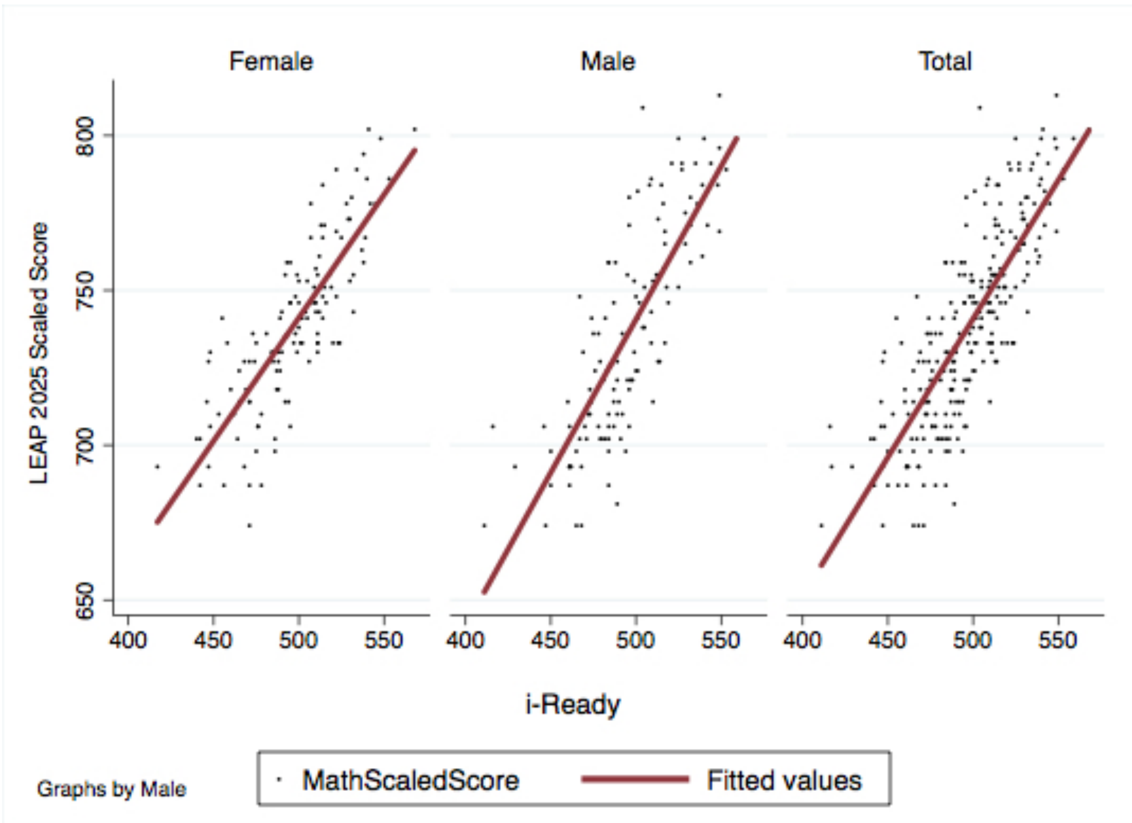
Sixth Grade

Concurrent Validity of Spring i-Ready



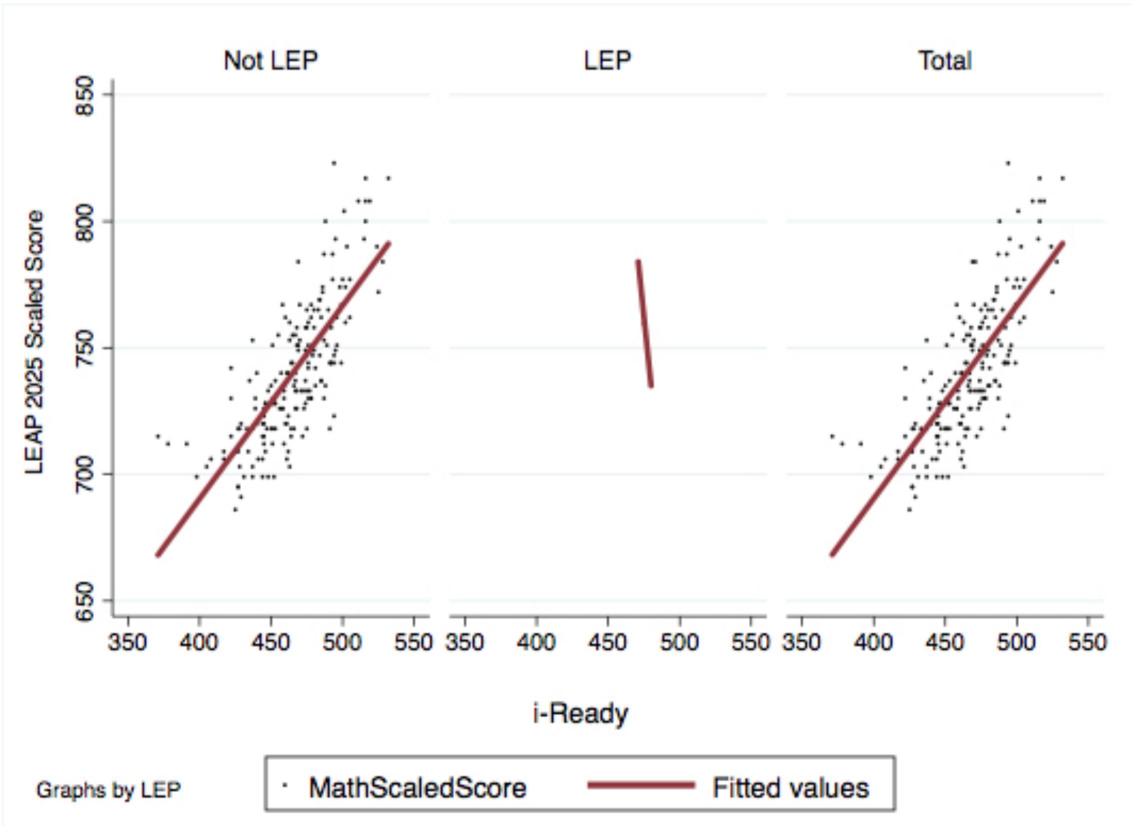
Eighth Grade

