Properties of curriculum-based measurement for mathematics: an investigation of the average growth, variability, and precision of three forms

Arianna Doss
*University at Albany, State University of New York*, adoss426@gmail.com

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Psychology Commons

**Recommended Citation**

This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive. Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
Properties of Curriculum-Based Measurement for Mathematics: An Investigation of the
Average Growth, Variability, and Precision of Three Forms

By

Arianna Doss

A Dissertation
Submitted to the University at Albany, State University of New York
In Partial Fulfillment of
the Requirement for the Degree of
Doctor of Psychology

School of Education
Department of Educational and Counseling Psychology
Fall 2021
ABSTRACT

CBM for mathematics assesses growth in accuracy and fluency of basic math skills using content from a student’s curriculum. CBM for mathematics can include single-skill measures (SSM), skill-based measures (SBM), and general-outcome measures (GOM). Past research on growth rates in CBM for mathematics has focused on GOMs and has relied on estimations of weekly growth rates, but more information on expected growth rates and their precision for these measures is needed for practice. The current study addresses this gap in the literature by examining weekly growth rates and their precision for one SSM containing multiplication problems, one SBM containing addition and subtraction problems, and a GOM, M-COMP, with third grade students. Results suggest that individual estimates of the $SEE$ varied widely across participants and between the SSM, SBM, and GOM. However, the SSM results were more precise than the other two measures. The average slope for the SSM was 3.50 DCPM/wk with an average $SEE$ of 6.49 DCPM, and 1.89 PC/wk with an average $SEE$ of 3.57 PC. The HLM results indicated that initial performance on the SBM and SSM varied by individual student. Additionally, the initial performance and rate of improvement over time varied significantly across classroom. Regarding the GOM, the 3-level model demonstrated that student initial performance varied by student as well. Implications for the consideration of SSMs when engaging in high-stakes decisions in schools are discussed.
# TABLE OF CONTENTS

**ABSTRACT** ......................................................................................................................... ii

**TABLE OF CONTENTS** ........................................................................................................ iii

**LIST OF APPENDICES** ......................................................................................................... v

**CHAPTER 1: INTRODUCTION** .............................................................................................. 1

Overview of Response to Intervention .................................................................................. 1

Impact of Federal Legislation ................................................................................................. 1

Curriculum-Based Assessment ............................................................................................... 2

Curriculum-Based Measurement ............................................................................................ 2

Decision-Making in Curriculum-Based Measurement for Mathematics .............................. 4

Statement of the Problem ...................................................................................................... 5

Purpose of the Study ............................................................................................................. 5

Importance of the Study ........................................................................................................ 5

**CHAPTER 2: REVIEW OF THE LITERATURE** .................................................................. 7

Response to Intervention ...................................................................................................... 7

Legislative History or Response to Intervention ................................................................... 8

Framework of Response to Intervention ............................................................................. 9

Curriculum-Based Assessment ............................................................................................. 13

Curriculum-Based Measurement .......................................................................................... 14

Purpose of the Study ........................................................................................................... 31

Research Questions ............................................................................................................. 31

**CHAPTER 3: METHODOLOGY** ......................................................................................... 33
# LIST OF APPENDICES

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Sample Subskill Probe- Multiplication</td>
<td>84</td>
</tr>
<tr>
<td>B. Sample Subskill Probe- Addition &amp; Subtraction</td>
<td>86</td>
</tr>
<tr>
<td>C. Single-Skill Math Facts Standard Directions</td>
<td>88</td>
</tr>
</tbody>
</table>
Chapter One: Introduction

Overview of Response to Intervention

Response to Intervention (RtI) is a framework consisting of several tiers of instructional environments to identify students who require more intensive supports and to monitor their progress in school. Tier I of the framework consists of instruction given to all students, Tier II provides more supports and usually involves small groups of students, and Tier III includes more intensive services for individual students (Brown-Chidsey & Steege, 2010). Students who do not respond to intervention may require more intensive intervention through a process using screening and progress monitoring data to inform how students progress through the tiers. Legally, many states require schools to use data to inform their decisions, but they can decide which tools they will use and how they will use them (No Child Left Behind Act of 2001, 2002).

Impact of Federal Legislation

Federal legislation, including the Elementary and Secondary Act (ESEA), which was reauthorized as the No Child Left Behind Act (NCLB), has made data-based decision-making a requirement in schools (Gresham, 2005; No Child Left Behind Act of 2001, 2002). NCLB was reauthorized as Every Student Succeeds Act (ESSA), which maintains the importance of using data and holds schools accountable for student achievement (Mathis & Trujillo, 2016). Legislation regarding students with disabilities also changed, including the Individuals with Disabilities Education Improvement Act (IDEIA) of 2004. IDEIA included the use of data to evaluate whether a student responds to intervention as a possible means of determining if a student has an educational disability (Individuals with Disabilities Education Act, 2004). To abide by the legislation
and collect data to guide decision-making, teachers needed to use formative assessment to assess student progress within the curriculum.

**Curriculum-Based Assessment**

Curriculum-based assessment (CBA) samples content directly from a student’s curriculum to assess progress. Using CBA, instructors create tests and administer them repeatedly to help inform their teaching (Fuchs & Deno, 1991; Tucker, 1987). CBA is a broad construct, and includes specific frameworks of collecting data, such as curriculum-based evaluation (CBE) and curriculum-based measurement (CBM; Christ et al., 2014). CBE entails a scripted process of hypothesis testing to identify effective interventions, including initial assessment, hypothesis generation, direct assessment of skills, and results interpretation. CBE requires the use of more specific measures, such as CBM, which, in contrast, is used only to assess student performance (Christ, et al., 2014; Howell et al., 2008; Kelley et al., 2008).

**Curriculum-Based Measurement**

CBM provided the missing link teachers needed to implement RtI and abide by legal mandates to collect data to inform instructional decisions. CBM utilizes standardized procedures and materials that sample a student’s curriculum (Deno, 2003). It was created to allow teachers to make day-to-day decisions about their instruction based on students’ idiographic and nomothetic performance over time (Salmon-Cox, 1981). Although CBM was created prior to the changing legislation, it has been integrated into the RtI framework and is now used near universally to screen and monitor progress (Fuchs & Vaughn, 2005; Hosp et al., 2007; Rowe et al., 2014; Shinn, 2007). CBM instruments have been created for literacy, mathematics, and writing, but most
research to date has focused on a specific CBM, oral reading fluency (Ardoin, et al., 2013). Researchers have studied the psychometric properties of CBM for reading and its relation to student performance (Ardoin & Christ, 2009; Ardoin et al., 2013; Christ, 2006; Deno et al., 2001; Francis et al., 2008; Good & Shinn, 1990; Hintze & Christ, 2004; Hintze et al., 1994; Parker & Tindal, 1992; Stecker et al., 2005). Guidelines for data-collection and analysis for CBM for reading have been developed to aid teachers in collecting and using data (Ardoin et al., 2013). However, in contrast to the abundance of research on reading CBM, CBM for mathematics is significantly underdeveloped.

CBM for mathematics can be used to measure growth in acquisition/accuracy and fluency of basic math skills. Fluency with basic mathematical computational skills is important because as students answer problems quickly, they inherently practice more problems and can conserve their cognitive resources for higher-level tasks (Dahaene, 1997; Haring et al., 1978). To measure fluency, three common scoring systems include correct digits in the solution (CD-S), digits correct per minute (DCPM), and correct problems (CP). In literacy, teachers are encouraged to use or create CBM tests based on subskills, called Specific Subskill Mastery Measurements (SSM), or based on an entire year of curriculum, called General Outcomes Measurement (GOM). In contrast, for CBM for mathematics, single-skill measures (SSM), as opposed to subskill mastery measures, are used. They include problems of all one type, such as multiplication. For CBM for mathematics, it has been difficult to develop GOMs, so another type of measurement, called skill-based measures (SBM), has been introduced (Kelley et al., 2008) to be used in conjunction or as a replacement to GOMs. SBMs target a select few skills within one mathematics domain, such as computation.
**Decision-Making in Curriculum-Based Measurement for Mathematics**

The typical magnitude and precision of slope has been studied to help teachers make instructional decisions using student growth, or rate of learning, to determine when to modify interventions (Fuchs, 2004). Teachers can use decision-rules derived from this research (e.g., when enough data accrues to make a decision) to decide if a student is making adequate progress. For example, they can either compare the slopes of a student’s trend line to an aimline, or expected growth rate under intervention, or they can look at individual data points and decide whether enough data points are above or below the aimline, suggesting a change to the intervention or goal (Ardoin et al., 2013; Parker & Tindal, 1992). Instructors can increase the intensity of the intervention for students who do not progress and can increase goals for students who do progress (Ardoin et al., 2013; Deno et al., 2002). Research on growth slopes is abundant for CBM for reading but is lacking for CBM for mathematics.

More research is needed to identify the stability of growth slopes for math CBM to develop guidelines on how long to monitor progress before changes to instruction can occur with reasonably accuracy. Additional research is also needed on expected growth rates for CBM for mathematics to help teachers compare their student data to existing data. Researchers have studied growth rates by focusing mostly on GOMS, particularly using M-CBM, from AIMSweb (NCIS Pearson, Inc., 2018), as well as other unspecified GOMs (Fuchs et al., 1993; Graney, et al., 2009; Keller-Margulis, et al., 2014). Additional research focusing on SSMs and SBMs is necessary. Furthermore, another AIMSweb GOM, called M-COMP, has not yet been studied by any researchers aside from the test publishers. The test publishers’ data was collected two to three times per
month and actual rates of improvements, typically known per week of measurement, were not provided (AIMSweb Technical Manual, 2012).

**Statement of the Problem**

Although several studies have investigated growth rates for CBM for mathematics, limited research has been conducted using weekly data specific to SSMs, SBMs, and M-COMP. Presently, average growth and the precision of slopes for these well-used instruments are largely unknown.

**Purpose of the Study**

The purpose of the present study is to add to the current literature regarding growth rates for CBM for mathematics. This study’s purpose is to examine weekly growth rates and their precision for one SSM containing multiplication problems, one SBM containing addition and subtraction problems, and M-COMP, which are measures of CBM for mathematics, with third grade students.

**Importance of the Study**

Legislation requires that schools use data to inform their decision-making regarding student progress. RtI provides a framework for delivering interventions to students and CBM is a useful tool for gathering data. While most of the current literature has focused on CBM for reading, information on expected growth rates for CBM for mathematics is needed for teachers to compare student data and decide when to make instructional changes. Additionally, research on the stability of growth slopes for CBM for mathematics is necessary to inform instructors on how long to collect data before analysis. Existing literature on CBM for mathematics has focused on GOMs, widely
excluding M-COMP, and has relied on estimations of weekly growth rates. Therefore, the present study is designed to address these gaps in the extant literature.
Chapter Two: Review of the Literature

The purpose of the current study is to investigate expected weekly growth rates and their associated variability with curriculum-based measurement (CBM) for mathematics for third graders. The following chapter provides background information on the current study, beginning with Response to Intervention (RtI), including its legislative history and general framework. The use of an RtI framework created a need for formative assessment, which was filled by CBM, as CBM allowed for teachers to frequently screen and monitor student progress. The origins of CBM are discussed along with curriculum-based assessment (CBA) and curriculum-based evaluation (CBE). Research on CBM for reading and for mathematics is reviewed, with particular attention to research on expected growth rates.

Response to Intervention

RtI is a systematic framework that involves the use of data-based decision-making to improve student learning (Batsche et al., 2006). Educational reforms have influenced the progression of public education towards using evidence-based procedures – their application organized within the RtI framework – which emphasizes using data to both isolate student instructional deficits and use response to treatment to make future diagnostic/intervention decisions. This tiered model of service developed as an efficient alternative to the “test-and-place” model of cognitive assessment and diagnostic decision-making, which had been criticized along multiple domains, including the entire validity of the process brought into question (Fuchs et al., 2003; Meciak et al., 2015). The adoption of RtI requires newer forms of academic assessment that can be used to both summarize, screen, and progress monitor.
Legislative History of Response to Intervention

Legal mandates from the early 2000s have made the use of evidence-based practices, and valid assessment of a student’s learning while receiving such supports, a requirement in schools (Gresham, 2005). The Elementary and Secondary Act (ESEA) of 1965 was reauthorized as the No Child Left Behind Act (NCLB) of 2001. NCLB, created under the George W. Bush administration, held schools accountable for ensuring that students demonstrated proficiency in basic academic skills by using scientific research-based practices to improve student achievement (No Child Left Behind Act of 2001, 2002). President Bush also created the President’s Commission on Excellence in Special Education (PCESE) in 2001 to provide recommendations for improving all areas of special education. The commission’s report recommended that schools focus on prevention and intervention, where students receive special education instruction as a direct and measured response to students’ success in general education instruction.

