Developing and Investigating the Promise of Early Measurement Screeners

Authors

Ben Clarke, Marah Sutherland, Christian T. Doabler, Taylor Lesner, David Fainstein, Kelsey Nolan, Britt Landis & Derek Kosty

Publication History

Received 18 Sep 2020, Accepted 15 Apr 2021, Published online: 13 Jul 2021

Full Reference

Clarke, B., Sutherland, M., Doabler, C. T., Lesner, T., Fainstein, D., Nolan, K., Landis, B., & Kosty, D. (2021). Developing and investigating the promise of early measurement screeners. *School Psychology Review*, Advance online publication. doi: 10.1080/2372966X.2021.1919493

Funding Source

Research reported here was supported by the National Science Foundation through Grants **1503161** and **2010550**. Any opinions, findings, and conclusions or recommendations expressed in these materials are those of the author(s) and do not necessarily reflect the views of National Science Foundation.

Abstract

This study investigated the technical characteristics of four early measurement curriculum-based measures (EM-CBMs) designed to assess concepts related to linear measurement and iteration. The sample consisted of 221 first grade students. Data were collected at two time points approximately 10 weeks apart. Reliability and concurrent and predictive validity correlations were in the low to moderate range. We discuss study results related to screening for risk status including limitations to the current work and future directions for research.

Brief Impact Statement

The importance of measurement in mathematics development is garnering increased attention. Exploring measures to screen students for risk status is critical to enable schools to allocate resources to students in need of intervention services. The findings in this manuscript represent a first exploratory attempt to develop screening measures in the area of measurement.

Keywords: measurement, mathematics, screening

Developing and Investigating the Promise of Early Measurement Screeners

Despite the importance of mathematics to long-term academic success (Morgan et al., 2009), significant numbers of students fail to demonstrate an understanding of basic mathematics (NAEP, 2017). To address this problem, calls have been made to focus on increasing students' understanding of number and number systems (NMAP, 2008). As such, in the early elementary grades researchers have developed and evaluated interventions targeting whole number understanding (e.g. Clarke et al., 2014; Dyson et al., 2013). While those efforts have been successful, advances in whole number intervention have been at the neglect of other critical areas of mathematics.

One such area that has been drastically understudied is the domain of measurement. The Measurement and Data strand of the Common Core State Standards appears as early as kindergarten, where students are expected to learn how to describe measurable attributes of an object (e.g., height, weight) and directly compare two objects based on that attribute (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). This area of mathematics is conceptualized as distinctive from other areas such as whole number understanding in several key ways (National Research Council, 2009). Although whole number skills underlie the application of measurement, understanding foundational measurement concepts requires mathematical proficiency specific to this domain. For example, a student measuring the length of an object in units is subdividing a continuous quantity, thus making it "countable" and necessitating whole number skills (Clements & Sarama, 2004). However, this process also requires students to (a) understand length as an attribute that spans a fixed distance and can be subdivided, (b) understand that length can be represented as a smaller unit that is iterated along the full length of the object, and (c) understand that the origin when measuring an objects' length is 0, with each additional unit representing "1" more unit (National Research Council, 2009). These concepts within measurement are often challenging for young students to grasp, yet are essential for building foundations for advanced mathematics skills ranging from understanding units in the whole number system (ten ones as a unit of ten) to fraction understanding (CCSS Writing Team, 2018; Siegler et al., 2010; Van de Walle, 2019). For example, when considering the teaching of fractions, research findings point to the effectiveness of the measurement model to teach fraction understanding, where fractions are represented as lengths on a number line, over the more commonly-used part-whole interpretation (Fuchs et al., 2013). Early practice with units and rulers can also help students more easily access the measurement model of fractions and understand fractions as units (e.g., "2/5" is two copies of the unit "1/5"; National Research Council, 2009).

Measurement also provides an ideal platform for students to apply skills within number and operations to everyday, real-world situations such as measuring and estimating time, distance, weight, and speed (Clements & Sarama, 2004). The ability to understand and engage with concepts of measurement underlies STEM (Science Technology Engineering Mathematics) fields (Beckmann, 2008; Billstein et al., 2004; Shaughnessy, 2007) and enables students to engage in the complex statistical investigations (Confrey et al., 2012) necessary for success in those fields. Longitudinal and experimental studies demonstrate the importance of early measurement understanding to overall mathematics development. Early math achievement, including understanding of measurement, is highly predictive of later mathematics (Duncan et al., 2007) and persistent mathematics difficulty (Morgan et al., 2014). Specific measurement tasks are predictive to a broad array of mathematics concepts and skills (Schneider et al., 2018). In addition, learning of early measurement facilitates (Fyre et al., 2013) and impacts learning of whole number concept (Vasilyeva et al., 2020). Given the important ties between early proficiency in measurement and later educational outcomes (Claessens & Engel, 2013), it is critical that students learn foundational measurement concepts early on in their educational experience. Prominent calls have been made to expand the focus of early mathematics instruction beyond whole number to include critical concepts related to other aspects of mathematics including measurement (Frey et al., 2009). The following section, we provide an overview of these skills in first grade and their link to other central ideas in measurement and other domains of mathematics.

Linear Measurement

Linear measurement is the process of quantifying the distance between two end points of an object (Cross et al., 2009; Reys et al., 2014). One foundational skill that students must develop in linear measurement, listed in the Common Core State Standards (2010) for Grade 1, is comparing the length of objects to determine which object is longer or shorter. Students begin working toward this skill in kindergarten by comparing sets of objects using visual strategies to determine which set has more or fewer objects (Clements & Sarama, 2004). In first grade, students are expected to extend these strategies to comparisons of length within linear measurement. To accomplish this, they must rely on known concepts of linear measurement, including understanding that lengths span fixed distances, and that as objects are moved or rotated their lengths stay consistent (Stephan & Clements, 2003). In addition, students must develop understanding of key measurement vocabulary, such as “longer” and “shorter” to specify these comparisons. The skill of comparing object lengths extends students’ early comparison skills from visually comparing sets of discrete objects to comparing continuous lengths (CCSS, 2010). Recognizing length as an object attribute is also a prerequisite skill to understanding that length may be subdivided into equal-sized units. A second important concept that students must understand within linear measurement is transitivity, where a third object is used to compare the lengths of two other objects. This skill is also an objective within the CCSS (2010) for Grade 1. For example, a student comparing the lengths of two immovable objects may use his or her pencil as a referent by holding or marking the length of one object on the pencil, and then lining up the pencil with the other object to determine which is object is longer. Reasoning transitively is considered to be a prerequisite skill to formal measurement (Boulton-Lewis, 1987; Kamii & Clark, 1997), as students draw upon this concept when using a ruler to compare the length of two objects.

