8-2014

Predicting High-Stakes Tests of Math Achievement using a Group-Administered RTI Instrument: Validating Skills Measured by the Monitoring Instructional Responsiveness: Math

Jeremy Thomas Coles
University of Tennessee - Knoxville, jcoles1@utk.edu

Recommended Citation

Coles, Jeremy Thomas, "Predicting High-Stakes Tests of Math Achievement using a Group-Administered RTI Instrument: Validating Skills Measured by the Monitoring Instructional Responsiveness: Math. " PhD diss., University of Tennessee, 2014.
https://trace.tennessee.edu/utk_graddiss/2811

To the Graduate Council:

I am submitting herewith a dissertation written by Jeremy Thomas Coles entitled "Predicting High-Stakes Tests of Math Achievement using a Group-Administered RTI Instrument: Validating Skills Measured by the Monitoring Instructional Responsiveness: Math." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in School Psychology.

R. Steve McCallum, Major Professor

We have read this dissertation and recommend its acceptance:

Sherry M. Bell, William L. Seaver, Jennifer A. Morrow, Brian E. Wilhoit

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Predicting High-Stakes Tests of Math Achievement using a Group-Administered RTI Instrument: Validating Skills Measured by the Monitoring Instructional Responsiveness: Math

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Jeremy Thomas Coles

August 2014

Dedication

To my fiancée, Mackenzie, I cannot express how important your love and support was throughout graduate school; you gave me purpose and clarity when I needed it most. To Steve, Susan, and Ashley, you are already family to me. To my parents, Brian and Valerie, you believed in me before I believed in myself. To my best friend and little brother, Braden, now it’s your turn to follow your dreams.

Acknowledgements

This dissertation would not have been possible without the help of Dr. Steve McCallum. His critical analysis and suggestions ensured that each idea was carefully evaluated and focused. I would like to thank Dr. Sherry Bell. Her extensive knowledge of education and psychology helped me bridge ideas across disciplines. I would like to thank Dr. Morrow and Dr. Seaver for their knowledge and assistance in statistical methodology. I would like to thank Dr. Wilhoit for his ability to put everything into perspective.

I would like to thank Dr. Michael Hopkins and Dr. Angela Hilton-Prillhart who, along with Drs. McCallum and Bell, helped created the MIR probes. I would like to thank Dr. Scott Graves, Dr. Christopher Skinner, and Dr. Jerry Morton. I would not be where I am today without their guidance and support as I navigated graduate school and internship. Most importantly, their reassurance provided clarity during the most uncertain and difficult times along this path. Finally,I would like to thank the members of the research group who helped organize the probes and enter data. Specifically, I would like to thank Elizabeth Hayes, Lakmal Walpitage, Kelly Smyth, Brooke Browarnik, and Emily Taylor. I cannot express how grateful I am for the time you spent assisting me with this project.

Abstract

Three universal screeners and nine progress monitoring probes from the Monitoring Instructional Responsiveness: Math (MIR:M), a silent, group-administered math assessment designed for implementation with an RTI Model, were administered to 223 fifth-grade students. The growth parameters of the overall MIR:M composite and two global composites (math calculation and math reasoning) identified significant variation in student growth, within significant linear and quadratic trajectories. However, there were significant differences in the nature of the growth trajectories that have applied educational implications. In addition, growth parameters across the three composites provided significant predictive potential when using the Tennessee Comprehensive Assessment Program (TCAP) Achievement Test, a high-stakes, end of the year assessment of academic achievement, as the criterion measures ($p < .001$). Furthermore, these parameters were predictive at the classroom and student level. Differential predictive potential of the parameters and the composites provide additional information about the nature of the MIR:M data. Altogether, the growth modeling and the predictive modeling provide evidence to support two practical uses of the MIR:M.