Assessment within an RtI framework can help identify students who require intervention, prior to or after being identified with a significant instructional deficit resistant to intervention (i.e., a learning disability). The report also noted that although the Individuals with Disabilities Education Act (IDEA) led to some success for students with disabilities, more reforms were needed to better support them (President’s Commission on Excellence in Special Education, 2002). Later, under the President Barack Obama administration in 2015, NCLB was reauthorized as the Every Student Succeeds Act (ESSA), which still emphasizes accountability and requires states to use evidence-based interventions (Mathis & Trujillo, 2016). To do this, valid assessment is
necessary to gauge the need for intervention and monitor progress in both regular and special education.

Congress responded to the PCESE report by revising the IDEA of 1990. The new act was called the Individuals with Disabilities Education Improvement Act (IDEIA) of 2004. IDEIA encouraged the use of RtI in schools and specifically, its use in determining whether a student has a specific learning disability. When determining eligibility, the law dictates that schools “…May use a process that determines if the child responds to scientific, research-based intervention as a part of the evaluation procedures” (Individuals with Disabilities Education Act, 2004, p. 118). IDEIA went further to give public school districts the power to adopt RtI, regardless of the policy of the state. Although the language of PCESE and IDEIA appear to focus on students in special education, RtI is widely considered a general education initiative that serves all students (Canter et al., 2008). To use RtI correctly, the practices must originate in general education (Brown-Chidsey & Steege, 2010). This is commonly envisioned in practice as a three-tiered model of service delivery that prioritizes measurement and decision-making with all students in all environments (Brown-Chidsey & Steege, 2010; Fuchs, Mock, et al., 2003; Shinn, 1989). In other words, RtI uses general and special education resources to both prevent and respond to significant instructional deficits across environments.

Framework of Response-to-Intervention

The development of RtI occurred prior to the previously mentioned legislation, in response to problems in the education system that had been developing over decades. These problems included a disconnect between general and special education service delivery, little emphasis on prevention and early intervention in schools, limited use of
research-based instruction and intervention, and a mismatch between Specific Learning Disability identification and theory and eligibility procedures and consequent interventions (National Association of State Directors of Special Education [NASDSE], 2005).

Curriculum-based measurement was initially conceptualized as a procedure to assist teachers in making instructional decisions based on student performance within the general classroom environment (Deno, 1985). However, others argued for the further expansion of CBM within an RtI framework to conduct universal screening, determine special education eligibility, and monitor progress within the RtI framework (Busch & Reschly, 2007; Shinn, 2007; Stecker et al., 2008). Shinn (2007) stated that due to the research-base of CBM, and its explicit design as a measure of intervention responsiveness, it “seems logical that CBM would be the primary tool in an RtI process when there are concerns about a student’s basic skills” (p. 608). The use of CBM has been demonstrated in evaluating both a student’s response to intervention and in differentiating student need within a screening process (Fuchs & Vaughn, 2005; Hintze, et al., 2000; Shinn, 2007). It is now universally used within RtI and is widely accepted among teachers (Rowe et al., 2014).

The RtI framework contains multiple tiers that represent a continuum of supports ranging from the universal level to the individual level (Barnes & Harlacher, 2008). Some models contain differing numbers of tiers, but the three-tiered model is the most widely accepted (Brown-Chidsey & Steege, 2010; Fuchs & Fuchs, 2005; Fuchs, Mock, et al., 2003; Ikeda et al., 2002). Tier I is conceptualized as the general education curriculum provided to all students at the universal level, typically with triannual screening. Ideally,
about 80% of students are successful in this tier, allowing for efficient distribution of resources to students who do not respond to Tier I. Tier II services are provided to students who are unsuccessful within the general education curriculum alone. Ideally, about 15% of students require Tier II services, which often occur at the small group level and entails modest additional support. Typically, students receiving such services are progress monitored weekly. Ideally, no more than 5% of all students are unsuccessful at Tier II and therefore require more intensive Tier III services at the individual level. Those who do not respond to Tier I and Tier II services may undergo comprehensive academic evaluation to determine whether they meet requirements for special education (Brown-Chidsey & Steege, 2010) and likely require individualized, intensive services that cannot be maintained in the general education environment. At all levels, formative assessment must be conducted to measure student progress.

Assessment in Tier I commonly involves universal screening three times a year in the fall, winter and spring, using relevant forms of CBM (Brown-Chidsey & Steege, 2010; Good & Kaminski, 2002). Screening data should be used to identify which students need additional instruction at additional tiers of service, or whether large groups of identified non-responders require a reconceptualization of Tier I instruction. Specifically, RtI implementers must compare each student’s performance on screening instruments to a criterion or normative standard.

In Tiers II and III, assessment occurs regularly in the form of progress monitoring, also using CBM. Progress monitoring assesses a student’s responsiveness to instruction through quantifying his or her rate of improvement over time (Deno et al., 2009). For students in Tier II, researchers have made different recommendations for the
frequency of this type of assessment. These recommendations include reviewing growth every eight weeks, every four weeks, every week, or twice a week, and using one or multiple probes per measurement session to determine if students are adequately responding to an intervention (Brown-Chidsey & Steege, 2010; Coyne & Harn, 2006; Marston et al., 2007; Vaughn et al., 2007). More broadly, researchers have typically suggested collecting between three to 20 data points, with seven being the most common (Ardoin et al., 2013). Ardoin and colleagues (2013) recommended that at least five data points should be collected over at least three weeks. No matter the schedule, it is imperative that progress monitoring of students in Tier II, who typically receive small-group intervention within the classroom at least twice a week, occurs more frequently than universal screening. This helps to document change and helps teachers decide whether they should modify their instruction or if the student should move up or down a tier.

In Tier III, progress should be monitored more frequently than in Tier II, such as daily or weekly, and with CBMs that target smaller instructional units (i.e., subskills; Stecker, 2007). Again, there is no set standard for the frequency of data collection in Tier III, but like in Tier II, frequent collection will permit calculation of a slope, which is used to determine if the student responds to intensive individualized instruction and allows for efficient instructional changes. This guarantees that RtI implementers use formative assessment, which is ongoing assessment used to make decisions about student progress while the students participate in the intervention (Howell & Nolet, 2000). To use formative assessment, teachers must set goals, monitor progress, modify instruction, and evaluate the effects of those modifications against goals for the rate of learning (Deno,
2003). As opposed to data collected at the end of the school year, grading period (i.e., summative), for a clinical evaluation, formative data is especially useful to students who are academically performing below expectations (Deno et al., 2002).

When RtI was originally conceptualized in the 1980’s (Deno, 1985), no such formative measures existed to adequately screen and progress monitor students. Extant standardized measures for decision-making were summative in nature (e.g., broadband academic achievement tests). CBM emerged to fill this need for direct, quick, formative assessment.

**Curriculum-Based Assessment**

CBM evolved from an earlier conception of assessment called curriculum-based assessment (CBA). CBA involves making instructional decisions based on how a student performs within his or her school curriculum (Deno, 1987). Three fundamental features of CBA are that (a) test materials are taken directly from a student’s curriculum; (b) students are tested repeatedly over time; and (c) scores are used to inform decisions regarding instruction (Fuchs & Deno, 1991; Tucker, 1987). CBA was conceptualized to allow teachers to understand how a student performed on content within the curriculum, in contrast to norm-referenced tests (Christ et al., 2014). CBA is a term used to describe several diverse assessment practices rather than one specific set of procedures (Hintze et al., 2006).

After comparing different models of CBA, Jones and colleagues (1998) concluded that all models included the following steps: analyze the curriculum, determine student’s current level of functioning, set specific target behaviors and criteria for success, design assessment instruments, collect and display data, and make educational
decisions. To utilize CBA in its traditional sense, a teacher would evaluate students’ skill development across the curriculum, sampling across skills and creating locally valid measurements (Hintze, et al., 2006). CBA is an umbrella construct, which includes curriculum-based evaluation (CBE) and CBM (Christ et al., 2014). Both CBE and CBM involve use of curriculum samples to assess student performance and guide instruction.

**Curriculum-based evaluation.** CBE is a formalized process of CBA, which includes initial assessment, hypothesis generation, direct assessment of skills, and results interpretation (Howell et al., 2008; Kelley et al., 2008). Initial assessment can involve survey-level assessment or previously collected data and is intended to identify whether the student demonstrates a skill deficit (Christ, et al., 2014; Howell & Nolet, 2000). After that, the examiner develops hypotheses about possible causes for the student’s skill deficit and tests those hypotheses through assessment of specific skills in the curriculum. This process is generally conceptualized as a flow chart, beginning with more broad assessment and culminating in specific measurement of subskill deficits (Howell & Nolet, 2000). Formative decision-making in CBE can involve the use of CBM, or another type of CBA, as required (Christ, et al., 2014). As CBE is an assessment framework rather than a specific assessment measure, researchers have not investigated its reliability and validity (Howell et al., 2008).

**Curriculum-Based Measurement**

In contrast, CBM is a generalized measurement system with a set of standardized procedures and materials for administration and scoring that is used to assess student performance across curriculums (Deno, 1985; Deno, 2003). CBM is considered a form of CBA because the materials were originally derived from the materials used in
classrooms (Deno, 2003). Contemporary CBM materials are intended to be construct valid for measuring progress within most major curriculums used by schools by targeting core basic skills. CBM has a strong research-base, beginning in the 1980s, and is technically adequate in terms of its reliability and validity as an indicator of basic skills (Christ & Ardoin, 2009; Christ et al., 2014; Christ et al., 2013; Deno, 1985; Deno, 2003; Deno, et al., 1982; Good & Jefferson, 1998; Hintze et al., 2000; Menthe, et al., 2008; Shinn, 1989). CBM can be used to screen students to identify those who are at-risk, to progress monitor growth in the presence of intervention, and to plan instruction (Christ et al., 2014; Deno, 2003; Deno et al., 2001; Fuchs et al., 1993; Fuchs, et al., 1994).

CBM can be used as both an idiothetic and nomothetic assessment. Idiothetic assessment compares the performance of a single person to him or herself, whereas nomothetic assessment takes others’ performances into account (Haynes & O’Brien, 2000). CBM can provide information on a student’s performance as compared to his or her previous performance or as compared to other students’ performances.

CBM can be used within normative and criterion-referenced frameworks as well (Shinn, 1989). In norm-referenced assessment, student performance is compared to the performance of a similar group of students who completed the same measure (i.e., the norming group; Tallmadge, 1981). Criterion-referenced assessment involves items taken from an instructional domain and uses a standard of performance to determine mastery (Shinn, 1989).

CBM was created to give teachers data to make decisions about their teaching (Salmon-Cox, 1981). Prior to the creation of CBM, teachers relied on informal inventories to assess student knowledge (Deno, 1985). They also relied on commercially
developed, standardized achievement tests for student evaluation, but these did not help teachers make day-to-day decisions about modifying instruction because they could not be used formatively to evaluate changes in student growth conditional on changes in the instructional environment, nor could they be used for mass screening (Deno, 1985; Salmon-Cox, 1981). These limitations led to the development of CBM to guide instructional decision-making. CBM is a form of CBA that is meant to be practical and valid across curriculums (Hintze, Shapiro, & Lutz, 1994). The use of CBM has been shown to increase teacher effectiveness in terms of student achievement by raising awareness as to how students are responding to core instruction in schools (Fuchs, et al., 1984).

**CBM for reading.** CBM has been developed for literacy, math, and writing. By far the most well-developed of CBM instruments is for reading, with the bulk of research focusing on oral reading fluency (Ardoin et al., 2013). From this research, probe equivalencies have been established and decision-making guidelines generated. CBM for reading procedures are sensitive to changes in growth and the use of these procedures has been shown to increase student performance (Fuchs, et al., 1989; Stecker et al., 2005; Deno et al., 2001; Hintze et al., 1994; Shinn & Marston, 1985; Shinn, et al., 1989). Student performance can improve through CBM procedures because teachers change their teaching behavior based on data by altering instruction more often or creating more specific goals for students (Fuchs et al., 1989; Stecker et al., 2005). Recommendations have been provided for the number of data points for teachers to collect, which helps teachers know when to analyze data before continuing with instruction or modifying it (Ardoin et al., 2013).
Additionally, several studies have reported data on slope and precision of slope for reading CBM (Ardoin & Christ, 2009; Ardoin et al., 2013; Christ, 2006; Francis et al., 2008; Good & Shinn, 1990; Hintze & Christ, 2004; Parker & Tindal, 1992; Shinn, et al., 1989). For example, Christ and colleagues (2013) suggested that CBM be collected twice a week for approximately 10 weeks. This data is useful for those making assessment choices for schools and developing policy regarding RtI. Unfortunately, the research base for CBM for mathematics is not nearly as well developed as CBM for reading. The same rules from reading should not apply to mathematics without more research and investigation of error rates, rates of improvement, and recommended lengths of progress monitoring for mathematics.