Iteration of Length Units

Students must develop conceptual understanding of measurement as covering space (Cross et al., 2009). With this understanding comes the idea that an object's length can be subdivided into equal-sized units, and also measured by taking a unit and placing it end-to-end along the length of an object (also known as *unit iteration*; Cross et al., 2009; Stephan & Clements, 2003). Conceptual understanding of unit iteration includes understanding that when measuring with units, each unit must be the same size and placement of units should not include gaps or overlaps (Clements & Sarama, 2004). These principles of unit iteration are important for students to master as they underlie conceptual understanding of formal measurement tools such as a ruler, and are listed as a Grade 1 standard in the CCSS (2010). For example, students must understand that the use of numerical values to express the length of an object is the same as iterating a unit along the length of an object a given number of times. Conceptual understanding of unit iteration and the meaning of numerals on a ruler also sets up students for success when they encounter number line representations of fractions and fraction computation in the later elementary grades (Siegler et al., 2010).

Assessment of Early Measurement Skills

In Response to Intervention or Multi-Tiered Systems of Support (MTSS) models, screening functions as a first step into the provision of more intensive services including small group instruction (Albers & Kettler, 2014; Authors et al., 2015), providing necessary supports for struggling learners. The most commonly used tools for screening are curriculum-based measures (CBM). CBMs were first developed to enable the monitoring of student growth over time (Deno & Mirkin, 1977). The goal of using CBMs for progress monitoring guided their design as simple and efficient assessments to administer, as well as being reliable and valid (Deno, 1985). These design parameters also enable CBM to be used for screening within MTSS as schools administer CBMs (~10 minutes or less/ student) to determine which students will need supplemental support (Fuchs et al., 2007; Gersten et al., 2012). CBMs also addressed shortcomings with single skill or mastery measurement including determining instructional skill hierarchies, retention and generalization of skills, shifts in measurement focus, and a lack of technical adequacy data (Fuchs & Deno, 1991). Fuchs (2004) proposed a three-stage research approach when developing and validating CBMs for eventual use as progress monitoring assessments. In the first stage, technical features of a static score of the measure are explored. This includes acquiring and analyzing psychometric data (e.g., inter-rater reliability, criterion validity) from an assessment at a single point in time. Work in this stage can serve as an initial exploration of the measures potential use in screening by examining predictive validity between CBM measures and criterion measures administered at later time points in the study or school year. Because stage 1 is the initial step in CBM research, it is imperative that a range of tasks and formats are developed and explored to determine which approach may best measure the underlying construct of interest. For example, while the use of oral reading fluency is now well established as the standard approach to measure reading, several measures were developed and explored as potential CBM reading measures. Tasks and formats included reading isolated word lists, cloze tasks, providing word definitions, and reading words in context (Deno et al., 1982) with promising measures further investigated (Deno, 1985). In the second stage, technical features of slope are examined to gauge capacity of the measure to index growth over time. The sensitivity and reliability of the CBM slope must be sufficient to describe a rate of skill acquisition (e.g., words read in one minute during a passage reading fluency task increases a predictable amount of words per week). In the last stage, the utility of the measure to improved instructional decision making is investigated. The third stage is vital because data utilization is the ultimate goal of CBMs (Lembke et al., 2016). Taken together, CBMs are deemed sufficient for progress monitoring if their technical adequacy (stage 1), linkage between increasing CBM scores and academic skill development (stage 2), and utility for instructional decision-making (stage 3) are empirically validated.

Since the original development of CBM, researchers have developed variations of CBM-like measures to screen for risk and monitor progress in mathematics (Foegen et al., 2007). One CBM approach is to focus on robust indicators, or general-outcome measures, of performance within an academic domain. Robust indicator measures are designed to be generalizable across contexts and grade levels, sampling from one task that strongly correlates with an important math construct (Foegen et al., 2007; Fuchs, 2004). Corresponding to the focus on whole number understanding and whole number interventions, typical screening systems for the early elementary grades have focused on specific aspects of number understanding with commonly used measures designed to tap into concepts related to understanding of magnitudes and strategic counting (Fuchs et al., 2007; Gersten et al., 2012). Calls have been made to expand screening efforts into other critical domains of mathematics (Methe et al., 2011). Although there have been sustained and successful attempts to develop measures in more advanced domains of mathematics such as algebra (e.g. Foegen, 2008) little work has been done to develop and validate general outcome measures focused on areas outside of number in the early elementary grades with limited exceptions in the areas of geometry (e.g. VanDerHeyden et al., 2011).

The purpose of this study was to conduct a Stage 1 (Fuchs, 2004) initial exploration of potential general outcome measures in the area of early measurement with a focus on utility as screening instruments. We developed a set of four first grade measures intended to assess students' conceptual knowledge of measurement skills. The measures were designed according to CBM design principles (Deno, 1985), targeted foundational concepts of linear measurement and principles of iteration, and were designed to align with the Measurement and Data Analysis domain of the Common Core State Standards (2010) for first grade. Concepts were selected based on their importance to student understanding and growth in early measurement and their potential capacity to serve as a general outcome measures. While Stage 1 research focuses on the development of measures and examining data collected at one point in time, this study includes two data time points allowing an initial examination of the capacity of the measures to model growth over time (Stage 2) and investigate additional psychometric properties. Research questions for the study were as follows:

1. What are the descriptive statistics (mean, range, distribution) of each experimental measure?

2. What are the psychometric (test-retest and alternate form) reliability properties of each experimental measure?

3. What are the psychometric (concurrent and predictive) validity properties of each experimental measure?

Method

This study analyzed data collected during a federally-funded design and development project (Doabler et al., 2015) to test the promise of an intervention program (Precision Math) focused on concepts of measurement and data analysis (Doabler et al., 2019). The study was conducted in a mid-size school district of approximately 18,000 students during the 2017-2018 school year. In the winter, data were collected at two time points approximately 10 weeks apart. The design of the study allowed for examining psychometric properties at one point in time (at each of the two data collection time points) along with investigating additional research questions including test-retest, predictive validity, and capacity to model growth afforded by the second data collection time point.

Participants

Schools and Students. Participants included 223 first grade students in ten first grade classrooms, within five elementary schools. Demographic data was available for 221 of the 223 students. Of these students, 85% were White, 10% were Hispanic, 3% were more than one race, 2% were Asian, and <1% were African American. Approximately 5% of participating students received special education services, and 53% were female.