Table of Contents

List of Tables

CHAPTER I

LITERATURE REVIEW

Because the Individuals with Disabilities Education Improvement Act of 2004 (IDEIA) allows educators to use the Response to Intervention (RTI) Model to help identify specific learning disabilities (SLDs), curriculum-based measures (CBMs) are typically used in the assessment and monitoring of academic progress. Although CBM-type measures are now used by educators across the country, most possess one or more serious flaws (e.g., inefficient individualized administration; unidimensional structure, inability to guide instruction). One experimental measure developed to address these limitations has yielded promising psychometric properties and utility (i.e., Monitoring Instructional Responsiveness: Math, MIR:M; Hopkins, McCallum, Bell & Mounger, 2010); however, additional validity data are needed before MIR: M will be accepted for widespread use. The purpose of this study is to continue the investigation into the psychometric integrity of the MIR: M by determining: (a) the extent to which the probe properties (i.e., intercept and slope) account for data variation, operationalized by "model fit" statistics from all (12) MIR:M administrations, with consideration of variable administration times (fixed interval vs. variable interval); and (b) the relative predictive power of various MIR: M scores (e.g., slopes and intercepts taken from all (12) probes from one year for the MIR:M Total composite core and two Global composite scores) when a large-scale end-of-year math composite score (Tennessee Comprehensive Assessment Program, TCAP; Tennessee Department of Education, 2011) is used as the criterion. The TCAP scores were chosen as the criterion because of the strong need to compare academic performance on CBM measures used within a RTI framework with scores taken from a credible high-stakes standardized test.

This literature review includes a: (a) brief history of measures commonly used to gauge progress in math, both high-stakes end-of-year tests and CBM-type instruments; (b) comparison of CBMs and high-stakes tests and a discussion highlighting the need to show the relationship between the two; (c) description of the limitations characterizing the validity estimates of CBM measures; and (d) discussion of how these limitations can be addressed in math CBMs in general and one experimental CBM measure in particular, the MIR:M. The literature review is followed by the statement of the problem and research questions, which, when answered, will address the strengths and weaknesses of the MIR:M.

High Stakes Assessment

The No Child Left Behind Act (NCLB; 2001) increased accountability for academic growth by requiring all students to make adequate yearly progress (AYP). Specifically, AYP required all students in grades 3 through 8 to be at or above standards of proficiency in reading and mathematics by the end of the 2013-2014 school year (Thum, 2003); although waivers have since been granted. Specifically, states must establish accountability systems based on a standardized assessment to include at least 95% of the student population and 95% of each subgroup (e.g., students on free/reduced lunch, minority students, students in special education, etc.). Graduation rates must be included for high schools, and states must establish separate objectives for reading and mathematics. In addition, schools and local education agencies (LEAs) are accountable for achievement and progress and must be based on substantial and continuous growth toward proficiency. Although the purpose of AYP is to encourage adequate progress for each academic area, data are reported typically in the aggregate; consequently, AYP has been conceptualized as setting a school level standard (Thum, 2003) with high-stakes standardized tests often used to establish adequate progress.

Accountability. Since end-of-the-year standardized tests are the primary instruments to determine accountability, these results have high-stakes implications for schools (Gulek, 2003). However, stakes and stakeholders have not always been adequately distinguished (Braden & Shroeder, 2004). That is, accountability is required by AYP (Thum, 2003), but NCLB fails to specify requirements for student promotion or retention and how progress affects teacher contracts; nor does it establish a link between teacher pay and performance (Braden & Schroeder). Adopting such consequences is at the discretion of the states and school districts. On the other hand, consequences for states and schools that fail to make AYP have been clearly specified within NCLB. In particular, states failing to meet AYP may have federal funds withheld. States must provide additional funding for schools that fail to meet AYP for two consecutive years; these schools must consider major changes. States are required to inform parents of the AYP of all schools. Parents are permitted to transfer students from failing schools to schools within the district that are making AYP, with transportation provided by the district. Finally, schools not meeting AYP are required to develop plans for improvement. As Braden and Shroeder (2004) point out, although consequences may extend to individual educators as a result of not achieving AYP, the direct consequences are placed first one the states, then on LEAs.

Test-based accountability was intended to provide incentive to increase motivation and performance of students, teachers, and administrators (Jacob, 2005); however, research on the impact of incentives on motivation has not provided much support for this notion (Deci, Koestner, & Ryan, 1999). Furthermore, according to Jacob, research on high-powered incentives may cause individuals to focus on the most salient aspects of a task (Holstrom & Milgrom, 1991). Similarly, within high-stakes accountability, schools may focus on the components that improve test scores and ignore other areas of education.