Predictive Validity of Fall i-Ready



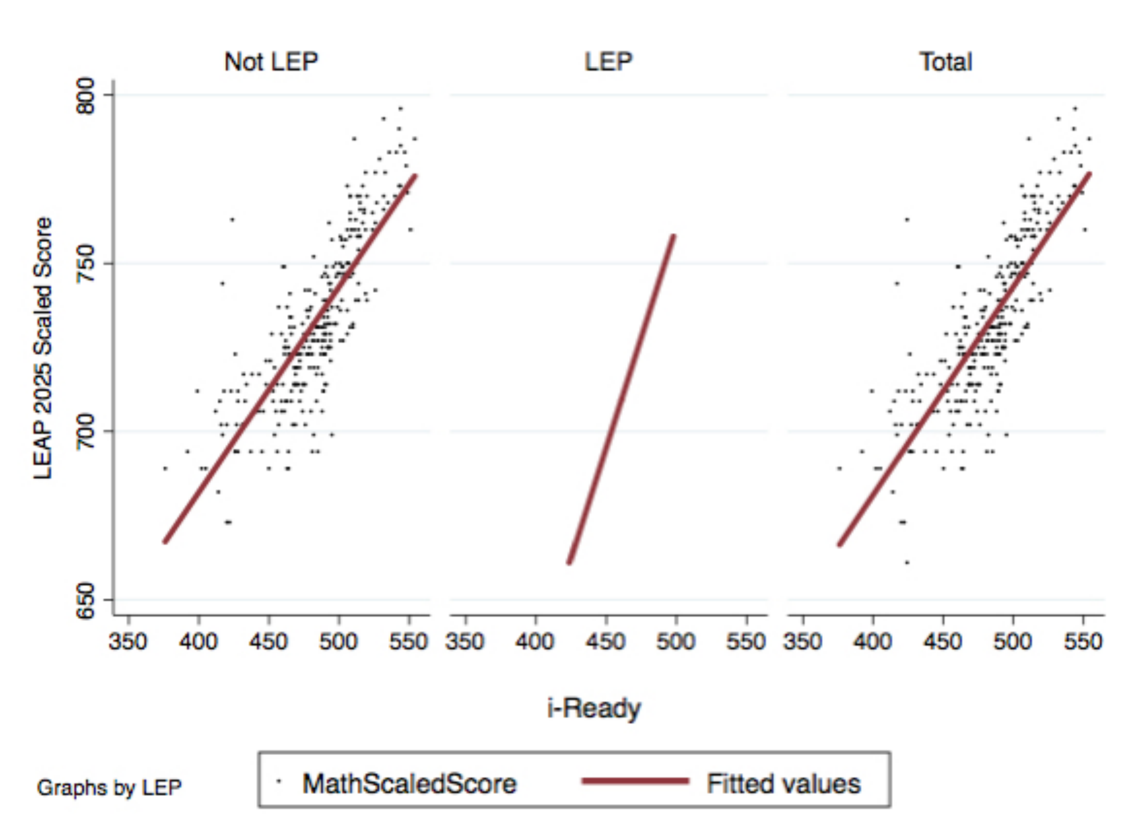
Fifth Grade

Predictive Validity of Fall i-Ready



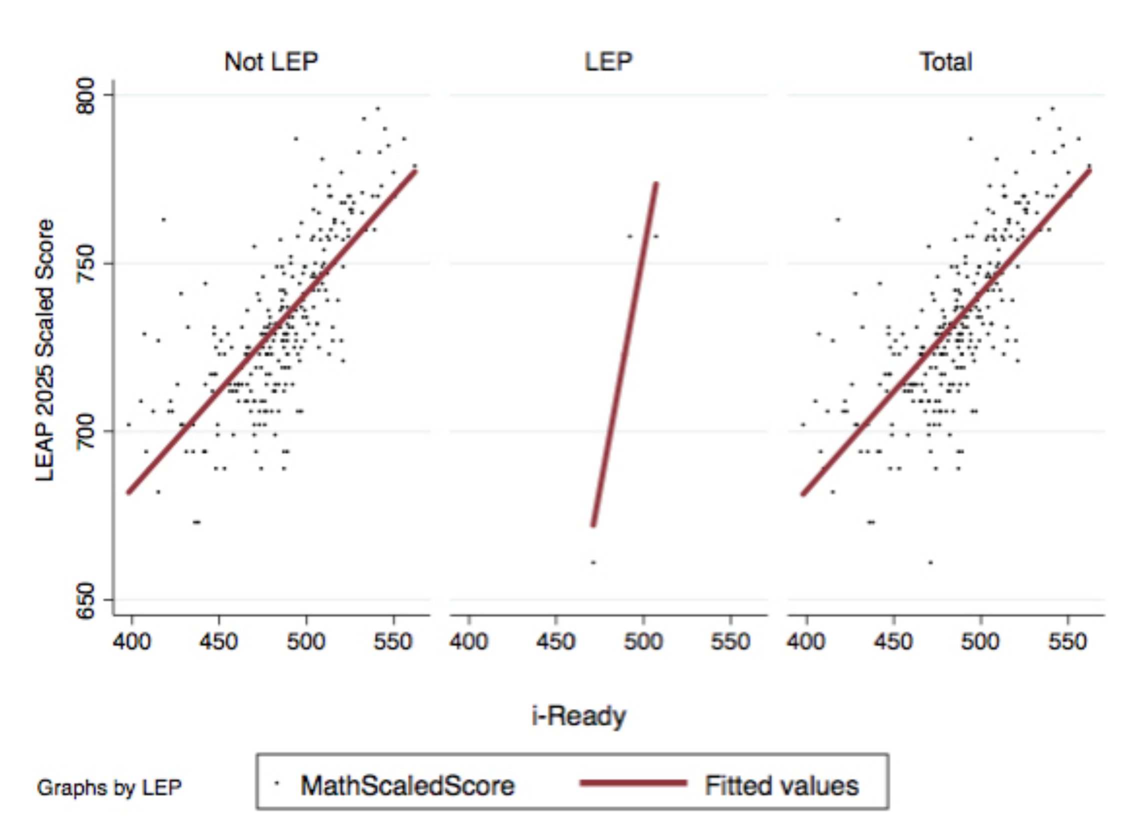