Measures

Early Measurement Curriculum-based Measures (EM-CBM; Clarke et al., 2017). Four EM-CBM subtests (Form A) were developed to align with the Grade 1 CCSS (2010) –

Measurement domain. Members of the development team included two of the grant PI's with extensive experience in mathematics assessment and intervention in the areas of whole number, an experienced special education teacher who authored multiple early intervention curricula including an intervention program focused specifically on building understanding of measurement concepts, and a small team of master's and doctoral graduate students in school psychology and special education programs. As part of the initial development process, the team met weekly. One challenge in the development of the EM-CBMs was determining how to best assess the critical concepts of linear measurement. The areas of length measurement and unit iteration were selected as key areas given their direct alignment with the Grade 1 CCSS (2010) for length measurement. Members of the development team created first drafts of the measures which were then revised based on feedback from other team members including specific feedback on assessment items and directions. As part of a summer academic intervention clinic, versions were tested informally and then underwent further revision (e.g. modifying sample items and directions to increase clarity) resulting in the final measures used as part of this study.

Each subtest consisted of 30 items and was timed so that students had one to two minutes to answer as many items as possible (two minutes only for the Length-Measurement subtest). The development team decided to include 30 assessment items per subtest by factoring in the number of items a student could reasonably complete in the time allotment for each subtest. The 30 items were designed to be approximately the same difficulty, though natural variation in items was expected. During test administration, assessors prompted students to attempt the next item after 6 seconds, with the exception of the Length-Measurement subtest where there was no formal rule due to students using a tool to measure. For Length-Measurement, assessors were instructed to use their professional judgment and encourage the student to move on if they were not actively working on an item. All subtests included one to two practice items with verbal feedback from the assessor confirming correct responses or providing scripted corrective feedback. Alternate forms (Form B) were developed for each subtest, consisting of the same items in a pre-randomized order.

The EM-CBMs were printed in binders with two to four items per page (see Figure 1). Assessors turned the pages for students and scored student responses on an iPad-based Qualtrics survey with a built-in timer. Subtest scores were computed using the total number of items correct. Incorrect items were not counted against students. All subtests had a rule where the assessment was discontinued if a student missed five consecutive items. Additionally, assessors were trained to administer verbal prompts such as encouraging students to "Pick just one" if they pointed to multiple items. The full measures, administration directions, and scoring rules are available from the first author. **Length-Comparison.** Each test item included a picture of three objects of varying heights, positioned on the same plane. Students were instructed to point to the shortest object. This subtest was intended to assess students' skills in comparing object lengths (CCSS.Math.Content.1.MDA.1). **Length-Measurement.** Each test item included a picture of two identical objects, with one object manipulated to be slightly longer than the other. Assessors instructed students to point to the shorter object, and students were given a base ten rod to help them measure. This subtest was intended to assess students' transitive reasoning (CCSS.Math.Content.1.MDA.1). **Iteration-Application.** Each assessment page contained a picture of a paperclip (approximately 1-inch long) and four horizontal lines of varying lengths. Students were instructed to say how long each line was in paper clip units (one, two, three, or four paper clips). This subtest was intended to assess students' application of iteration skills to measure lengths (CCSS.Math.Content.1.MDA.2). **Iteration-Conceptual.** Each test item included pictures of objects iterated along three horizontal lines of the same length. Two examples showed incorrect iteration (e.g., overlapping objects, objects spanning past the length of the line, etc.) and one example was correct. Students were given a hypothetical measurement scenario (e.g., "Johnny tried to measure a line using objects three times; Which time did he measure correctly?") and pointed to indicate their answer. This subtest was intended to assess students' conceptual understanding of iteration principles (CCSS.Math.Content.1.MDA.2).

Assessing Student Proficiency of Early Number Sense (ASPENS; Sopris; Clarke et al., 2011). The ASPENS is an individually administered, standardized test of early number sense. A composite score was derived from the three Grade 1 subtests assessing students' ability to (1) determine which of two numerals is greater (Magnitude Comparison), (2) identify the missing numeral in a string of three numerals (Missing Number), and (3) solve addition and subtraction facts crossing ten (Basic Arithmetic Facts and Base 10). Test-retest reliability (Fall to Winter, Winter to Spring, Fall to Spring) ranges from .77 to .84, and concurrent (.63) and predictive validity (Fall to Spring, Winter to Spring) with the TerraNova-3 is .57 and .63 respectively.

EasyCBM Math (Alonzo et al., 2006**).** EasyCBM Math is a multiple-choice, online mathematics assessment assessing all domains of the CCSS-M (2010). In Grade 1, internal reliability of easyCBM ranged from .81 to .84 and split-half reliability ranged from .72 to .81. Concurrent validity of the spring benchmark with the TerraNova-3 was .69 and predictive validity from the Fall and Winter to Spring TerraNova-3 scores was .60 and .70 respectively.

Procedures

Data was collected in the winter of first grade at Time 1 (T1) and approximately 10 weeks later at Time 2 (T2). The ASPENS was only administered at T1. EM-CBMs and EasyCBM Math measures were administered at T1 and T2. All measures were administered within 1-2 days of one other at each time point, and all the EM-CBMs were administered in the same testing session at each time point. At both time points, the EM-CBMs were administered in the same order for each student (Length-Comparison, Iteration-Application, Iteration-Conceptual, and Length-Measurement). The randomized Form B subtest was administered following the four Form A subtests. All data was collected by research personnel. Data collectors had assorted backgrounds though most had worked as data collectors previously on other large-scale mathematics projects. All data collectors completed individual reliability checkouts on the EM-CBMs and the ASPENS, meeting a standard of 90% item level inter-rater reliability or higher prior to administering measures in schools. The EM-CBM training lasted approximately three hours with opportunities for practice and orientation to scoring using iPads. Data collectors completed a second checkout in the field at the start of data collection to the same 90% criterion level.

Statistical Analyses

Univariate descriptive statistics were examined for all study measures. Pearson's *r* bivariate correlations were generated to examine reliability and relationships among the EM-CBMs and other measures at T1 and T2. For reliability correlations, 95% confidence intervals were calculated based on recommended formulas (Snedecor & Cochran, 1980). Bland-Altman plots were generated to examine proportional bias of the association between a student's estimated true score and the difference between scores on Forms A and B (Bland & Altman, 1986). Proportional bias is present if the difference between a student's score on Forms A and B increases or decreases in proportion to the student's estimated true score. Follow-up linear regression analyses were used to confirm whether bias was statistically significant.

Missing Data. At T1, assessment data was missing for four students due to students changing schools or classrooms (*n* = 3) or unavailable due to special education classes (*n* = 1). A total of 213 students completed assessments at T2, with reasons for missing data including changing schools or classrooms (*n* = 8), or being absent after multiple testing attempts (*n* = 2). Thus, the correlational data presented in the study included *n* = 219 students for T1 assessments and *n* = 213 for T2 assessments.