**CBM for mathematics.** In mathematics, CBM typically involves students solving computational problems (Deno, 2003; Hosp et al., 2007). These tasks, or probes, are simple to administer and have differing administration times. Different, but equivalent, CBM probes must be repeatedly administered over time to allow for comparison to monitor progress effectively (Deno, 2003).

CBM for mathematics can measure accuracy, or acquisition, and fluency. These are considered foundational stages in learning a skill (Haring & Eaton, 1978). Response accuracy is the first stage of learning and is considered a prerequisite to fluency. Fluency is the ability to respond to a stimulus (i.e., the math problem) quickly, correctly, and with minimal effort (Haring & Eaton, 1978), which proceeds acquisition.

**CBM for mathematics accuracy.** Instruction to improve accuracy may involve teaching strategies such as modeling accurate behaviors, providing direct feedback, and allowing for many opportunities for students to practice the target skill (Codding, et al.,
CBM may be valid for measuring math accuracy by capturing what proportion of problems in a response set is answered correctly.

There is no consensus on defining mastery for accuracy in mathematics (Burns, et al., 2006). Codding and colleagues (2017) indicated that functional acquisition ranges from 60 to 100%, Gickling and Thompson (1985) suggested a range of 70-85% for mastery of accuracy, whereas Burns and Klingbeil (2010) suggested 90%. Bloom (1968) considered mastery to be at 80% and specific interventions often use 80% as a criterion (Durr, et al., 1981; Ravthon, 2008). Once a student’s accuracy improves to a criterion, the teacher can alter his or her form of instruction. For example, a teacher might provide less immediate feedback, or give fewer demonstrations (Codding et al., 2017) but increase opportunities to practice. When learning to respond accurately, students may become dependent on time-consuming manipulatives or strategies and may only be able to arrive at the correct answer when using these procedures (Poncy, et al., 2007; Skinner & Schock, 1995; Stokes & Baer, 1977).

Once students demonstrate mathematical accuracy, teachers can then target fluency (Skinner, et al., 2005). Fluency is important because as students demonstrate the ability to solve problems quickly, other cognitive resources, such as working memory, become available for other tasks (Dahaene, 1997). When a student is fluent, he or she can practice more problems (Haring et al., 1978), allowing for generalization of the skill to higher order math. Both accuracy and fluency are necessary for developing more complex mathematics skills (Poncy, et al., 2007).

**CBM for mathematics fluency.** There are competing metrics present for evaluating fluency in CBM for mathematics: correct digits in the solution (CD-S), digits
correct per minute (DCPM) and correct problems (CP). CD-S involves scoring the number of digits correct in the solution, which includes critical processes in addition to the final answer. DCPM only counts digits in the final answer. Although this method does not account for problems that take longer than others (i.e., a division problem with two digits in the answer is worth the same amount as an addition problem with two digits in the answer, even though division takes longer than addition), DCPM is easier to score and more reliable than CD-S (Hosp et al., 2016). CP involves calculating the number of problems solved correctly and is important because it is ultimately necessary for a student to arrive at a correct answer.

While DCPM and CD-S are more sensitive to growth, CP might be better used for screening purposes, and is relevant for estimating accuracy (Fuchs, et al., 1990; Hosp et al., 2016). CP can also be used in math concepts and applications for problems where students must choose the correct answer instead of computing the answer (Stecker et al., 2005). Conversely, DCPM and CD-S may be more valid and representative of task effort because they award more points for more difficult problems, rather than one point per correct answer, no matter the item difficulty (Hosp et al., 2007; Hosp et al., 2016). A fourth way to measure fluency in CBM for mathematics has recently been developed and is used to score the AIMSweb collection of mathematics probes called Mathematics Computation (M-COMP). This method involves weighting items based on the length of time they typically take to complete in relation to the number of digits or problems correct. Items are assigned between one and three points per correct answer. Ultimately, when deciding upon a metric to evaluate CBM for mathematics, one should use the
metric used by test developers when comparing student performance to benchmarks or norms (Hosp, et al., 2016).

**CBM for mathematics test content.** Fuchs and Deno (1991) described two models for CBM, including Specific Subskill Mastery Measurement (SSM) and General Outcomes Measurement (GOM). SSMs assess discrete, underlying skills that, collectively, result in increased performance on the GOM, which assesses capstone skills. SSMs involve identifying a subskill hierarchy within an academic area based on the curriculum and creating test materials for each part of the hierarchy. Students repeatedly complete the tests until they demonstrate mastery, as set by a criterion. The teacher then moves through the hierarchy in his or her instruction and assessment. For example, in mathematics, a teacher may start with teaching and assessing single-digit addition without regrouping and then move on to single-digit addition with regrouping. A GOM measure would contain both types of problems as well as more advanced problems entailing addition that teachers will teach later in the year. An example of a GOM is AIMSweb’s M-COMP, mentioned previously.

SSMs break down curriculum into smaller subskills and are therefore sensitive to instructional changes over short periods of time and are useful for monitoring specific deficits (See Fig. 1; Christ et al., 2014; Fuchs & Deno, 1991). A limitation of SSMs is that they do not allow for assessment of long-term retention and generalization of subskills, as teachers typically move through the hierarchy without re-testing previously mastered skills (Fuchs & Deno, 1991). A second limitation is that because the skill is narrow in scope, it may be difficult to create, standardize, and empirically test large parallel sets of probes.
In contrast, GOMs do not utilize instructional hierarchies, but contain items testing students on full-year content. Therefore, it is expected that students will gradually improve on content over the course of the year (See Fig. 1). Measures of GOM allow teachers to change their methods, materials, and sequencing of content, in addition to rapidly and accurately identifying students who are both struggling and excelling. GOMs also allow for test difficulty to remain constant, because they always sample the same content across the curriculum, in comparison to SSMs, which vary in difficulty based on specific content and must be exchanged quickly as new skills are taught. Furthermore, GOMs allow for assessment of retention and generalization because they sample skills from across the academic year.

Research suggests that GOMs in literacy are more predictive of norm-referenced tests and more useful for screening and progress monitoring over the course of a year than SSMs, although they should complement each other within a comprehensive formative assessment system (Christ et al., 2014; Fuchs & Deno, 1991). A disadvantage of using GOMs is that they are less sensitive to assessing specific skills and short-term growth (Fuchs & Deno, 1991). Therefore, when used alone, they may lack rapid sensitivity in identifying student who need help in specific curricular areas and as measures of behavioral change to small units of instruction.
To comprehensively assess mathematics skills, it is recommended that both GOM and SSM data be collected (Burns & Klingbeil, 2010). In mathematics, however, there is disagreement over what constitutes a GOM, what comprises a subskill, and when to target which subskills. Mathematics is an academic area with multiple, often intact, topics (e.g., computation, application, and problem solving; addition, subtraction, multiplication, fractions, etc.), so it has been challenging to develop GOMs to measure a single task of generalized mathematics skills (Kelley et al., 2008). Kelley and colleagues (2008) noted that creating GOMs for mathematics has been difficult, so to supplement GOMs skills-based measures (SBMs) have been developed. Table 2 illustrates the differences between the three types of CBMs. SBMs use individual or multiple skills that encompass one domain. Within mathematics, numeracy, computation, problem solving, and application are considered domains. For example, a computation SBM might include addition, subtraction, and multiplication with problems of varying difficulties. These are in contrast to SSMs because SSMs target specific skills. Skills in SBMs are isolated or marked and require less coordination of multiple skills than do GOMs (Hosp et al., 2016;
SBMs are particularly important because math curricula are structured with skills and procedures that build upon previously learned procedures and skills (National Council of Teachers of Mathematics, 2000).

<table>
<thead>
<tr>
<th>General Outcome Measures (GOM)</th>
<th>Specific Subskill Mastery Measurement (SSM)</th>
<th>Skills-Based Measure (SBM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Content consists of multiple skills, some of which students have not yet learned</td>
<td>• Content consists of one subskill that students have learned</td>
<td>• Content consists of one or multiple skills that students have learned</td>
</tr>
<tr>
<td>• Allows for assessment of retention and generalization</td>
<td>• Does not allow for assessment of retention and generalization</td>
<td>• Does not allow for assessment of retention and generalization</td>
</tr>
<tr>
<td>• Growth is expected to be slower</td>
<td>• Growth is expected to be faster</td>
<td>• Growth is expected to be faster</td>
</tr>
</tbody>
</table>

Hosp and colleagues (2007) provided an example of a sequence of subskills for a third-grade curriculum including (a) multidigit addition without regrouping; (b) multidigit addition with regrouping; (c) multidigit subtraction without regrouping; (d) multidigit subtraction with regrouping; and (e) multiplication facts, factors to nine. Each of these constructs can be measured separately. They naturally share some covariance, but nonetheless represent distinct skills that do not come together like literacy subskills do when measuring ORF.

**CBM decision-making.** CBM allows for the calculation of growth rates in determining a student’s response to intervention. Because each probe is assumed to be of equal difficulty, slopes generated can measure a student’s rate of learning (Fuchs, 2004). Using slope, instructors will know that a student is not responding to intervention if his or
her observed growth rate is less than a specified expected growth rate, also known as the aimline, which is commonly based on normative pools of data (Deno, et al., 2002). For example, in reading, an aimline of three words correct per minute per week is seen as reflective of feasible but ambitious growth in intervention (Fuchs et al., 1993). Ensuring a student is achieving or surpassing their aimline assures that the student is “closing the gap” relative to their peers (Stanovich, 1986).

This data can be used to inform decisions about modifying intervention; if a student’s growth rate is below that of their aimline, the intervention should be intensified such that they meet this expectation for growth. If subsequent increases in intervention intensity require time and resources beyond that which the general education teacher can provide, this may become a rationale for considering special education. Teachers should use the data to modify the intervention for students who do not progress and should raise goals for students who do progress to increase student achievement (Deno et al., 2002). Given the potential magnitude of these decisions, it is imperative that the measurement system being employed is technically sound.

Ardoin and colleagues (2013) have summarized specific guidelines for decision-rules, defined as the time and means by which a practitioner determines a student’s responsiveness to intervention using CBM data. CBM data can be plotted on a graph, with the horizontal axis representing time and the vertical axis representing performance, such as digits correct. Teachers and RtI implementers must decide upon a decision-rule to determine whether the student made adequate progress and decide whether the intervention should be continued or modified. There are two types of decision-rules generally used in CBM: data point decision rule and trend line decision rule. These are
compared to the aimline, which is a linear slope connecting a student’s original performance level to the desired performance level at the end of the intervention period, determined by expected growth under intervention and stability of the measure. The linear slope summarizing a student’s actual performance over time is called the rate of improvement (ROI). ROI is a way to measure student growth within CBM.

While using data point decision rules, teachers must evaluate whether a student’s data are above or below the aimline. A common guideline is that instruction be modified if three to five consecutive data points are below the student’s goal line. If three to five consecutive data points are above the goal line, the goal line should be increased, as the student has made more progress than expected. Lastly, if three to five consecutive data points are inconsistent and fall both above and below the goal line, the intervention should be continued as is (Ardoin et al., 2013; Fuchs et al., 1992; Shapiro, 2004; Stecker et al., 2005). Although previously recommended and popular, such methods have been heavily criticized in more recent research (Hintze et al., 2018; Van Norman & Christ, 2016), which is a consequence of probe variability.

A trend line decision rule involves calculating a trend line and comparing it to the aimline. This can be done by generating an ordinary least squares (OLS) trend line for the data and comparing it to the aimline. If the slope of the trend line is less than the slope of the growth line, the intervention is intensified. If the slope of the trend line is greater than the slope of the growth line, the goal should be increased. Lastly, if the two slopes are similar, the intervention should be continued (Ardoin et al., 2013; Fuchs & Fuchs, 1986). Recommendations vary regarding exactly how long to wait before
applying such decision rules such that the instructional decision is sufficiently reliable (Ardoin et al., 2013; Christ et al., 2013).