Seventh Grade

Predictive Validity of Fall i-Ready



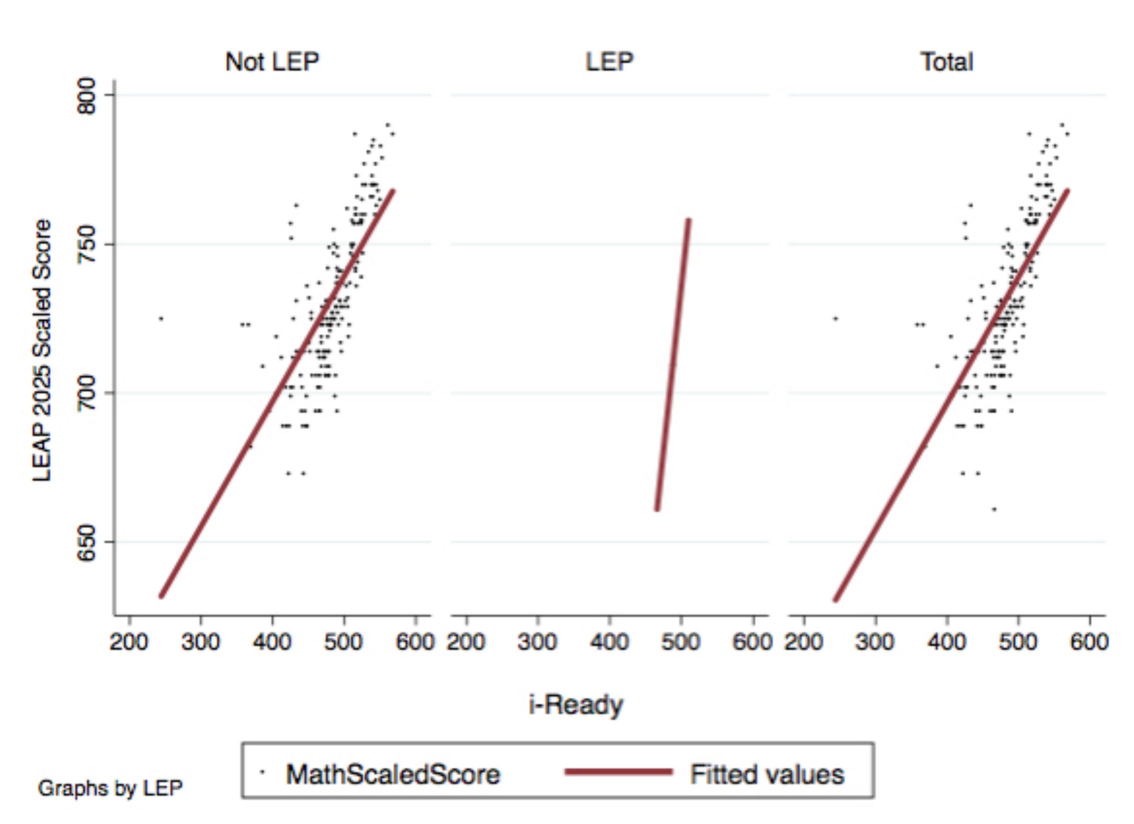
Seventh Grade

Predictive Validity of Winter i-Ready