Results

Descriptive Statistics and Distributions of Measures. Distributions of measures fell within the recommended bounds for normality with skew and kurtosis between -2.00 to 2.00 (Pedhazur, 1997), with the exception of the kurtosis values for Iteration-Conceptual T1 (kurtosis = 6.95), Length-Comparison T2 (kurtosis = 2.01), and EasyCBM T2 (kurtosis = 2.54). Descriptive statistics and bivariate correlations are displayed in Table 1. On average, students scored between nine and 17 items correct on the EM-CBMs across time points. Dependent-samples *t*-tests were used to compare EM-CBM scores from T1 to T2. Student scores significantly increased on Length-Comparison (T1: *M* = 16.07, *SD* = 4.86; T2: *M* = 17.07, *SD* = 4.24, *t*(211) = 3.05, *p* < .01, 95% CI [-1.65, -.35]), Iteration-Application (T1: *M* = 10.23, *SD* = 7.83; T2: *M* = 13.81, *SD* = 8.69, *t*(211) = -6.17, *p* < .001, 95% CI [-4.71, -2.43]), and Iteration-Conceptual (T1: *M* = 9.46, *SD* = 2.85; T2: *M* = 10.45, *SD* = 2.67, *t*(210) = -4.77, *p* < .001, 95% CI [-1.39, -.58]). Student scores decreased slightly on Length-Measurement though this result was not statistically significant (T1: *M* = 10.46, *SD* = 5.56; T2: *M* = 9.64, *SD* = 5.63, *t*(211) = 1.84, *p* = .067, 95% CI [-.06, 1.70]). *T*-tests were conducted to examine whether there was an increase in student scores from Form A to Form B at each testing administration. For T1, results are as follows: Length-Comparison (Form A: *M* = 15.51, *SD* = 5.37; Form B: *M* = 18.95, *SD* =

6.21; *t*(42) = -4.51, *p* < .001), Length-Measurement (Form A: *M* = 11.73, *SD* = 6.22; Form B: *M* = 11.04, *SD* = 5.83; *t*(44) = 0.94, *p* = .353), Iteration-Application (Form A: *M* = 7.98, *SD* = 7.25, Form B: *M* = 11.76, *SD* = 10.19; *t*(40) = -4.18, *p* < .001), Iteration-Conceptual (Form A: *M* = 9.78, *SD* = 2.70; Form B: *M* = 9.42, *SD* = 4.11; *t*(44) = .67, *p* = .507). For T2, results are as follows: Length-Comparison (Form A: *M* = 16.88, *SD* = 5.28; Form B: *M* = 20.25, *SD* = 6.08; *t*(31) = -7.08, *p* < .001), Length-Measurement (Form A: *M* = 11.03, *SD* = 6.18; Form B: *M* = 12.48, *SD* = 5.90; *t*(63) = -2.60, *p* = .012), Iteration-Application (Form A: *M* = 14.07, *SD* = 9.12; Form B: *M* = 17.69, *SD* = 9.15; *t*(57) = -4.30, *p* < .001) Iteration-Conceptual (Form A: *M* = 10.41, *SD* = 2.15; Form B: *M* = 11.65, *SD* = 3.16; *t*(53) = -2.95, *p* <.005).

Reliability. Test-retest reliabilities among the EM-CBMs from T1 to T2 (approximately 10 weeks) were low ranging from .33 to .48. Iteration-Application had the highest test-retest reliability (*r* = .48, 95% CI [.37, .58]), followed by Length-Comparison (*r* = .46, 95% CI [.34, .56]), Iteration-Conceptual (*r* = .42, 95% CI [.30, .52]), and Length-Measurement (*r* = .33, 95% CI [.20, .44]). Alternate form reliabilities were calculated at each time point (i.e. T1 to T1 and T2 to T2). Results were moderate to high with the exception of Iteration-Conceptual. These reliabilities were as follows: Length-Comparison (T1: *r* = .64, 95% CI [.41, .79]; T2: *r* = .90, 95% CI [.80, .95]), Length-Measurement (T1: *r* = .67, 95% CI [.47, .80]; T2: *r* = .73, 95% CI [.58, .82]), Iteration-Application (TI: *r* = .83, 95% CI [.70, .91]; T2: *r* = .75, 95% CI [.62, .85]), and Iteration-Conceptual (T1: *r* = .52, 95% CI [.26, .70]; T2: *r* = .37, 95% CI [.11, .58]). Bland-Altman plots and follow-up linear regression analyses comparing Form A and Form B revealed that proportional bias was present on the following subtests: Iteration-Application (T1) and Iteration-Conceptual (T1, T2). An example of the Bland-Altman plot with regression analyses is illustrated in Figure 2. Full proportional bias results are available from the first author.

Concurrent and Predictive Validity. With the exception of Length-Measurement, correlations among the EM-CBMs were significant at the *p* < .05 level and ranged from .22 to .41 at T1, and .20 to .44 at T2. Correlations between Length-Measurement and other EM-CBMs were not statistically significant except for at T1 with Length Comparison (*r* = .16, *p* < .05) and at T2 with Iteration Accuracy (*r* = .20, *p* < .001). For Length-Comparison, Iteration-Application and Iteration-Conceptual, concurrent validities at T1 with the ASPENS ranged from .25 to .43 (*p* < .001). At T2, concurrent validities with EasyCBM ranged from .24 to .48 (*p* < .001). For Length-Measurement, concurrent validities at T1 with the ASPENS and at T2 with EasyCBM were not significant (*r* = -.09 and *r* = -.02, respectively, *p* > .05). Predictive validities between the EM-CBMs at T1 and EasyCBM at T2 were low to moderate, ranging from .23 to .44 (*p* < .001), with the exception of Length-Measurement which did not significantly correlate with T2 EasyCBM (*r* = -.07, *p* > .05).

Discussion

Results from the current study were mixed. Descriptive statistics indicated gains for three of the four measures across time with the exception being Length-Measurement. Test-retest reliabilities were in the low range with stronger, yet still moderate, results for alternate form reliability. Proportional bias was evident for three of the alternate form comparisons (Iteration-Application at T1 and Iteration-Conceptual at T1 and T2). This is potentially due to practice effects with the tasks given that students completed one Form B subtest directly following the full battery of Form A subtests, and that these tasks were likely novel tasks that students had not encountered prior to testing. Additional support for the novelty factor, was evidenced by t-tests results showing significantly greater scores for Form B subtests on six of eight occasions. In addition, forms were not counterbalanced which may confounded results through an order effect.