Relatedly, a model from the 1990s, developed to determine which students should be considered to have an educational disability, can be applied to determine which students are unresponsive within an RtI framework using CBM. This model can be used in an environment where most students are experiencing academic growth. Children who demonstrate a deficit in mean level of performance and rate of progress from others can be considered for special education (Fuchs & Fuchs, 1998; Speece & Case, 2001). This is termed the “dual discrepancy” model, which recognizes that both the student’s current performance level as well as the student’s level of growth are suggestive of future performance (Fuchs & Fuchs, 1998).

Using the student’s performance level alone is not useful for low-performing students who may have shown growth. Furthermore, using the student’s rate of improvement alone does not provide useful information for high-performing students who may not have shown significant growth, but are still performing highly (Fuchs, Fuchs, et al., 2003). The dual discrepancy approach judges a student’s performance as compared to others in the instructional environment and can be used to establish which students are likely to continue to fall behind and require intensified services.

Non-responsiveness is demonstrated both when a student’s level performance is inferior to that of their peers and their growth slope suggests they will not catch up without assistance. Because CBM procedures are sensitive to growth and can be used repeatedly, CBM can be used within the dual discrepancy model (Fuchs, Fuchs, et al., 2003; Speece & Case, 2001). Speece & Case (2001) used CBMs to compare students
with dual discrepancy to students with discrepant cognitive ability and reading achievement. Students with dual discrepancy performed poorer on phonological processing. However, the dual discrepancy model is not as well researched, nor discussed in practitioner-oriented tutorials, to the degree that trend line and data point decisions rules are researched. This is likely due to its complexity in practice. It is cumbersome to implement and may not be feasible in the real world (Fuchs, Fuchs, et al., 2003; Speece & Case, 2001).

As discussed above, analysis of CBM growth slopes is a common way to determine the effectiveness of instruction or intervention because such data is sensitive to student growth and treatment effects (Fuchs, Fuchs, et al., 2003). This has been demonstrated for CBM for reading (Fuchs et al., 1989; Fuchs et al., 1994; Fuchs, et al., 1991), but not for math CBM. It is important to investigate the stability of growth slopes at the individual level, as such growth is used to make high stakes decisions regarding changes to a child’s instructional environment, and slope variability has been found to be directly related to the length of required progress monitoring (Christ et al., 2013).

This has been done, for example, by calculating the standard error of slope ($SEb$), the standard error of estimate ($SEE$), and/or the Root Mean Square Errors (RMSE) for reading CBM. $SEb$ estimates the variance of slope estimates from student to student, whereas the $SEE$ and RMSE are estimates of measurement precision (Ardoin et al., 2013). Ardoin and colleagues (2013) summarized the research on accuracy of growth rates for CBM for reading and reported that the $SEb$ ranged from 0.64 to 1.27, and the $SEE$ and RMSE ranged from 7.09 to 15.97 words read correctly per minute per week, which can be described as high error rates relative to average observed performance. In
other words, reading CBM may need to occur for many weeks before an observed trend line stabilizes.

This information can be used to judge the relative quality of different, comparable CBMS, as well be used to estimate the length of time needed to make sound instructional decisions and to temper recommendations. To date, no such research exists for math CBM. Therefore, practitioners need guidance on how long to monitor progress using math CBM before making decisions and how much error to expect surrounding estimated trend lines.

**Psychometrics of CBM for mathematics.** Research on typical growth rates for CBM for mathematics is also lacking. Fuchs and colleagues (1993) investigated growth patterns in CBM for mathematics using an unspecified GOM. During the first year, the GOM was administered weekly and during the second year, at least monthly. To find the weekly rate of improvement, an OLS regression was run between scores and calendar days to calculate the calendar day slope, which was multiplied by seven days. Fuchs and colleagues (1993) found that for an unspecified CBM for math, from first to sixth grades, average growth in digits correct per minute (DCPM) per week ranged from .20 to .77 in the first year and .28 to .74 in the second year of the study. Growth increased from first to fifth grade and then decreased from fifth to sixth grade. Graney et al. (2009) also found that CBM for mathematics growth rates in third through fifth grade students increased as grade level increased.

Some researchers have suggested that growth rates with CBM for mathematics may not always be linear (Fuchs et al., 1993; Graney et al., 2009; Keller-Margulis et al., 2014), as has been observed with their reading counterparts (Nese et al., 2013). Keller-
Margulis and colleagues (2014) found differing within-grade growth rates across benchmark periods on two measures of CBM for mathematics: M-CBM probes from the Monitoring Basic Skills Progress-Math Computation series and M-CAP probes from the Monitoring Basic Skills Progress- Math Concepts and Applications series, both from aimsweb. GOMs were administered in the fall, winter, and spring. Keller-Margulis and colleagues (2014) fit latent growth models for each probe type and scaled the slopes to indicate the amount of weekly change between benchmarking periods. Tests for nonlinearity for third grade math computation probes indicated more growth during the spring semester as compared to the fall semester, whereas tests for nonlinearity for third grade math concepts and applications probes indicated more growth during the fall semester.

Graney and colleagues (2009) administered an AIMSweb GOM and an M-CBM in the fall, winter, and spring over two years. A weekly growth index was computed by dividing the difference between the fall and winter and winter and spring scores by the number of weeks between the data collection periods. A greater increase in scores was reported on the GOM from winter to spring than from fall to winter in third through fifth grade over two years, with significant differences across grade levels and a significant interaction by grade level and time. More research is needed to further investigate weekly growth slopes within CBM for mathematics for shorter timeframes more representative of the length of intervention. Specifically, existing research has examined growth rates using GOMs, but research is lacking regarding SSMs (Fuchs et al., 1993; Keller-Margulis et al., 2014). Furthermore, previous studies have inferred weekly slope using monthly or triannual data collection, rather than more valid weekly data collection.
AIMSweb provides progress monitoring tools for math facts, called M-COMP, which is a GOM. M-COMP is a collection of CBM probes to aid in screening and progress monitoring of student in first through eighth grade. The national field test, with a sample of 7,703 students, yielded an alternate-form reliability of .89 at third grade, meaning that probes were of adequate consistency (AIMSweb Technical Manual, 2012). The standard error of measurement (SEM) is the variability of the measurement error that is related to the test scores of a group of examinees at a single point in time and the measure’s reliability (Harvill, 1991). The SEM for M-COMP was 5.80 at third grade (AIMSweb Technical Manual, 2012).

AIMSweb provides average student ROI based on a student’s initial performance as follows: Very Low (1st-10th percentile), Low (11th-25th percentile), Average (26th-75th percentile), High (76th-90th percentile), and Very High (91st-99th percentile). If a third-grade student performs at the 45th percentile (Average) in the fall, his or her fall-spring ROI of 1.10 corresponds to the 75th percentile (AIMSweb ROI Growth Norms, 2012). The reliability of the ROI of M-COMP scores was .75 in third grade. The SEM of the ROI of M-COMP scores was .29 in third grade. The students completed the progress monitoring probes two to three times per month (AIMSweb Technical Manual, 2012). These statistics demonstrate the utility of the AIMSweb probes for progress monitoring. However, there is a need for independent researchers to document the metrics of M-COMP in order to provide unbiased estimates and supplement the publisher’s. Additionally, research is needed on actual weekly growth rates for M-COMP.
Purpose of the Study

The purpose of the current study is to investigate weekly growth rates for third-grade students on an SSM containing multiplication problems, an SBM containing both addition and subtraction problems, and a GOM for CBM for mathematics. Although growth rates are established for CBM for reading, CBM for mathematics is lacking. The existing research on ROI for CBM for mathematics has focused on GOMs, but there is a need to study M-COMP to supplement information provided by the publisher. Additionally, previous research on the topic largely ignores SBMs, despite their frequent use in the published literature (e.g., Codding et al., 2007; Schutte et al., 2015) and popularity in recommend models of math RtI (Codding et al., 2017). The current study will add to the research-base on M-COMP, an example of a math GOM, and will also examine growth rates for SSMs and SBMs in math computation. While previous research has used statistical and mathematical procedures to estimate weekly growth rates in CBM for mathematics (e.g., from tri-annual screening data), the current study will use actual weekly student performance, thereby increasing the validity of estimates.

Research Questions

This investigation will utilize archival data to determine the expected weekly growth rates and error within various math CBMs for third graders. Specific research questions that will be addressed include:

1) What is the precision of weekly growth slopes in CBM for mathematics for third graders for both common subskills (i.e., the SSM and SBM) and for a common math GOM (i.e., M-COMP), as measured by common indicators of slope variability (i.e., standard error)?
2) What are typical weekly growth slopes in CBM for mathematics for third graders for the SSM, SBM, and M-COMP, and how do such slopes vary across students and across classrooms?

3) How does initial CBM performance predict slopes over time?
Chapter Three: Methodology

Overview

This will be an archival examination of data collected in a public elementary school by a team of graduate students. Background information about the study and its participants will be provided, as well as methods and procedures used in the current study.

Original Study Participants and Setting

Participants included 102 students from five third-grade classrooms in one elementary school located in the Northeast. Students in kindergarten through sixth grade attend the school. The most current demographic data was obtained based on the 2016-2017 school year (New York State Education Department, 2017). Fifty-three percent of students in the school were male. In terms of economic demographics, 66% of students from the school were eligible for free lunch and 8% were eligible for reduced-price lunch. Additionally, 76% were considered “economically disadvantaged” by the New York State Education Department. Four percent of students were considered English Language Learners and 14% were students with disabilities. In terms of proficiency, 38% of all third-grade students were proficient in English Language Arts based on state testing, as compared to 43% statewide. In mathematics, 43% of third-grade students in the school were proficient based on state testing, as compared to 48% statewide (New York State Education Department, 2017). Table 2 illustrates the percentage of students by ethnicity. The majority of the students were White.
The study was conducted during the fall semester of 2017. A team of graduate students group administered the mathematics assessment measures in the classrooms twice per week, on Monday and Thursday. **

**Instrumentation**

**Subskill Probes.** Teachers were consulted in September as to what math skills would be most useful to assess for growth, given the grade’s current position in the curriculum. Based on teacher input, experimenter-constructed two number three-digit by three-digit addition and subtraction probes (with regrouping) were administered twice a week for six weeks. These probes, which contained both addition and subtraction problems, were not considered SSMs, but rather SBMs, because they assessed more than one skill. The three-digit addition and subtraction probes were randomly generated through an Excel spreadsheet program and consisted of 56 problems in a four by six array, sampling from the entire universe of possible problems with possible solutions. Problems were presented in black Ariel 16point font on white paper. Addition and subtraction problems were presented in traditional vertical format (see Appendix A).

After teachers indicated they no longer were actively working on 3x3 addition and subtraction, and data previewing indicated the likely leveling of slopes and high measurement error, single-digit multiplication probes, which included multiples of two

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>49</td>
</tr>
<tr>
<td>Black or African American</td>
<td>15</td>
</tr>
<tr>
<td>Multiracial</td>
<td>15</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>14</td>
</tr>
<tr>
<td>Asian or Native Hawaiian/Other Pacific Islander</td>
<td>7</td>
</tr>
</tbody>
</table>
through nine, were introduced and used twice a week for four weeks (See Appendix B; Poncy & Duhon, 2017). As these probes contained only one type of problem, they were considered SSMs. The subskill assessment probes were group-administered twice per week and required students to complete as many math facts problems as possible in two minutes. Probes were scored based on both digits correct and problems correct per minute.

**AIMSweb Mathematics Computation (M-COMP).** M-COMP was used as a GOM and was group-administered once per week, typically on Thursdays. It required students to complete mathematics computation problems for eight minutes. M-COMP includes 33 equivalent probes for third graders with 37 problems per probe of mixed skill and difficulty. M-COMP is considered a “capstone” CBM that samples math knowledge broadly and is theorized to grow linearly and slowly throughout the school year, reflecting cumulative knowledge (AIMSweb Technical Manual, 2012). One probe was randomly assigned to each class each week and probes were counterbalanced so that each classroom completed a different probe each session. Probes were administered once a week because less growth per session was expected and the requisite session time was longer (8 minutes). Correct answers were awarded one, two, or three points based on AIMSweb weighted scoring procedures and the total number of points earned was calculated.

M-COMP has been demonstrated to have a median alternate-form reliability of .88 across grades one through eight and a mean reliability of .89 at third grade. It also has been reported to have a criterion validity of .73 in third grade as compared to the Group Mathematics Assessment and Diagnostic Evaluation (G-MADE; AIMSweb
Technical Manual, 2012). The test publishers for M-COMP reported the reliability and SEM of the ROI by finding the split-half correlation between ROIs. They adjusted using the Spearman-Brown formula (AIMSweb Technical Manual, 2012). Additional psychometrics regarding slope were reported in Chapter 2.