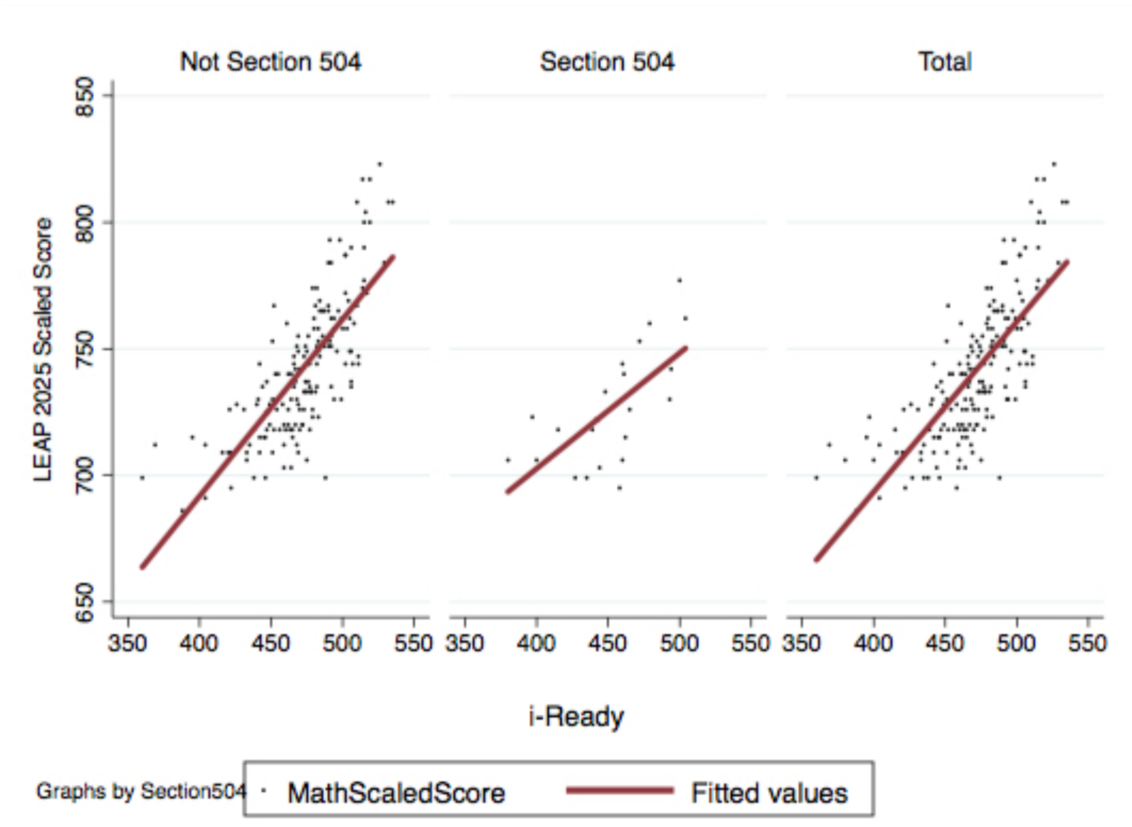
Seventh Grade

Concurrent Validity of i-Ready



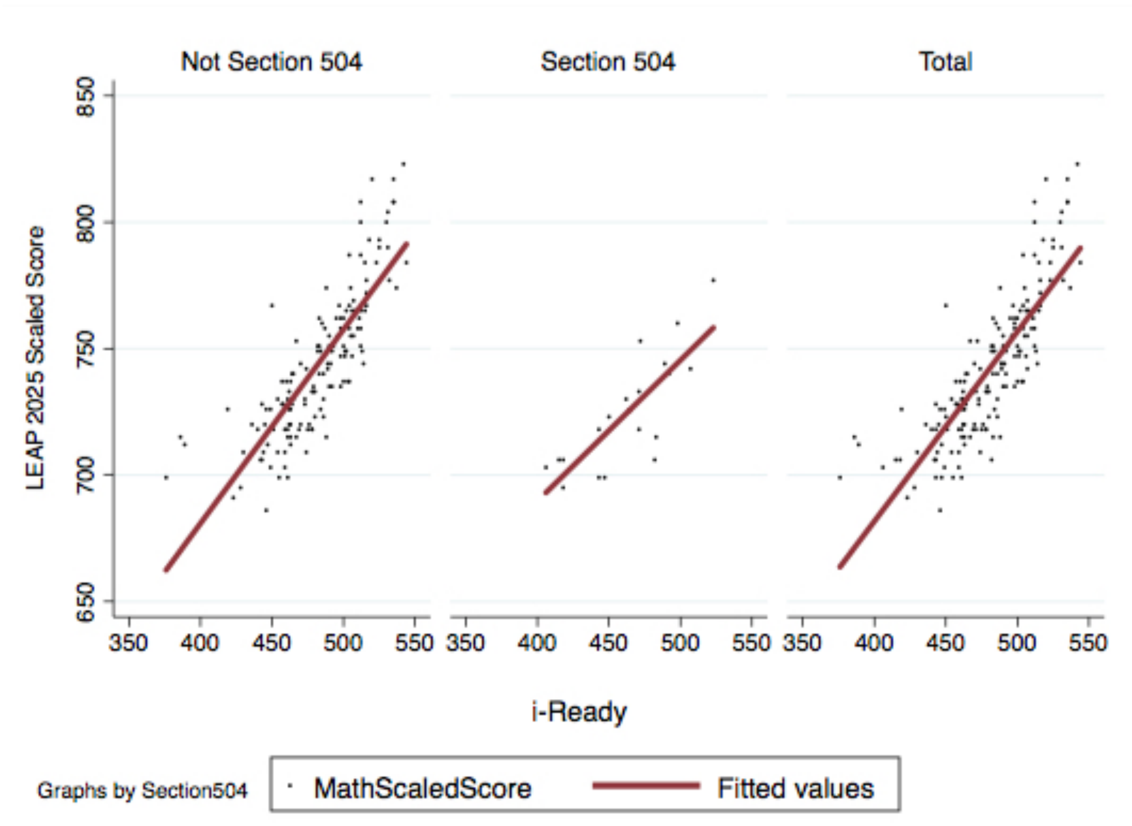
Fifth Grade

Predictive Validity of Winter i-Ready



Fifth Grade

Concurrent Validity of Spring i-Ready



Eighth Grade

Predictive Validity of Winter i-Ready