Concurrent and predictive validities with the ASPENS and EasyCBM were in low to moderate range for three of the four measures with the exception again being Length-Measurement. Across the board, results were not as strong compared to what has been typically found with math CBM and in particular with early numeracy measures (Foegen et al., 2007; Gersten et al., 2012). Prior to discussing the results it should be noted that the low to moderate reliability evidence makes the drawing of conclusions regarding validity data tenuous. Thus, although the remainder of the discussion focuses on issues raised and implications from the current study, we temper our conclusions with the recognition that substantive subsequent work is needed in the area. Our discussion is intended to highlight considerations for the field to advance future research to build upon this initial exploration of early measurement CBM.

Given the research reported here was conducted as part of an initial stage 1 (Fuchs, 2004) effort to develop CBMs in a novel area, a primary question for consideration is whether we selected the right critical constructs (i.e. linear measurement and iteration) and if so were they operationalized in a manner that captured student understanding. There are multiple ways to assess any construct and future research should explore alternative ways to measure constructs like iteration. For example, in the current study we conceptualized linear measurement skills as a distinct construct, yet it is possible that some measurement skills such as iteration could also be captured using whole and rational number screening tasks. For example, students iterating a unit along the length of an object, as was the case in the Iteration-Application task, has overlap with number line tasks where students are determining the placement of numbers based on the iteration of single units on a number line (Booth & Siegler, 2008; Clarke et al., 2018; Sutherland et al., 2021). Rather than screening for measurement skills specifically, a more efficient approach might include tapping into these skills using tasks that also are inherently tied to whole or rational number understanding through number lines. Concurrent and predictive validity correlations were low to moderate and lower than those found when measuring constructs related to number sense in first grade (Gersten et al., 2012). In part, this may be due to the alignment of the EM-CBMs with the criterion measures used in the study (ASPENS and EasyCBM) which focus largely on number understanding. Future research should include a broader range criterion measures including criterions more directly aligned to the constructs of interest. Given that the work in this study was framed as initial stage 1 research, future research should explore additional constructs and task formats.

The original approach to this study was to select constructs and develop measures that would serve as general outcome measures for early measurement. However, based on results of the study and how measurement is currently taught in schools it is possible that assessment of measurement concepts may be better captured within a mastery measurement paradigm. While CBMs are typically designed as general outcome measures and are designed to reflect broader understanding and growth across an academic year, the topic of measurement is often taught in one unit or in a shortened or condensed period of time. As such, EM-CBMs may have higher utility when used as indicators of measurement skill mastery during the instructional period when those skills are taught. Recent research has investigated the utility of subskill mastery measures for instructional decision-making in mathematics, including screening, intervention planning, and progress monitoring decisions (VanDerHayden & Broussard, 2019). These measures are reliable with a single one- to two-minute probe and contain less random error than broader math CBMs, making them strong candidates for measuring performance variability both within and between students (e.g., Hintze et al., 2002). Timely administration of subskill mastery measures, in addition to general outcome CBMs, may be an effective approach to screening for risk in mathematics, given that overall mathematics proficiency requires successive mastery of numerous discrete skills (VanDerHayden et al., 2019). Screening practices which incorporate subskill mastery measures, used to predict risk and growth over a shorter time frame during which those skills are most instructionally relevant, may effectively address the limitations of widely-used general outcome CBMs in accurately predicting mathematics risk (VanDerHayden et al., 2017). Subskill mastery measures have been shown to outperform multiskill CBMs in accurately predicting lack of proficiency in grade-level mathematics by the end of the school year, with significantly lower rates of false negatives (VanDerHayden et al., 2017).

Subskill mastery measures may be more sensitive and instructionally useful in monitoring response to mathematics intervention. General outcome math CBMs have been criticized for their lack of sensitivity to detect small but meaningful changes in performance, which limits their utility for determining the short-term effectiveness of interventions (VanDerHayden & Broussard, 2019). Subskill mastery measures are more sensitive to growth, potentially facilitating decisions about data-based instructional adjustments (VanDerHayden & Burns, 2018) and providing an earlier indicator of progress that is strongly associated with growth in broader mathematics competency (VanDerHeyden et al., 2012). VanDerHayden and Burns (2018) argue that both subskill mastery measures and more distal measures, including general outcome CBMs, are important indicators of student progress – subskill measures assess progress toward short-term goals and mastery of specific instructional objectives, whereas general outcome measures assess broader growth toward competency in mathematics. Future research should further explore how the EM-CBMs investigated in this study and future EM-CBMs including those conceptualized and developed as subskill mastery measures may be used in conjunction with general outcome CBMs to identify and monitor the growth of students at risk for mathematics difficulties. Given findings related to educators engaging in data-based decision making (Bosch et al., 2017), research should examine and contrast measures from a broader perspective than an exclusive focus on psychometric properties. For example, a mastery measurement approach may have stronger technical properties but increases the complexity of data such that instructional decision making is not improved. As specified in the Fuchs (2004) stages of CBM research framework, questions of this nature would fit well within the focus of stage 3 on the utility of the measures to enhance student outcomes via instructional decision making.

The EM-CBMs were designed using CBM design parameters (Deno, 1985) so they could have potential use in screening and progress monitoring but much of the work in the area of measurement requires the hands on application of skills that may not be amenable to being assessed in a timed format – such as using manipulatives or a ruler to measure the length of an object or physically iterating a set unit to measure an object. Because the measures were timed and of limited duration, we decided to not use manipulatives for three of the four measures (the exception being Length-Measurement). However, it should be noted that the subtests where students did not use manipulatives had higher predictive validities. It is possible that the use of the manipulative in the Length-Measurement subtest resulted in students with greater conceptual understanding of measurement taking more time to carefully measure with the manipulative, completing fewer problems (albeit more accurately) in the given administration time. Students with lower measurement skills may have guessed the shorter object, resulting in a greater number of problems completed but with less accuracy. To test this theory, we conducted an exploratory alternate scoring of the Length-Measurement subtest using accuracy scores only (taking items correct/total items) and found that the measure significantly correlated with the

ASPENS and both easyCBM administrations ($r = .18$ to $.25$, $p < .001$). Caution in interpreting this result should be exercised as it is exploratory and places greater emphasis on accuracy (i.e. a student who completed only a few items but accurately is considered to have greater understanding of the construct than a student who completed more items but with some degree of error) than fluency. As such, future research should investigate alternate ways to design measures in the domain of measurement including the use of manipulatives, untimed measures or measures with a greater time limit, and alternate ways of scoring or use of a broader composite score while also striving to maintain CBM design features. Since the purpose of screening is to make a dichotomous decision (at-risk or not at-risk) at common screening points, future research should be conducted across a school year (i.e. Fall to Spring) and report metrics related to classification accuracy.