**Institutional Review Board**

An Institutional Review Board (IRB) exemption was obtained for the examination of an extant database. The database was generated as part of broader instructional consultative work between the school and a faculty member. Prior to data collection, passive assent forms were sent to the parents and guardians of all participants and no adults declined their child’s participation.

**Procedure**

An effort was made to conduct assessments twice per academic week on a regular schedule. During the first session each week, typically a Monday, students were administered a subskill measure (SBM or SSM). During the second session each week, commonly a Thursday, students were administered both the subskill measure and the GOM (M-COMP). Therefore, the subskill measure was administered twice weekly and the GOM was administered weekly. Directions adapted from M-CBM standard directions were read to students prior to administration of the subskill measure (See Appendix C). AIMSweb M-COMP instructions were read to students prior to administration of the M-COMP (AIMSweb Math Facts Testing Directions, 2002). After data collection each week, scores were inputted into a computer program and sent to teachers electronically. That is, the data was used so that teachers could adapt instruction.
The paper copies of the probes were also scanned and sent electronically to teachers weekly, allowing them to engage in error analysis.

**Data Analysis**

The purpose of this investigation was to derive basic properties of math CBM weekly growth slopes that would be relevant to instructional decision-making within an RtI system. This includes the $\text{SEE}$, $\text{SEb}$, and average rate of growth for the three CBMs. Both digits correct, and problems correct per minute were scored. For the purpose of analysis, both measures were treated as continuous variables with a true zero. There were no determined consequences of making this assumption, as the range of values was wide.

Descriptive statistics for these measures have been studied. VanDerHeyden and Burns (2005) measured digits correct over four months on experimenter-created single and mixed skill CBM for mathematics probes for third graders. Their work revealed a mean of 27.9 digits correct in January with a standard deviation of 9.9, 37.9 digits correct in February with a standard deviation of 13.4, 35.3 digits correct in March with a standard deviation of 13.0, and 38.8 digits correct in April with a standard deviation of 12.3. Burns and colleagues (2006) administered an experimenter-created SSM over four weeks and found a mean slope of 2.63 digits correct with a standard deviation of 1.60 for third graders. Research by Fuchs and Fuchs (1993) examined slope rates in students at the elementary level using CBM for mathematics probes. They found a mean score of 28.92 with a standard deviation of 2.92 for third graders for both digits correct and problems correct. The mean slope was .42 for digits correct and 0.31 for problems correct. The kurtosis of the distribution of slopes was -1.17 for digits correct and -0.19
for problems correct and the skewness of the distribution was 0.11 for digits correct and 0.27 for problems correct.

Growth rates with CBM for mathematics have not always been found to be linear. For example, some growth rates have demonstrated quadratic trends (Fuchs et al., 1993). Despite this, researchers and publishers have described growth rates in CBM for mathematics as linear by noting growth as digits of improvement per week. They do this by subtracting the student’s first score from the student’s last score and dividing that by the number of weeks that have passed. Researchers have cautioned against making assumptions of linear growth throughout the year; therefore, no assumptions were made about linearity prior to analysis (Fuchs et al., 1993; Graney et al., 2009; Keller-Margulis et al., 2014).

**Calculation of the SEE.** A regression was conducted on each student and the OLS method – the most common procedure in both research and applied work – was used to calculate the SEE as a measure of reliability. OLS attempts to minimize the sum of the squares of the differences between each student’s performance and the performance predicted by the regression (Cohen, et al., 2003); the SEE being the average absolute value of the residuals. Distributional information from the calculation of the SEE for each student is reported in a graph. Outliers were reported. The SEE was calculated using the formula:

\[ SEE = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-2}}. \]

The average SEE, weighted by number of observations, and its standard deviation was reported. The SEE is useful because it is commonly reported in the CBM literature and provides a sense of the plausible range of observed values in slopes.
Calculation of the $SEb$, growth slopes, and variance. Data was imported into the computer software program *Hierarchical Linear and Nonlinear Modeling* 7 (Raudenbush et al., 2011). Multilevel modeling, referred henceforth as hierarchical linear modeling (HLM), was used because the data is structured by groups of units, clustered together. In other words, it was anticipated that individual observations within students covary together to some degree, mathematically quantified as the intra-class correlation ($ICC$). Furthermore, the scores of individual students nested within classrooms may also bear similarity. Hierarchical linear modeling is a means by which to account for these nesting effects when estimating models. The observation points were considered level-1, individual students were considered level-2, and classrooms were considered level-3 (Raudenbush & Bryk, 2002; Singer & Willett, 2003).

There are two ways to analyze hierarchical data without using HLM, but these methods have limitations. One way is to disaggregate the data and use the number of lower-level units to assess the impact of higher-level units, but this does not account for the shared variance between variables. The other way is to model relationships at the aggregate level of analysis, but this does not acknowledge any lower-level variance that may contribute to the outcome measure (Hofmann, 1997).

HLM overcomes the limitations of the disaggregate and aggregate methods of analysis because it models both within and between group variance simultaneously and investigates the influence of groups on individual outcomes (Hofmann, 1997). HLM is a complex form of the generalized linear model that accounts for the variance that is shared between the variables at different levels within the hierarchy of data (Woltman et al., 2012). In this way, HLM controls for the violation of the independence assumption,
which necessarily occurs when using OLS on time-series data (Raudenbush & Bryk, 2002). There is dependency between the error terms, but it is uncommon in the literature to control for dependency and autocorrelation and it is never done within an applied setting.

A common result of this violation is the underestimation of the standard error (Raudenbush & Bryk, 2002). HLM models individual and group level residuals separately; in OLS they are not estimated separately. Therefore, it allows for investigation of relationships both within a hierarchical level and across levels (Hofmann, 1997). It uses these regression relationships to identify the relationships between predictor and outcome variables. There are limitations to HLM, which include that it requires a larger sample size to maintain adequate power and it typically removes groups with missing data at higher levels (Woltman et al., 2012). The present study had five classes in level-3 of the model, which is insufficient and a limitation of the current study (Moeyaert et al., 2013).

The data was run for all students as one group. Major assumptions of the respective models were checked. Linearity was assumed, as this is common in the field (Christ, 2006; Christ, et al., 2014). Fixed and random effects for intercept ($\beta_0$) and slope ($\beta_1$) were reported for each model. Fixed effects for intercept and slope are the aggregate intercept and slope for the participants in a group, whereas random effects are estimates of variance across students, accounted for by the clustering at each level (e.g., level 1, 2, and 3; Van Norman, et al., 2018). Fixed effects and random effects intercepts and slopes were reported for all three groups. This analysis was used to derive the average rate of
improvement for the CBMs, reliability, as well as speak to the variability of these ROIs across students and classrooms. The three-level model was defined as

Level-1:

$$SCORE_{ij} = \pi_{0i} + \pi_{1i}(OCCASION_{ij}) + e_{ti}$$

Level-2:

$$\pi_{0i} = \beta_{00} + u_{0i}$$

$$\pi_{1i} = \beta_{10} + u_{1i}$$

Level-3:

$$\beta_{0j} = \gamma_{000} + u_{0j}$$

$$\beta_{1j} = \gamma_{100} + u_{1j}$$

where $SCORE_{ij}$ is an individual CBM score for individual, $i$, at each occasion, $t$, in classroom $j$. $\pi_{0i}$ is the initial performance (intercept) for each student and $\pi_{1i}$ represents the growth rate per week, or the slope. $e_{ti}$ represents unaccounted for residual error. $\beta_{00}$ and $\beta_{10}$ represent fixed effects, which are permitted to vary across individuals, whereas random effects are represented by $u_{0i}$ and $u_{1i}$. In turn, individual level effects, $\gamma_{000}$ and $\gamma_{100}$, are permitted to vary by classroom, represented by $u_{0j}$ and $u_{1j}$. The full maximum likelihood method of estimation was used and ICC was reported.

**Interrater Reliability**

Experimenters were trained in administration and scoring procedures during one hour-long session. Interrater agreement was assessed by comparing the total number of digits/problems/points the examiner scored as correct to the total number of digits/problems/points an independent rater scored as correct on each measure for the session for 20% of sampled probes. The number of digits/problems/points that both the
examiner and rater record as correct was divided by the total number of probes examined and multiplied by 100.
Chapter Four: Results

Overview

Analyses were conducted to find the average rate of growth, as well as the distribution of slopes and magnitude of the residuals (SEE) for all three probe types, including the SSM, SBM, and the GOM. Assumptions of the models were reviewed. Approximately 9.5% of the data was missing. Interrater reliability based on 20% of randomly selected probes ranged from 97 to 99%. The distribution of the residuals at level 1, 2 and 3 were reviewed as well. Review of the standardized $\tau$ matrix indicated that the level 2 random effects were not highly correlated. They ranged from -0.10 to 0.27 across DCPM and PC for the SSM and SBM.

Precision of Slope - SEE

The OLS method was used to calculate the SEE. As the OLS method is sensitive to outliers, the average and median number of residuals was calculated. For the SBM, the average number of residuals were 0.67 for DCPM and 0.68 for PC. For the SSM, the average residuals were 0.65 for DCPM and 0.56 for PC. The median was 1 for DCPM and PC for both measures. For the GOM, the average number of residuals was 0.49 and the median was 0. Overall, there were very few outliers that would indicate an assumption violation.

SBM. The SEE for individual students were calculated in R (R Core Team, 2017) with the plyr package (Wickham, 2011), which were then aggregated to provide insight into overall group trends to demonstrate results between the three CBMs. A smaller $SEE$ is more desirable than a larger one, as a smaller $SEE$ is indicative of a more precise measure with less error about the observed slope (Christ, 2006). For the SBM, which
consisted of addition and subtraction problems, the average $SEE$ was 5.74 DCPM and the median was 5.06 (See Table 3). Individual $SEE$s varied widely, $MIN = .94$, $MAX = 12.47$, $SD = 2.60$. A 95% confidence interval (CI) for the average $SEE$ estimate would be [5.19, 6.29]. The average slope was -0.22 DCPM/wk, indicating near zero growth over time among all participants, and a large difference of 5.96 when compared to the mean $SEE$ of 5.74, suggesting a wide residual distribution relative to growth. To put this in perspective, at an average slope of -0.22 DCPM, an 80% CI, which would be a more realistic confidence band for applied practice, would be [-0.92, 0.48]. An 80% CI is considered more realistic, as 90% is not needed when making low-stakes decisions.

The average $SEE$ was 3.57 when growth was measured as PC and the median was 2.48, with individual estimates ranging, $MIN = 0.82$, $MAX = 11.45$, $SD = 2.76$. A 95% CI for this estimate is [2.99, 4.15]. The average slope was 0.20 PC. The difference between the average $SEE$ and the slope was 3.37, which again suggests relatively large residuals about the trend line. To put this in perspective, at an average slope of 0.20 PC per week, an 80% CI would be [-0.29, 0.69].

**SSM.** The average $SEE$ for the SSM, consisting of multiplication problems, was 6.49 DCPM, with a median of 5.85. Individual estimates ranged from a minimum of 0.76 to a maximum of 13.27 and the $SD$ was 2.84. A 95% CI for this estimate is [5.89, 7.09]. The average slope was 3.50 DCPM/wk with a difference of 2.99 from the mean $SEE$, again suggesting large variability about the trend line, but more stability than observed with the SBM. To put this in perspective, at an average slope of 3.50 DCPM/wk, an 80% CI would be [2.87, 4.13].
For PC, the average SEE was 3.57 and the median was 3.55. Individual estimates ranged from a minimum of 0.45 to a maximum of 7.00 and the SD was 1.52. A 95% CI for this estimate was [3.25, 3.89]. The slope was 1.89 PC/wk and the difference between the slope and mean SEE was 1.68. To put this in perspective, at an average slope of 1.89 PC, an 80% CI would be [1.56, 2.22].

**GOM (M-COMP).** Lastly, for M-COMP, the GOM, the average SEE was 7.60 points with a median of 6.63. Individual SEE ranged, $MIN = 1.90$, $MAX = 19.60$, $SD = 4.12$. A 95% CI for this estimate was [6.73, 8.47]. The average slope was 0.68 points/wk and there was a difference of 6.92 between the average slope and the mean SEE, which is large.