Efforts to increase accountability have yielded mixed results (Springer, 2008). After studying statewide assessment data over 20 years, Lee and Reeves (2012) concluded that accountability implementation did not have a consistent impact on student achievement as there were varying results across academic subjects and grades. In particular, comparing student gains before and after the implementation of NCLB, Lee and Reeves found that students experienced greater gains in math post-NCLB but not necessarily in reading. Similarly, Dee and Jacob (2011) found that imposing accountability had an impact on math achievement but not reading achievement. These inconclusive results should not be surprising since, historically, major resource allocation decisions have not caused significant gains on student achievement (Hanushek, 1997).

While the impact of imposing accountability has yielded mixed results, disaggregating students based on prior skills and demographics has highlighted variability in performance among students. Reback (2008) found that students in the margin of passing and failing experienced the most short-term gains with minimal impact on other students. Springer (2008) found that students with the lowest initial skills benefited the most from explicit accountability plans without negatively impacting students with more developed initial skills. In direct contrast to Springer's findings on higher-performing students, other researchers found that the gains of the lower-performing students appeared to come at the expense of their higher-performing peers (Deere & Strayer, 2001; Figlio & Rouse, 2006; Reback, 2008). In other words, when schools increased the scores of lower-performing students, the scores of higher-performing students were less than expected. In addition, Hanushek and Raymond (2005) found that accountability systems increased student gains and decreased the Hispanic-White achievement gap but increased the Black-White achievement gap.

The evidence is not only mixed as to the impact of accountability on overall student achievement but also its impact on various subgroups. Further complicating the pursuit of accountability are the differences between students in regular education and students in special education. Comparing regular education and special education teachers, Christenson et al. (2007) found significant differences in teachers' perspectives on promotion decisions of their students, especially in regards to high-stakes testing. This disconnect indicates that applying the same promotion criteria for students in special education as students in regular education is problematic as students in special education may have not met the same standards. Furthermore, universal expectations may complicate accountability standards since educators are expected to ensure all students make adequate progress, regardless if those expectations are unreasonable for students in special education.

One of the principal consequences for schools failing to make AYP is that they must allow eligible students the opportunity to transfer to schools meeting AYP within the district. Therefore, if the initial school was reducing students' academic proficiency, then the opportunity to choose a better school should have a noticeable impact on later proficiency. Similar to the literature on accountability, research on school choice has shown variable results (e.g., Cowen, Fleming, Witte, Wolf, & Kisida, 2013; Cullen, Jacob, & Levitt, 2005; Cullen, Jacob, & Levitt, 2006; Dobbie & Fryer, 2011; Rouse, 1998). For example, Cowen et al., found that students who participated in a school choice voucher program were more likely to graduate, enroll in college, and maintain enrollment in college, even when controlling for demographic variables. Conversely, Cullen et al. (2006) identified students who participated in a lottery of school choice, and found that students who won the lottery did not have significantly better academic outcomes than those that lost the lottery; however, students who entered the lottery had better academic outcomes than those who did not. Therefore, Cullen et al. concluded that students who entered the lottery shared common attributes, and the school itself had minimal impact on achievement; however, this study was limited to high school students.

With goals similar to the Cullen et al. (2006) study, Dobbie and Fryer (2011) studied the impact of a school lottery on entering kindergarteners and entering sixth graders at a charter school that used high-stakes testing and accountability to recruit and incentivize teachers. Controlling for various demographic variables, Dobbie and Fryer found significant gains between the students who won the lottery and attended the school compared to those who lost the lottery. Furthermore, Dobbie and Fryer found that after a few years in their respective schools, the Black-White achievement gap in mathematics was reduced significantly for the elementary students and the language arts achievement gap was reduced significantly for both the elementary and middle school students. This may highlight an important interaction of increased accountability within high-performing schools and the variables that contributed to students initially entering the lottery (Cullen et al., 2005; 2006). Therefore, the benefits of accountability are more pronounced when students and their parents are motivated to enter students into the lottery. Nonetheless, the research on school choice offers mixed results for schools but promising results for parents who choose to take advantage of choice.