Student performance may have been confounded with the novelty of the task and the confounded with intervention delivery. The data collection team anecdotally noted that during the first administration multiple students did not understand the task but at the second administration students seemed to grasp task demands. The differences in understanding task demands may account for the lower test-retest reliabilities between time 1 and time 2. The issue of novelty also raises an interesting question about whether we should assess a skill that is not a primary focus of teaching or intervention but still considered critical for students to understand. At this point, we would not advocate for the use of measures studied within this research by practitioners in the field. However, we consider research in this area to be worthy of investigation by the research community to further our understanding of student development in measurement and to lay the groundwork for advances in screening, progress monitoring, and their link to intervention services.

References

Albers, C. A., & Kettler, R. J. (2014). Best practices in universal screening. In P. Harrison & A. Thomas (Eds.), Best practices in school psychology: Data-based and collaborative decision making (pp. 121–131). National Association of School Psychologists.

Alonzo, J., Tindal, G., Ulmer, K., Glasgow, A. (2006). EasyCBM online progress monitoring assessment system. University of Oregon. http://easycbm.com

Beckmann, S. (2008). Mathematics for elementary teachers. Pearson Addison Wesley.

Billstein, R., Libeskind, S., & Lott, J. W. (2004). A problem solving approach to mathematics for elementary school teachers. Pearson Addison Wesley.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Booth, J. L., & Siegler, R. S. (2008). Numerical magnitude representations influence arithmetic learning. Child Development, 79(4), 1016–1031. https://doi.org/10.1111/j.1467-8624.2008.01173.x

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(1), 46–60. https://doi.org/10.1111/ldrp.12122

Boulton-Lewis, G. M. (1987). Recent cognitive theories applied to sequential length measuring knowledge in young children. British Journal of Educational Psychology, 57(3), 330–342. https://doi.org/10.1111/j.2044-8279.1987.tb00861.x

Claessens, A., & Engel, M. (2013). How important is where you start? Early mathematics knowledge and later school success. Teachers College Record, 115, 1–29.

Clarke, B., Doabler, C. T., & Nelson, N. J. (2014). Best Practices in Mathematics Assessment and Intervention with Elementary Students. In P. Harrison & A. Thomas (Eds.), Best Practices in School Psychology: Data-Based and Collaborative Decision Making (pp. 219–232). National Association of School Psychologists.

Clarke, B., Gersten, R. M., Dimino, J., & Rolfhus, E. (2011). Assessing student proficiency of number sense (ASPENS) [Measurement instrument]. Cambium Learning Group, Sopris Learning.

Clarke, B., Strand Cary, M. G., Shanley, L., & Sutherland, M. (2018). Exploring the promise of a number line assessment to help identify students at-risk in mathematics. Assessment for Effective Intervention, 45(2), 151–160. https://doi. org/10.1177/1534508418791738

Clarke, B., Sutherland, M., & Doabler, C. T. (2017). Exploring the promise of a number line assessment to help identify students at-risk in mathematics. Center on Teaching and Learning, University of Oregon.

Clements, D. H., & Sarama, J. (2004). Learning and teaching early math: The learning trajectories approach. Taylor & Francis.

Common Core State Standards Initiative (CCSS). (2010). Common core standards for mathematics. Retrieved December 15, 2010, from http://www.corestandards.org/the-standards/mathematics

Common Core State Standards Writing Team (CCSS Writing Team). (2018). Progressions for the Common Core State Standards in Mathematics (August 10 draft). Institute for Mathematics and Education, University of Arizona. http://mathematicalmusings.org/wp-content/uploads/2018/08/ccss_progression_nf_35_2018_08_10.pdf

Confrey, J., Maloney, A., Nguyen, K., Lee, K. S., Panorkou, N., Corley, D., Avineri, A., Nickell, J., Neal, A., Varela, S., & Gibson, T. (2012). TurnOnCCMath.net Learning Trajectories for the K-8 Common Core Math Standards. NC State University, College of Education. http://turnonccmath.net/index.php

Cross, C. T., Woods, T. A., & Schweingruber, H. A. (2009). Mathematics learning in early childhood paths toward excellence and equity. National Academies Press. http://www.nap.edu/catalog/12519/mathematics-learning-in-early-childhood-paths-toward-excellence-and-equity

Doabler, C. T., Clarke, B., Kosty, D., Turtura, J. E., Firestone, A. R., Smolkowski, K., Jungjohann, K., Brafford, T. L., Nelson, N. J., Sutherland, M., Fien, H., & Maddox, S. A. (2019). Efficacy of a first-grade mathematics intervention on measurement and data analysis. Exceptional Children, 86, 77–94. https://doi.org/10.1177/0014402919857993

Doabler, C. T., Smolkowski, K., Clarke, B., Fien, H., Gause, M., & Nelson, N. J. (2015-2019). Precision Mathematics: Using interactive gaming technology to build student proficiency in the foundational concepts and problem solving skills of measurement and data analysis (Project No 1503161, awarded $2,999,702). National Science Foundation, Division of Research On Learning, Directorate for Education & Human Resources, Research on Learning in Formal and Informal Settings, Discovery Research K-12 (DRK-12), NSF 13-601, CFDA No. 47.076 http://www.nsf.gov/awardsearch/show-Award?AWD_ID=1503161.

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

Deno, S. L., Mirkin, P. K. (1977). Data-based program modification: A manual. Leadership Training Inst. for Special Education. http://files.eric.ed.gov/fulltext/ED144270.pdf

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. Exceptional Children, 49(1), 36–45. https://doi.org/10.1177/001440298204900105

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. Developmental Psychology, 43(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Dyson, N. I., Jordan, N. C., & Glutting, J. (2013). A number sense intervention for low-income kindergartners at risk for mathematics difficulties. Journal of Learning Disabilities, 46(2), 166–181. https://doi.org/10.1177/0022219411410233

Foegen, A. (2008). Progress monitoring in middle school Mathematics: Options and issues. Remedial and Special Education, 29(4), 195–207. https://doi.org/10.1177/0741932507309716

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in Mathematics. The Journal of Special Education, 41(2), 121–139. https://doi.org/10.1177/00224669070410020101

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. Educational Measurement: Issues and Practice, 28(3), 39–53. https://doi.org/10.1111/j.1745-3992.2009.00154.x

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

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. Exceptional Children, 57(6), 488–499. https://doi.org/10.1177/001440299105700603

Fuchs, L. S., Fuchs, D., Compton, D. L., Bryant, J. D., Hamlett, C. L., & Seethaler, P. M. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. Exceptional Children, 73(3), 311–330. https://doi.org/10.1177/001440290707300303