<table>
<thead>
<tr>
<th>Probe Set</th>
<th>Metric</th>
<th>$SEE M$</th>
<th>$SD$</th>
<th>Slope $M$</th>
<th>% $M$ of the Slope $M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSM</td>
<td>DCPM</td>
<td>6.49</td>
<td>2.84</td>
<td>3.50</td>
<td>185% larger</td>
</tr>
<tr>
<td></td>
<td>PCPM</td>
<td>3.57</td>
<td>1.52</td>
<td>1.89</td>
<td>189% larger</td>
</tr>
<tr>
<td>SBM</td>
<td>DCPM</td>
<td>5.74</td>
<td>2.60</td>
<td>-0.22</td>
<td>2609% larger</td>
</tr>
<tr>
<td></td>
<td>PCPM</td>
<td>3.57</td>
<td>2.76</td>
<td>0.20</td>
<td>1785% larger</td>
</tr>
<tr>
<td>GOM</td>
<td>PC</td>
<td>7.60</td>
<td>4.12</td>
<td>0.68</td>
<td>1118% larger</td>
</tr>
</tbody>
</table>

Overall, individual estimates of the SEE varied widely across participants and between the SSM, SBM, and GOM, suggesting that the SEE is not a “fixed” value across students who engage in progress monitoring. A sample participant’s slope and $SEE$ values across all CBM types are presented below in Figure 2. The $SEEs$ ranged from moderate to high relative to the slope values presented below. Implications for these values are expanded upon in the discussion.
Figure 2. Sample Student DCPM by SEE and Slope. This figure shows the SEE and slope of one sample student in DCPM for all three probe types.

HLM

For DCPM, the hierarchical three-level linear model consisted of $N_{level1} = 557$ observations, $N_{level2} = 86$ students, and $N_{level3} = 5$ classes. At level-1, 45 individual observations were missing (7.48%) and results were calculated using listwise deletion.

The ICC were calculated at the student level (level-2) and classroom level (level-3). ICC for the SBM-DCPM was 0.49 and 0.56-PC at level-2 and 0.17 and 0.77 at level-3. ICC for both the SSM-DCPM and SSM-PC was 0.60 at level-2 and 0.00 and 0.01 at level-3. Lastly, ICC for the GOM was 0.65 at level-2 and 0.04 at level-3. Overall, scores varied and were strongly attributed to students, but not as much by teachers.

SBM. Results for the SBM are reported first. The error variance of the model for DCPM, $\sigma^2$, was 40.24. The reliability estimate for the random level-1 coefficient was .65 for the intercept and .003 for time, $\pi_j$. The reliability of the random level-2 coefficients was .58 for the intercept and .75 for the slope estimate. The estimation of the fixed effect
for the intercept, or the aggregate intercept across participants was significant, \( r_0 = 17.97 \) DCPM, \( t(4) = 13.84, p < 0.01 \). The fixed effect for slope was not significant. The final estimation of random effect for intercept, or the estimate of variance contributed by the clustering at level-2, was significant, \( r_0 = 40.02, p < .001 \) (See Table 4). This suggests that student’s initial performance varied by student. The random effect for slope, \( r_1 = 0.02 \), was small, and was not significant. Within one SD (68% of all scores), initial performance scores fall within the range of 11.64 to 17.97 + 24.30 DCPM. Regarding the variance-covariance matrices, at level-2, the standard deviation was 6.33 DCPM, which is large.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Variance Component</th>
<th>( \chi^2 )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept, ( r_0 )</td>
<td>40.02</td>
<td>236.52</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Time, ( r_1 )</td>
<td>0.02</td>
<td>69.23</td>
<td>&gt; 0.500</td>
</tr>
</tbody>
</table>

At level-3, the standard deviation was 2.22 DCPM, indicating that teachers varied on average 2.22 DCPM at their intercept. This is a small difference across classrooms. The estimation of the random effect for intercept at level-3 was \( u_{00} = 4.91 \) and was significant, \( \beta_{00} = 2.22, p = .018 \) (See Table 5). The random effect for slope was also significant, \( \beta_{10} = 1.12, p < .001 \). This indicates that both the initial performance and rate of improvement over time for DCPM varied significantly across classrooms.
For PC, the three-level linear model consisted of $N_{level1} = 558$ observations, $N_{level2} = 86$ students, and $N_{level3} = 5$ classes. At level-1, 44 cases were missing (7.31%) and results were calculated using listwise deletion.

The error variance of the model, $\sigma^2$, was 18.99. The reliability estimate for the random level-1 coefficient was .76 for the intercept and .057 for time, $\pi_1$. The reliability of the random level-2 coefficients was .96 for the intercept and .75 for the slope estimate. The estimation of the fixed effect for the intercept, or the aggregate intercept across participants, was not significant. The fixed effect for slope was also not significant. The final estimation of random effect for intercept, or the estimate of variance accounted for by the clustering at level-2, was significant, $r_0 = 31.88, p < .001$ (See Table 6). This suggests that the initial performance level of students, like the DCPM metric discussed previously, varied by student. The random effect for slope, $r_1 = 0.18$, was small, and was not significant. Regarding the variance-covariance matrices, at level-2, the standard deviation was 5.65 PC, which is a large difference.

### Table 5

*Results of Multilevel Analyses for Level-3 Random Effects for SBM for DCPM*

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>SD</th>
<th>Variance Component</th>
<th>Df</th>
<th>$\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept, $\beta_{00}$</td>
<td>2.22</td>
<td>4.91</td>
<td>4</td>
<td>11.94</td>
<td>0.018</td>
</tr>
<tr>
<td>Time/Intercept, $\beta_{10}$</td>
<td>1.06</td>
<td>1.12</td>
<td>4</td>
<td>20.61</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

### Table 6

*Results of Multilevel Analyses for Level-2 Random Effects for SBM for PC*

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Variance Component</th>
<th>$\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept, $r_0$</td>
<td>31.88</td>
<td>347.28</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Time, $r_1$</td>
<td>0.18</td>
<td>68.30</td>
<td>&gt;0.500</td>
</tr>
</tbody>
</table>
The standard deviation at level-3 was 7.30 PC, indicating that teachers varied on average 7.30 DCPM at their intercept. This is a large difference across classrooms. The estimation of the random effect for intercept at level-3 was $\beta_{00} = 53.32$ and was significant, $\beta_{00} = 7.30, p < .001$ (See Table 7). The random effect for slope was also significant, $\beta_{10} = 0.54, p < .001$. This indicates that the rate of improvement over time for PC, like that of DCPM, varied significantly across classrooms.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>SD</th>
<th>Variance Component</th>
<th>Df</th>
<th>$\chi^2$</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept, $\beta_{00}$</td>
<td>7.30</td>
<td>53.32</td>
<td>4</td>
<td>115.10</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Time/Intercept, $\beta_{10}$</td>
<td>0.73</td>
<td>0.54</td>
<td>4</td>
<td>20.03</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

SSM. Results from the SSM are reported next. The error variance of the model for DCPM, $\sigma^2$, was 49.82. The reliability estimate for the random level-1 coefficient was .76 for the intercept and .64 for the slope, time, $\pi_f$. The reliability of the random level-2 coefficient was .03 for the intercept and .68 for the slope estimate. The estimation of the fixed effect for the intercept, or the aggregate intercept across participants, was significant, $r_0 = 20.93$ DCPM, $t(87) = 19.71, p < 0.001$. The fixed effect for slope was also significant, $r_1 = 3.50$ DCPM, $t(87) = 7.14, p < 0.001$. The final estimation of random effect for intercept, or the estimate of variance contributed by the clustering at level-2, was significant, $r_0 = 73.78, p < .001$ (See Table 8). This suggests that student’s initial performance varied by student. The random effect for slope was significant, $r_{1} = 10.93, p < .001$. Within one SD (68% of all scores), initial performance scores fall within the
range of 12.27 to 29.59. Regarding the variance-covariance matrices, at level-2, the standard deviation was 8.59 DCPM, which is large.

Table 8
Results of Multilevel Analyses for Level-2 Random Effects for SSM for DCPM

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Variance Component</th>
<th>( \chi^2 )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept, ( r_0 )</td>
<td>73.78</td>
<td>355.95</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Time, ( r_1 )</td>
<td>10.93</td>
<td>201.92</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

The standard deviation at level-3 was 0.44 DCPM, a small difference across classrooms. The estimation of the random effect for intercept at level-3 was \( \beta_{00} = 0.19 \) and was not significant, \( \beta_{00} = 0.44, p > .500 \) (See Table 9). The random effect for slope was significant, \( \beta_{10} = 2.24, p = 0.004 \). This indicates that while the initial performance did not vary, the rate of improvement over time for DCPM varied significantly across classrooms.

Table 9
Variance Components for Level-3 Random Effects for SSM for DCPM

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>SD</th>
<th>Variance Component</th>
<th>( Df )</th>
<th>( \chi^2 )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept, ( \beta_{00} )</td>
<td>0.44</td>
<td>0.19</td>
<td>4</td>
<td>2.45</td>
<td>&gt;.500</td>
</tr>
<tr>
<td>Time/Intercept, ( \beta_{10} )</td>
<td>1.50</td>
<td>2.24</td>
<td>4</td>
<td>15.53</td>
<td>0.004</td>
</tr>
</tbody>
</table>

For PC, the three-level linear model consisted of \( N_{level1} = 563 \) observations, \( N_{level2} = 88 \) students, and \( N_{level3} = 5 \) classes. At level-1, 53 cases were missing (9.41%) and results were calculated using listwise deletion.

The error variance of the model, \( \sigma^2 \), was 14.89. The reliability estimate for the random level-1 coefficient was .71 for the intercept and .058 for time, \( \pi_t \). The reliability of the random level-2 coefficients was .02 for the intercept and .67 for the slope estimate. The estimation of the fixed effect for the intercept, or the aggregate intercept across
participants, was significant, $r_0 = 11.01$ PCPM, $t(87) = 19.22, p < 0.001$. The fixed effect for slope was also significant, $r_1 = 1.89$ PCPM, $t(87) = 7.23, p < 0.001$. The final estimation of random effect for intercept, or the estimate of variance accounted for by the clustering at level-2, was significant, $r_0 = 20.45, p < .001$ (See Table 10). The random effect for slope was also significant, $r_1 = 2.77, p < .001$. Regarding the variance-covariance matrices, at level-2, the standard deviation was 4.52 PC, which is small.

### Table 10

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Variance Component</th>
<th>$\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept, $r_0$</td>
<td>20.45</td>
<td>314.61</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Time, $r_1$</td>
<td>2.77</td>
<td>173.59</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

The standard deviation at level-3 was 0.17 PC, which is a small difference across classrooms. The estimation of the random effect for intercept at level-3 was $\beta_{00} = 0.03$ and was not significant, $\beta_{00} = 0.17, p > .500$ (See Table 11). The random effect for slope was significant, $\beta_{10} = 0.62, p = 0.004$. This indicates that the rate of improvement over time for PC, like that of DCPM, varied significantly across classrooms.

### Table 11

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>SD</th>
<th>Variance Component</th>
<th>Df</th>
<th>$\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept, $\beta_{00}$</td>
<td>0.17</td>
<td>0.03</td>
<td>4</td>
<td>2.14</td>
<td>&gt;.500</td>
</tr>
<tr>
<td>Time/Intercept, $\beta_{10}$</td>
<td>0.79</td>
<td>0.62</td>
<td>4</td>
<td>15.67</td>
<td>0.004</td>
</tr>
</tbody>
</table>

**GOM.** The GOM are reported last. A three-level model was run. The error variance of the three-level model for PC, $\sigma^2$, was 77.25. The reliability estimate for the
random level-1 coefficient was .76 for the intercept and .008 for time, $\pi_t$. The estimation of the fixed effect for the intercept was significant, $r_0 = 41.96$ PC, $t(4) = 24.54$, $p < .001$. The fixed effect for slope was not significant. The final estimation of random effect for intercept, or the estimate of variance contributed by the clustering at level-2, was significant, $r_0 = 140.16$, $p < .001$ (See Table 12). This suggests that student’s initial performance varied by student. The random effect for slope, $r_1 = 0.02$, was small, and was not significant. Within one SD (68% of all scores), initial performance scores fall within the range of 30.12 to 53.80. Regarding the variance-covariance matrices, at level-2, the standard deviation was 11.84 points, which is a large difference.

### Table 12

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Variance Component</th>
<th>$\chi^2$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept, $r_0$</td>
<td>140.16</td>
<td>359.62</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Time, $r_1$</td>
<td>0.02</td>
<td>80.84</td>
<td>0.453</td>
</tr>
</tbody>
</table>

At level-3, the standard deviation was 2.04, which is a small difference across classrooms. The estimation of the random effect for intercept at level-3 was $\beta_{00} = 4.15$ and was not significant. The random effect for slope was $\beta_{10} = 1.12$ and was not significant either. This indicates that the initial performance and rate of improvement over time for PC did not vary significantly across classrooms.