Measuring Teacher Accountability. As noted by Braden and Shroeder (2004), NCLB did not place accountability for increasing student achievement on any single stakeholder. Naturally though, as teachers have an inherently larger responsibility on student achievement, specific policies have placed explicit accountability on teachers. As a result, such decisions as teacher promotion, retention, and raises have become prevalent within the high-stakes testing environment. For example, Chingos and West (2011) found that as a response to greater accountability, effective teachers are promoted to positions of leadership (e.g., principals) while less effective teachers are placed in low-stakes positions within the school. In addition, effective teachers were more likely to remain in high-stakes teaching positions in lower-performing schools, where leadership positions have less impact on students. Chingos and West conclude that this is a result of administrators making decisions to best increase student test scores.

As a result of decision-making based on high-stakes assessment, some researchers have focused on teachers' perceptions of high-stakes assessment and accountability policies. Guskey (2007) studied the ranking of 15 different measurements of academic utility and found that state assessments were ranked 14th. While teachers appear to have relatively low perceptions of the assessment themselves, their perceptions of decisions based on the high-stakes data may be even more extreme. As Schoen and Fusarelli (2008) explained, teachers and school leaders have become fearful of the negative consequences of high-stakes testing and are reluctant to deviate from what has already worked, rather than attempting to identify novel techniques and procedures. Schoen and Fusarelli expanded upon the research of Hagel and Brown (2002) and hypothesized that this environment of fear has lead a decrease in innovation and risk-taking and an increase in a "high-threat" school environment. Contract negotiations have amplified these fears with government and private company takeover of schools becoming increasingly common (Cooper & Sureau, 2008). Moreover, Braden and Shroeder (2004) identified the following unintended consequences of high-stakes accountability that pertain to teachers: limiting academic focus, academic demoralization, anxiety, targeting borderline students while ignoring other students, and increased cheating. Altogether, many unintended consequences of NCLB have impacted teachers, and their attitudes towards high-stakes are significantly negative.

Given these changes and the subsequent impact on teachers, it is important to identify how teacher effectiveness is measured and whether the effectiveness has significant impact on test scores. Overall, specific characteristics of effective teachers have been difficult to isolate. Aaronson, Barrow, and Sander (2007), found that the most commonly observed characteristics of teachers (e.g., degree, rankings of college attended, gender, experience) had very little impact on the quality of teachers, and much of the variation in teacher quality came from unknown characteristics. Despite the difficulty identifying *what* makes an effective teacher, models have been developed that can identify *who* makes an effective teacher. Specifically, value-added models (VAM) isolate the contributions of teachers and schools on student achievement (McCaffery, Lockwood, Koretz, Louis, & Hamilton, 2004) and account for various influences on achievement and growth (e.g., family characteristics, basic demographics, prior achievement) to determine the value added by school and teacher influences (Meyer, 1997). Different, but related techniques have been used to obtain valued-added estimates, although McCaffery et al., argue that the common models are all extensions of a multivariate, longitudinal mixed-model.

Taken together, various VAMs have produced some promising results on teacher effectiveness. Aggregating a number of studies on teacher effectiveness, Hanushek and Rivkin (2012) found that quality teachers have an average impact of .13 standard deviations for students' reading achievement and .17 for students' math achievement. Put another way, teachers in the 75th percentile of effective teachers will increase student performance .2 standard deviations (Hanushek and Rivkin). Furthermore, according to Hanushek (2011), by quantifying the value of higher achievement, replacing the least effective teachers with average teachers would increase the future earnings of students by nearly $100 trillion and place nationwide scores at the top of worldwide test scores. In other words, the gains that students would make would have a significant monetary impact as a result of increasing test scores.