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

Hintze, J. M., Christ, T. J., & Keller, L. A. (2002). The generalizability of CBM survey-level mathematics assessments: Just how many samples do we need? School Psychology Review, 31(4), 514–528. https://doi.org/10.1080/02796015.2002.12086171

Kamii, C., & Clark, F. B. (1997). Measurement of length: The need for a better approach to teaching. School Science and Mathematics, 97(3), 116–121. https://doi.org/10.1111/j.1949-8594.1997.tb17354.x

Lembke, E. S., Carlisle, A., & Poch, A. (2016). Using curriculum-based measurement fluency data for initial screening decisions. In K. D. Cummings & Y. Petscher (Eds.), The fluency construct: Curriculum-based measurement concepts and applications (pp. 91–122). Springer. https://doi.org/10.1007/978-1-4939-2803-3_4

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

Morgan, P. L., Farkas, G., Hillemeier, M. M., & Maczuga, S. (2016). Who is at risk for persistent mathematics difficulties in the United States. Journal of Learning Disabilities, 49(3), 305–319. https://doi.org/10.1177/0022219414553849

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. Journal of Learning Disabilities, 42, 306–321. https://doi.org/10.1177/0022219408331037

National Mathematics Advisory Panel (NMAP). (2008). Foundations for success: The final report of the National Mathematics Advisory Panel. U. S. Department of Education. National Research Council. (2009). Mathematics learning in early childhood: Paths toward excellence and equity. National Academies Press. http://www.nap.edu/catalog.php?record_id=12519

Pedhazur, E. J. (1997). Multiple regression in behavioral research: Explanation and prediction (3rd. ed.). Wadsworth. Reys, R. E., Lindquist, M. M., Lambdin, D. V., & Smith, N. L. (2014). Helping children learn mathematics (11th ed.). John Wiley and Sons, Inc.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics (Vol. 2, pp. 957–1010). Information Age Publ. http://books.google.com/books?id=W4GnocmF02IC

Schneider, M., Merz, S., Stricker, J., De Smedt, B., Torbeyns, J., Verschaffel, L., & Luwel, K. (2018). Associations of number line estimation with mathematical competence: A meta-analysis. Child Development, 89(5), 1467–1484. https://doi.org/10.1111/cdev.13068

Siegler, R. S., Carpenter, T., Fennell, F. S., Geary, D., Lewis, J. R., Okamoto, Y., Thompson, L., & Wray, J. (2010). Developing effective fractions instruction for kindergarten through 8th grade: A practice guide (Report No. NCEE #2010-4039). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/fractions_pg_093010.pdf

Snedecor, G. W., Cochran, W. G. (1980). Statistical methods. Iowa State University Press. https://trove.nla.gov.au/work/10694367?

Stephan, M., & Clements, D. H. (2003). Linear, Area, and Time Measurement in Prekindergarten to Grade 2. In D. H. Clements & G. Bright (Eds.), Learning and Teaching Measurement (2003 Yearbook) (pp. 3–16). National Council of Teachers of Mathematics.

Sutherland, M., Clarke, B., Nese, J., Strand Cary, M., Shanley, L., Furjanic, D., & Williams, C. (2011). Investigating the utility of a kindergarten number line assessment compared to an early numeracy screening battery. Early Childhood Research Quarterly, 55, 119–128. https://doi.org/10.1016/j.ecresq.2020.11.003

Van De Walle, J. A. (2019). Elementary and middle school mathematics: Teaching developmentally. Pearson.

Vanderheyden, A., Codding, R., & Martin, R. (2017). Relative value of. School Psychology Review, 46(1), 65–87. https://doi.org/10.17105/SPR46-1.65-87

VanDerHeyden, A. M., & Broussard, C. (2019). Construction and examination of Math subskill mastery measures. Assessment for Effective Intervention, Advance online publication. https://doi.org/10.1177/1534508419883947

VanDerHeyden, A. M., Broussard, C., & Burns, M. K. (2019). Classification agreement for gated screening in Mathematics: Subskill mastery measurement and classwide intervention. Assessment for Effective Intervention. https://doi.org/10.1177/1534508419882484

VanDerHeyden, A. M., Broussard, C., Snyder, P., George, J., Lafleur, S. M., & Williams, C. (2011). Measurement of kindergartners' understanding of early Mathematical concepts. School Psychology Review, 40(2), 296–306. https://doi.org/10.1080/02796015.2011.12087719

VanDerHeyden, A. M., & Burns, M. K. (2018). Improving decision making in school psychology: Making a difference in the lives of students, not just a prediction about their lives. School Psychology Review, 47(4), 385–395. https://doi.org/10.17105/SPR-2018-0042.V47-4

VanDerHeyden, A. M., McLaughlin, T., Algina, J., & Snyder, P. (2012). Randomized evaluation of a supplemental gradewide mathematics intervention. American Educational Research Journal, 49(6), 1251–1284. https://doi.org/10.3102/0002831212462736

Vasilyeva, M., Laski, E. V., Veraksa, A., & Bukhalenkova, D. (2020). Leveraging measurement instruction to develop kindergartners' numerical magnitude knowledge. Journal of Educational Psychology. https://doi.org/10.1037/edu0000653

Witzel, B., & Clarke, B. (2015). Focus on Inclusive Education: Benefits of Using a Multi-tiered System of Supports to Improve Inclusive Practices. Childhood Education, 91, 215–219. https://doi.org/10.1080/00094056.2015.1047315

Table 1

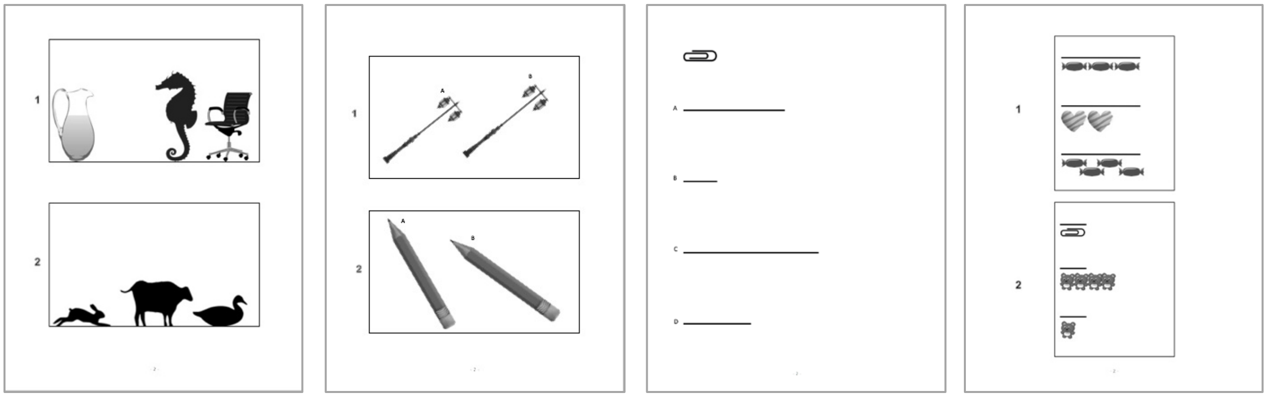
Descriptive Statistics and Bivariate Correlations for All Study Variables (T1 n = 219; T2 n = 213)