**Summary**

Overall, results investigated the relative error and growth for all three CBM measures. For the SBM, there was no growth in DCPM over time and a wide residual distribution relative to growth with an average $SEE$ of 5.74 DCPM. For PC, the average slope was slightly larger at .20 PC/wk, but still yielded a wide residual distribution with
an average $SEE$ of 3.57 PC. The rate of improvement for the SBM varied significantly across classrooms for DCPM and PC, unlike the other CBM measures.

Regarding the GOM, the average slope was 0.68 points/wk and the average $SEE$ was 7.60 points, also indicating a wide residual distribution about the trendline. Results were more precise for the SSM. The average slope for the SSM was 3.50 DCPM/wk with an average $SEE$ of 6.49 DCPM, and 1.89 PC/wk with an average $SEE$ of 3.57. The SSM results indicated similar variability about the trendline as the other two CBM measures but was more precise.

The HLM results indicated that initial performance on the SBM and SSM varied by individual student as measured by DCPM and PC. Additionally, the initial performance and rate of improvement over time varied significantly across classroom as measured by DCPM and PC. Regarding the GOM, the 3-level model demonstrated that student initial performance varied by student as well. The initial performance and rate of improvement over time did not vary significantly across classrooms for the GOM.
Chapter Five: Discussion

The purpose of the current study was to examine expected weekly growth rates and error rates within three CBMs given to third grade students to provide data to influence instructional decisions. These findings inform how many weeks one may need to progress monitor such that instructional decisions are reliable and offers insight into the quality of these commonly used measures as well as expected rates of improvement. These properties of single-skill probes, and M-CBM more generally, are not well understood, and research in reading CBM vastly outpaces research for math assessment. The OLS method was used to calculate the SEE for the measures, which is a common indicator of reliability within the CBM field (e.g., Christ, 2006). SEEs were calculated to investigate the error for individuals, replicating the analysis that would be used in practice. The average rate of growth and SEₜ were calculated using HLM to account for the clustered levels of data: observation points within students, and within classrooms. Linearity was assumed, as it is the most practical way for teachers to functionally use data.

There was significant variability of growth slopes, as indicated by the HLM analysis, and individual SEEs, as indicated for the estimated OLS trend lines, for all CBMs across students. Some students demonstrated more precise slopes than others, making decisions about tailoring instruction to those students more reliable. Characteristics about these students are unknown. For the students with less precise slopes, teachers would need to continue progress monitoring longer and wait for the slopes to become more stable to make confident adjustments to instruction. However, this would take time and therefore would not be helpful in real time, as teachers would
not know which slopes were more or less reliable until after progress monitoring for a specified amount of time. It is also possible that these students do not follow a linear pattern and that is why they have less precise slopes.

Additionally, for the SBM, initial student performance varied among students and rate of improvement varied by classroom. This indicates that teachers had an effect on student learning. Finally, for the SSM, there was notable growth across time and more precise slopes than in CBM for reading and relative to the other two measures. Therefore, teachers should be more confident in making instructional changes based on SSM data than for the SBM or GOM. Overall, the data paints a complex picture.

Reliability not only varied by measure, but by student. Initial performance and growth over time, in contrast, would be expected to vary across students, as dictated by the three-tiered model (Codding et al., 2017; Fuchs et al., 2003).

**Average Slope Across CBMs**

**Math Skill-Based Measure.** For the SBM, average slopes were -0.22 DCPM per week and 0.20 PC per week. This means that the mean was effectively 0 digits and problems correct per minute over six weeks, a seemingly disappointing result given the amount of instruction that occurred over the course of assessment. However, it is similar when compared to previous studies using addition and subtraction problems. For example, Graney and colleagues (2009) reported a mean of weekly growth of 0.26 in the fall, although data was not administered weekly. However, this likely speaks more to the validity of the assessment tool than the effectiveness of the instruction. That is, the tool was not effectively capturing the growth that was occurring. Given this tool offered a limited scope of mixed skills, it simply may not have been sensitive enough to the nature
of discrete math skill instruction, which is supported by the contrasting results for the SSM. Based on the SBM results overall, teachers should be less confident in the precision of the SBM when considering making changes to instruction based on the data, and the use of such a measure should be carefully considered.

Math General Outcome Measure. For the GOM, the average slope was 0.68 points per week. Therefore, like the SBM, results indicated minimal to no average growth over time. There is not much existing published data to compare to this finding, so teachers would have to compare this to the ROIs in the GOM’s publisher’s literature (AIMSweb ROI Growth Norms, 2012). For third grade, an average ROI was 0.89 to 1.08 from fall to winter, 0.49 to 0.63 from winter to spring and 0.80 to 0.88 from fall to spring. Again, more than a lack of student growth, this more likely speaks to the mismatch between assessment and intervention. That is, the assessment wasn’t capturing the growth that was occurring.

Math Single-Skill Measure. For the SSM, there was considerably more growth with 3.50 DCPM per week and 1.89 PC per week over the course of the data collection. This shows this latter measure was sensitive to instruction in the classroom, whereas the other two did not appear to be so over the short duration of school monitored in this study. Single-skill measures yielded more desirable properties that would make them more valid choices for progress monitoring instruction. That is, given high levels of growth over short periods of time, it would be easier to establish a rate of improvement and determine response to intervention. However, this finding may also be partially the result of testing effects, as the limited content of the SSM was practiced repeatedly.
**Summary of Slopes of Improvement.** The current results for the SBM and GOM in math indicate that students’ growth either did not grow at all, or more likely the nature of the measure simply did not fit the nature of instruction. As these are fluency-based measures, it is possible the curriculum was mismatched, promoting conceptual knowledge and accuracy, in contrast to fluency of calculation (Coddington et al., 2017). However, it is important to note that some students did continue growing across all measures. For example, one student demonstrated growth of 5.21 DCPM and 1.74 PC on the SSM and 4.51 DCPM and 4.36 PC on the SBM throughout the study. The same student’s slope was 1.13 points per week on the GOM. During the study, it was expected that teachers continue teaching or offering fluency practice on the previous skills during the monitoring period, but it is possible that some teachers did not do this. The current study did not measure the fidelity of the teachers’ teaching practices, but some students grew, despite the possibility that teachers may not have continued teaching or offering fluency practice. This demonstrates a possibility of how educational gaps can widen over time between students within classes. In reading, this phenomenon is called the Matthews Effect, which dictates that students who grow in reading early are met with more success later, while those who have difficulty initially also have difficulty later (Stanovich, 2009).

**Precision of Slope - SEE**

Research guiding interpretation of the *SEE* for CBM for mathematics is scarce, although there are studies on this topic that focuses on CBM for reading. Researchers previously found that *SEE* for CBM for reading ranged from 7.09 to 15.97 WRCM (Ardoin et al., 2013). Others found that optimal levels of *SEE* were within 2 to 6 WRCM
The present study found wide variability across CBMs and across students. SEEs ranged from 3.57 to 7.60 across the three CBMs, the smallest representing the SSM, and the largest representing the mixed skill measures. Individual SEEs varied from .94 to 12.47 DCPM and .82 to 11.45 PC.

**SEE of the Math Skill-Based Measure.** For the SBM, the average SEE was 5.74 DCPM and 3.57 PC, which is within the range Christ (2006) found and is notably more precise than the range Ardoin & Christ (2009) reported for R-CBM. Although this suggests the SBM may be more precise than reading CBMs, it must be taken in context of the slope. That is, the error for the SBM was much higher relative to the slope (which was near zero), relative to the findings of Christ (2006). It also is important to reiterate that the SEE varied wildly by student. Given that this error of observed slope wouldn’t be known a priori, this limits the ability to confidently predict when instructional decisions with SBMs could be made.

**SEE of the General Outcome Measure.** For the GOM, the average SEE was 7.60 points, which was higher than the range for either the SBM or the SEE (Ardoin & Christ, 2009; Christ, 2006). Much like the SBM, although this value was desirable relative to findings in reading, it was much larger than the average slope of .68 points/wk. Therefore, determining whether a child has truly improved over shorter periods of time would be difficult.

**SEE of the Single-Skill Measures.** Lastly, for the SSM, the average SEE was 6.49 DCPM and 3.57 PC, which is more precise than the ranges mentioned prior for reading (Ardoin & Christ, 2009; Christ, 2006). Christ (2006) demonstrated that optimal levels of SEE fell within 2-6 WRCM. However, SEEs under 8 WRCM were not found in
the literature (Christ, 2006). Other studies yielded weekly SEE\textsubscript{s} ranging from .31 to .48 WRCM (Ardoin & Christ, 2009). The greater precision of the SSM in the current study may have been because it only contained one skill and is therefore a more reliable measure of skill fluency than a mixed skill measure, such as an SBM or GOM. Another possibility is that the students performed better due to the active instruction taking place as they completed the SSM. Teachers can be more confident in the precision of the SSM relative to the SBM and GOM and can make more confident decisions regarding altering interventions based on the SSM data.

**HLM**

HLM was used in the present study to model how students’ skills improved over time, which was a secondary analysis that provided insight on how students grow under typical curricular instruction, and how that growth may vary. HLM was used to account for and illustrate differences between the three levels of data: individual observation points, individual students, and classrooms. While the measures in the current study are inherently different and have different scales than CBM in reading, overall the weekly growth for the SBM and the GOM in the present study was shallower than previous growth estimates in reading (Ardoin & Christ, 2009; Christ, 2006).

**SBM.** For the SBM, student performance as measured by DCPM varied significantly based on random effects of slope and intercept. The standard deviation of the intercept was 6.33, where the model was centered, which is large relative to the fixed effect, 17.97. Individual student performance was wide-ranging, suggesting that students demonstrated different levels of knowledge upon entering the progress monitoring window. Specifically, teachers should be aware that within one SD (68\% of all scores for
the standard normal distribution), initial performance scores may fall between 11.64 and 24.3 DCPM. That is, upon entry into third grade, student required instruction that might fall across tiers.

Interesting, and unexpectedly, the rate of improvement over six weeks also varied significantly across classrooms, and varied more across students than across classrooms. This indicates that although teachers utilized the same curriculum and assessments, they appeared to have wide ranging effects on the performance of their students. There may be several reasons for the differences in rates of improvement by classroom related to the teacher, including teaching style, teaching experience and teacher/student fit. There also may have been differing skill levels of individual students across classrooms. For PC, similarly to DCPM, student initial performance and rate of improvement across classrooms varied significantly. This also suggests that DCPM, the more common metric and the one previously endorsed for use (Burns et al., 2006), performed similarly to PC. Problems correct is by far the easier metric to score. Given the comparable results, parsimony might dictate teachers use the latter, if they are using SBMs.

It should be noted that the archival data used in the current study was from the fall and results may have been different if the data had been from a different time of year. Teachers collecting data using SBMs should keep in mind that previous research has demonstrated that mixed computation probes for third graders demonstrate more growth during the spring semester than the fall (Keller-Margulis et al., 2014). Knowing that the SSM was a more precise measure as compared to the SBM and GOM, teachers could consider avoiding SBMs and instead consider administering multiple SSMs. For example, if teachers want to collect data on addition and subtraction fluency, they should
consider administering a separate SSM for addition and subtraction. This will increase the representative sample for individual operations, providing a purer metric of computational fluency.

SSM. For the SSM, student initial performance as measured by DCPM also varied significantly based on random effects of slope and intercept, similarly to the SBM. For the SSM, the variance component or range of slopes was higher than the SBM at 4.91 DCPM. The variance of the intercept was similarly large. Teachers should keep in mind when utilizing SSMs that their student’s slopes may vary such as in this study, from about 0 to 12 DCPM per week over six weeks, despite tier I instruction. The variability in performance may create an obstacle for teachers. They may need to review the probes and conduct error analyses to determine common errors among students. Teachers could then focus lessons on addressing those errors. Teachers could also differentiate instruction by breaking students into groups based on their performances, either initially or during progress monitoring. One way to differentiate to students coming in at differing acquisition and fluency levels is to preplan class-wide explicit timing and Cover Copy Compare interventions (McLaughlin & Skinner, 1996; Van Houten, & Thompson, 1976). Teachers can use the average growth rates from the current study and compare them to metrics found in previous literature. Regardless, significant differences in performance emerged quickly and again speaks to how stark differences in math performance emerges amongst classmates when supports are not quickly provided.