Variable	1	2	3	4	5	6	7	8	9	10	11	*M* (*SD*)
1. Length-Comparison (T1)												16.02 (4.87)
2. Length-Measurement (T1)	.16*	-										10.43 (5.51)
3. Iteration-Application (T1)	.41**	.07	-									10.21 (7.83)
4. Iteration-Conceptual (T1)	.28**	.02	.22*	-								9.41 (2.94)
5. Length-Comparison (T2)	.46**	.11	.33**	.14*	-							17.00 (4.33)
6. Length-Measurement (T2)	.09	.33**	.07	.07	.12	-						9.66 (5.62)
7. Iteration-Application (T2)	.40**	.02	.48**	.24**	.44**	.13	-					13.76 (8.71)
8. Iteration-Conceptual (T2)	.25**	.21**	.16*	.42**	.36**	.20**	.25**	-				10.38 (2.76)
9. ASPENS Composite (T1)	.43**	-.09	.42**	.25**	.37**	.02	.51**	.17*	-			36.13 (22.17)
10. EasyCBM (T1)	.34**	-.14*	.37**	.16*	.30**	.07	.46**	.12	.58**	-		21.40 (5.37)
11. EasyCBM (T2)	.44**	-.07	.39**	.23**	.34**	-.02	.48**	.24**	.56**	.56**	-	27.68 (5.21)

Note. Correlations calculated using pairwise deletion. *M* = mean, *SD* = standard deviation. T1 = Time 1, T2 = Time 2. **p* < .05, ***p* < .01.

Figure 1.

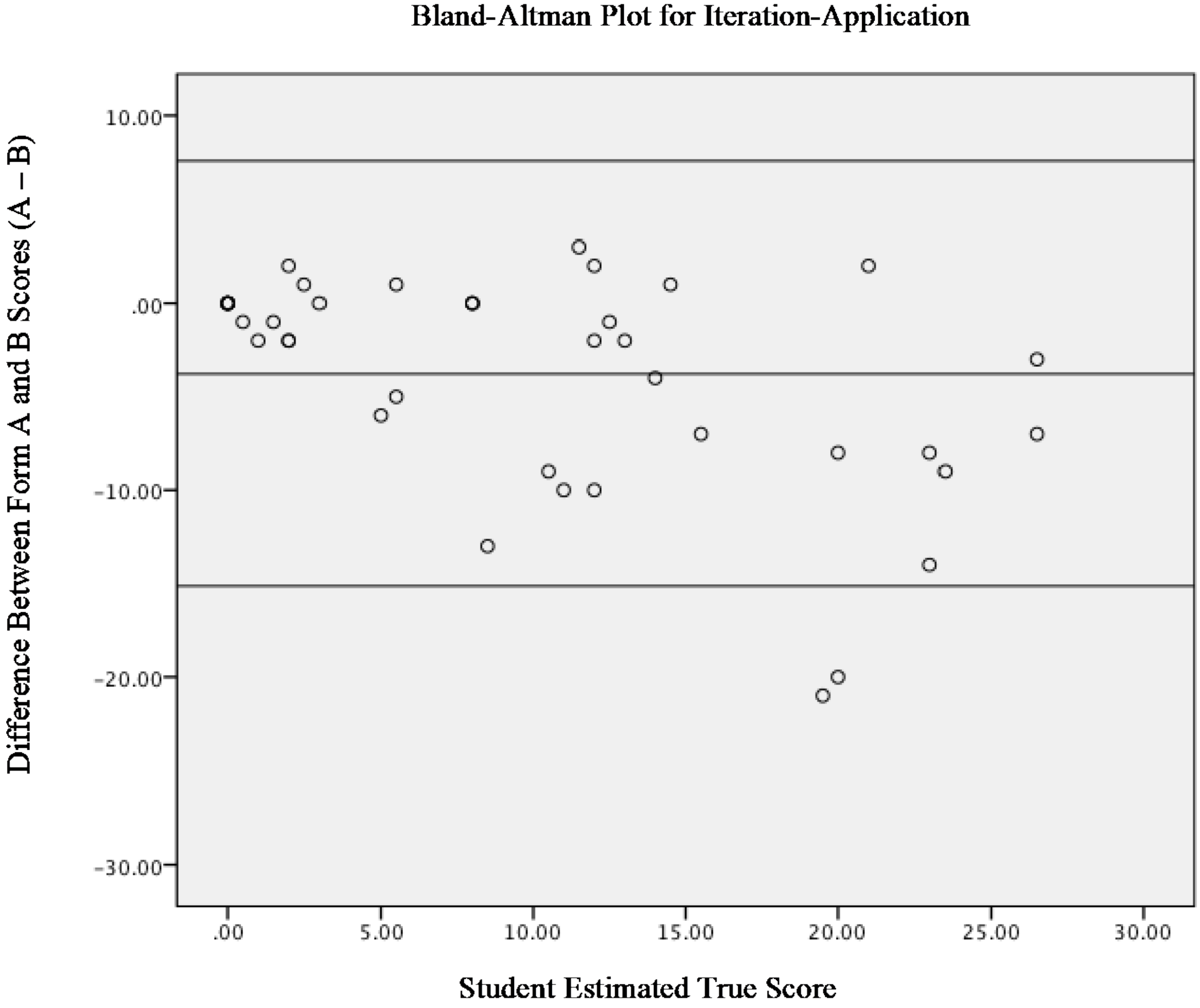
The First Test Page of Each EM-CBM Subtest.



Note. In order from left to right: Length-Comparison, Length-Measurement, Iteration-Application, Iteration-Conceptual.

Figure 2.

Proportional Bias Detected for Alternate Forms.



Note. To illustrate an example of proportional bias detected for alternate forms, the Bland-Altman plot for Iteration-Application is shown here. Proportional bias is present such that as a student's estimated true score (x-axis) increases, the difference between scores on Forms A and B (y-axis) becomes more negative. Regressing the difference between scores on the estimated true score indicated that proportional bias was statistically significant (B = -.368, p < .001).