Overall, despite the range of differences, the SSM demonstrated the most desirable properties of the three measures. This is likely because it measured only one skill. If CBM in mathematics progress monitoring data is utilized in relation to RtI to
determine which students qualify for more intensive supports, including special
education, then teachers and districts should heavily consider that SSMs appear to be the
most precise measure of skill acquisition in comparison to SBMs and GOMs. Although
unexamined in the current study, it also may be the case that sampling the collective set
of individual skills via SSMs may prove more efficient than their briefer sampling in the
framework of an SBM or GOM. That is, because single skills require less time to achieve
stability, both as individual measures and when used for progress monitoring, it may
make sense to rely on multiple SSMs rather than an individual SBM or GOM.

GOM. For the GOM based on the two-level model, student initial performance
as measured by PC varied significantly, similarly to the SBM and SSM. However, based
on the HLM results for the three-level model, initial student performance and rate of
improvement did not vary significantly across classrooms, in contrast to the SBM.
Further, estimates of slope indicated teachers did not have a significant effect on student
outcome over time, as measured by the GOM. It is possible that more positive results
were not seen due to the short-term nature of the present study. Had data been collected
over the entire school year, more useful differences may have been uncovered.

Although use of GOM has occasionally been discussed as means to progress
monitor overall growth towards grade level math goals, its wide sampling of skills likely
makes it inappropriate as a measure of short-term progress monitoring, especially if
intervention focuses on one computational skill at a time, which is the recommend
approach (Codding et al., 2017; Foegan et al., 2007). Differences may also have existed if
the data was collected during different times of the year as well. For example, past
research has shown that GOMs demonstrate more growth from winter to spring than fall
to winter (Graney et al., 2009), although this research would suggest the very shallow growth observed might be even flatter, at least on average.

**Limitations and Future Directions**

The results of the current study should be interpreted with consideration of several limitations. The study used an AIMSweb GOM, therefore the results are not readily generalizable to other existing GOMs, including Fastbridge CBMmath Automaticity, CBMmath Concepts and Applications (CAP), and CBMmath Process. CBMmath CAP is a computer-administered and scored GOM, in contrast to M-COMP (Brown, 2017). AIMSweb Math Concepts and Applications (M-CAP) is another GOM. It focuses on problem-solving skills as opposed to M-COMP, which focuses on computational skills. M-CAP is administered and scored similarly to M-COMP (AIMSweb Mathematics and Applications Administration and Technical Manual, 2009). Future research should investigate other published and self-generated CBMs to provide more information about slopes and precision of slopes and student growth and how they compare across probe sets.

The study results are applicable to the sample investigated, which were third grade students in an elementary school in the Northeast. Future studies can consider other samples, such as students in other geographical areas, students receiving special education services, or students receiving English as a New Language instruction. More information about the characteristics of the participants could be included. Additionally, the probes in the present study focused on a narrow range of operations, including addition, subtraction, and multiplication. A typical third-grade math curriculum covers a significantly wider range of topics, which were not included in the study. The study also
strayed from real-world conditions in that progress monitoring occurred during typical instruction, not during intervention. In practice, progress monitoring typically occurs under intervention conditions.

The current study assumed linear growth, which may be considered a limitation. Assuming linearity may impact cases where the $SEE$ was large. If the data was truly curvilinear, the residuals would be large. This is a possibility and therefore a limitation, but this would occur in practice as well.

Although teachers in the present study had access to the weekly data regarding student performance on the CBMs, the study did not include data on whether or how teachers modified their instruction based on the data. Future research should explore how teachers utilize data in making instructional decisions. Protocols for teachers to modify instruction can be created and fidelity can be studied to measure the affect the instruction has on student growth. Future studies could also incorporate teacher observation to investigate how the curriculum was taught.

**Conclusion**

Much of the existing research in CBM has focused on the psychometric properties of R-CBM and how it relates to a variety of student outcomes (e.g., Ardoin & Christ, 2009; Ardoin et al., 2013; Christ, 2006; Deno, et al., 2001; Francis et al., 2008; Good & Shinn, 1990; Hintze & Christ, 2004; Hintze et al., 1994; Parker & Tindal, 1992; Stecker et al., 2005). Both $SEE$ and HLM have been used in research on R-CBM (Shin et al., 2000; Stage, 2001). The existing M-CBM research focuses on GOMs and therefore largely ignores SSMs and SBMs, despite their common use in research and practice. The current study’s primary objective aimed to explore the reliability of slope, as has been
conducted for progress monitoring measures for reading, on three measures: an SSM with multiplication problems, an SBM with addition and subtraction problems, and M-COMP. A secondary purpose was to estimate weekly growth rates for third-grade students. Previous studies estimated weekly growth rates from tri-annual screening data, whereas in the present study weekly growth rates were calculated from probes administered once or twice weekly (Graney et al., 2009; Keller-Margulis et al., 2014). Both questions speak to how teachers can use progress-monitoring data to inform their instructional choices.

Results demonstrated that growth slopes and individual SEEs varied significantly for all CBMs and across weekly growth rates for third-grade students. This finding will make it difficult for teachers to compare their students’ growth to previous research and determine appropriate interventions based on student performance. Furthermore, results from the present study showed that student performance at pre-test in the beginning of data collection also varied for all CBMs. Students demonstrated differing levels of knowledge at the beginning of data collection, which may pose a challenge to teachers’ abilities to differentiate instruction to a wide range of student levels. In the beginning of the school year, teachers may wish to briefly review the previous year’s content prior to introducing new material to help decrease the initial gap in performance.

Results also indicated that growth was relatively slow overall, at least for the SBM. For the SBM, growth varied by classroom, indicating that teacher factors or classroom composition had an effect on student learning for the SBM. The SSM was a more reliable and precise measure in comparison to the SBM and GOM, with the most robust SEE. The SSM possessed the most desirable properties of the three measures and required less time to achieve stability. Therefore, SSMs should be considered more
confidently over SBMs or GOMs during data analysis when it comes to high-stakes decisions in schools.
References


http://dx.doi.org/10.1016/j.jsp.2012.09.004


http://dx.doi.org/10.1177/15345084070320040401


http://dx.doi.org/10.1002/pits.20353


http://dx.doi.org/10.1016/j.jsp.2007.06.003


http://dx.doi.org/10.1177/001440299105700603


instructional hierarchy. In N. G. Haring, T. C. Lovitt, M. D. Eaton, & C. L.
Hansen (Eds.), The fourth R: Research in the classroom (pp. 23-40). Columbus,
OH: Merrill.

Haring, N. G., Lovitt, T. C., Eaton, M. D., & Hansen, C. L. (1978). The fourth R:
Research in the classroom. Columbus, OH: Merril.

and practice, 10, 33-41.


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

Psychology in the Schools, 43, 45-56. http://dx.doi.org/10.1002/pits.20128

Hintze, J. M., Owen, S. V., Shapiro, E. S., & Daly, E. J. (2000). Generalizability of oral
reading fluency measures: application of G theory to curriculum-based
measurement. School Psychology Quarterly, 15, 52-68.
http://dx.doi.org/10.1037/h0088778

sensitivity of curriculum-based measurement in reading. The Journal of Special


http://dx.doi.org/10.1177/014920639702300602


Rowe, S. S., Witmer, S., Cook, E., & daCruz, K. (2014). Teachers’ attitudes about using curriculum-based measurement in reading (CBM-R) for universal screening and

http://dx.doi.org/10.1080/15377903.2014.938793


http://dx.doi.org/10.1016/j.jsp.2014.12.003


Appendix A

Sample Subskill Probe - Multiplication

Multiplication to 81

<table>
<thead>
<tr>
<th>Name:</th>
<th>Date:</th>
<th>Set</th>
<th>Probe</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

```
9 x 8  x 8  x 8  x 2  x 6  x 4  x 8  x 7
3 x 3  x 6  x 7  x 4  x 6  x 4  x 7  x 8
6 x 2  x 3  x 7  x 5  x 4  x 6  x 9  x 4
9 x 3  x 4  x 8  x 9  x 9  x 3  x 5  x 8
2 x 2  x 2  x 7  x 5  x 6  x 4  x 3  x 6
8 x 9  x 7  x 7  x 8  x 9  x 5  x 5
7 x 7  x 9  x 2  x 6  x 7  x 5  x 2  x 3
7 x 2  x 8  x 2  x 3  x 9  x 6  x 5  x 5
3 x 9  x 4  x 3  x 4  x 3  x 8  x 6  x 5
```
| Multiplication to 81 |
|---------------------|---------------------|---------------------|
| Name: _____________  | Date: _____________  | Set 1               |
| 4 x 8 8 x 7 2 x 4 3 | 6 x 4 9 x 8 8 x 5   |
| 3 x 5 8 x 8 2 x 4  | 5 x 4 3 x 8 8 x 5 7 |
| 6 x 9 7 x 3 2 x 4 8 | 2 x 9 9 x 9 9 x 7   |
| 7 x 9 5 x 6 2 x 4  | 5 x 6 6 x 6 4 x 2 8 |
| 9 x 3 9 x 7 5 x 5 4 | 6 x 9 9 x 3 x 4 x 2 |
| 2 x 3 7 x 9 4 x 9 3 | 6 x 6 8 x 8 8 x 5 5 |
| 6 x 9 9 x 4 3 x 3 6 | 8 x 5 8 x 3 x 2     |
| 7 x 8 x 6 x 8 x 6 4 | 5 x 3 x 4           |
| 9 x 6 3 x 5 x 2 7 4 | 9 x 7 x 7           |
| 6 x 6 7 x 4 2 x 3 9 | 6 x 8 x 7 x 8       |
# Appendix B

## Sample Subskill Probe - Addition & Subtraction

<table>
<thead>
<tr>
<th>Name:</th>
<th>Date:</th>
</tr>
</thead>
</table>

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>296</td>
<td>744</td>
<td>883</td>
<td>441</td>
</tr>
<tr>
<td>+ 256</td>
<td>- 583</td>
<td>+ 766</td>
<td>+ 236</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>194</td>
<td>878</td>
<td>692</td>
<td>291</td>
</tr>
<tr>
<td>+ 153</td>
<td>+ 168</td>
<td>+ 161</td>
<td>- 139</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>624</td>
<td>640</td>
<td>690</td>
<td>706</td>
</tr>
<tr>
<td>- 462</td>
<td>+ 359</td>
<td>- 169</td>
<td>+ 506</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>906</td>
<td>354</td>
<td>991</td>
<td>230</td>
</tr>
<tr>
<td>+ 333</td>
<td>+ 189</td>
<td>- 648</td>
<td>- 115</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>392</td>
<td>290</td>
<td>325</td>
<td>861</td>
</tr>
<tr>
<td>+ 211</td>
<td>- 245</td>
<td>+ 224</td>
<td>+ 610</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>486</td>
<td>913</td>
<td>774</td>
<td>409</td>
</tr>
<tr>
<td>- 447</td>
<td>- 123</td>
<td>+ 769</td>
<td>+ 182</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>371</td>
<td>596</td>
<td>798</td>
<td>638</td>
</tr>
<tr>
<td>- 324</td>
<td>- 299</td>
<td>- 635</td>
<td>- 510</td>
</tr>
</tbody>
</table>
Appendix C

Single-Skill Math Facts Standard Directions

1. Say, "We’re going to take a 2-minute math facts test. I want you to write your answers to these Addition <Subtraction> <Multiplication> <Division> problems. Look at each problem carefully before you answer it."

“When I say ‘BEGIN,’ write your answer to the FIRST problem (demonstrate by pointing) and work ACROSS the page. Then go to the next row.”

“Try to work EACH problem. If you come to one YOU REALLY DON’T KNOW HOW TO DO, put an 'X' through it and go to the next one.”

“If you finish the first side, turn it over and continue working. Are there any questions?” (Pause)

2. Answer any questions the students may have, then hand them their probes.

3. Say, “Here are your tests. Write your name and the date on the first page only in the space provided. Do not start working until I tell you to begin.”

4. Say, "BEGIN" and start your stopwatch/timer.

5. Walk around and monitor students to ensure they are not skipping problems, are working across the page, and continue to write answers to the problems during the test time.

If a student is excessively skipping problems they should know how to do, say to the student:

Try to work EACH problem. You can do this kind of problem so don’t skip.

If a student is not working across the page, say to the student.

Work across the page. Try to work each problem in the row.

If a student stops working before the test is done, say to the student.

Keep doing the best work you can.

6. At the end of 2 minutes, say, "Stop. Put your pencils down." Monitor to ensure students stop working.
7. Collect the probes and check that each student has written his/her name and the date. If any identifying information is missing, prompt the student to complete it.