Overall, while there appears to be some promising effects of imposing stronger accountability, implementing broad aspects of accountability rather than specific components that are most effective, may moderate the overall effectiveness. Even so, broad implementations appear to be the trend for the immediate future. Within this current trend, research indicates that teachers are experiencing a great deal of uncertainty. Given that high-stakes tests are administered once, at the end of the year, it is apparent that the majority of the year is spent without information from these tests, which could be used to impact each student's skill and growth. Furthermore, teachers are already limited in both time and resources, and what they do have available is needed for instructional time. Therefore, in order to obtain information of students' skills throughout the school year, an instrument must be efficient (i.e., brief) yet informative. In addition, because high-stakes tests results are considered important having brief CBMs which could predict end-of-year high-stakes tests would be very useful to teachers. Obviously, these measures should not only be brief, but should target more than one skill at a time, consistent with high-stakes criterion measures. Recent developments in instruments used in Curriculum-Based Measurement (CBM) appear to provide an efficient and effective means to assess multiple skills at multiple times during the academic year, but can they predict high-stakes test scores?

Curriculum-Based Measurement

Development of CBMs. Curriculum-based measurement stems from a broader set of assessments known as curriculum-based assessment (CBA) designed to measure basic academic skills (Shaprio, 2011). CBA addresses curriculum content that can inform academic instruction and is amenable to repeated testing (Hosp & Hosp, 2003; Tucker, 1987). In particular, these assessments are designed to evaluate the instructional needs and performance of students within their school's curriculum (Glicking & Havertape, 1981) and assist teachers in discovering an optimal level of instruction for each student (Tucker, 1985). This optimal level, or "window of learning," as Tucker points out, can be defined as "between frustration and boredom" (p. 201).

Curriculum-based assessments generally fall into one of two categories: mastery measures and outcome measures (Fuchs & Deno, 1991). Mastery measurement models break down global skills into measurable subskills that lead to short-term instructional goals of criterion-referenced mastery standards (Fuchs & Deno, 1991). These measures tend to be informal and are often teacher-made assessments used to measure a student's mastery of the classroom curriculum (Deno, 1992). Therefore, mastery measures align well with Tucker's definition of CBAs; they are based on the classroom curriculum and are meant to guide instructional planning through criterion-referenced goals and standards for each student and classroom. Mastery measures help to increase the overall efficiency and mastery of learning; however, the informal development and administration of these measures has drawbacks. Specifically, little is known of basic psychometric properties, such as reliability and validity, since these measures are usually developed by teachers and are tailored for particular subskills and students (Hosp & Hosp, 2003). Even when teachers followed more prescriptive procedures (Deno & Mirkin, 1977) but were given the option to choose from a collection of various measures as they saw fit, reliability and validity estimates were unknown because each measurement was distinct (Fuchs & Deno, 1991). In addition, the emphasis on short-term goals did not allow for an understanding of broader instructional issues such as the impact of instruction on student growth and identification of alternative instructional strategies (Fuchs &

Deno, 1991). Moreover, as these measures have essentially become an integrated part of curriculum and instruction, they cannot objectively measure instructional effectiveness (Deno, 1992). Therefore, unless each measure is studied and (perhaps) standardized, the extent to which a mastery measure can provide a valid and reliable assessment of a skill and the effectiveness of instructional practices will likely remain unknown.

Outcome measures, on the other hand, take on a very different role than mastery measures and can account for the shortcomings of informal assessments. Outcome measures use a standardized set of repeated procedures, which remain constant over an extended period of time while focusing on the proficiency of various global outcomes of a curriculum (Fuchs & Deno, 1991). Therefore, outcome measures differ from the mastery measures in that they move from specific and non-standardized measures of subskills to broader standardized measures of global skills. Fuchs and Deno (1991) identified a number of advantages of using outcomes of global skills as the strategy allows educators to: (a) focus on the broad end-of-the-year outcome; (b) avoid having to deconstruct curriculum into a sequence of smaller instructional tasks that can be time consuming and prone to error; (c) separate the content and instruction which allows teachers to test different methods and content of instruction; and (d) measure retention and generalization of skills since the general outcome skill should increase with each subskill.

Although Fuchs and Deno (1991) proposed multiple types of outcome measurement systems, their major focus was to describe measures which could assess basic curriculum (e.g., math, reading, and writing) and inform instructional planning. As a result, curriculum-based measurement was derived from this outcome measurement model of curriculum-based assessment as it provides repeated assessment of global skills (Hosp & Hosp, 2003), but is delineated from other outcome measurement systems because the focus is on measuring basic skills from a curriculum. This can lead to adapted instruction (Deno, 1985) by identifying skills that need targeted (Burns, MacQuarrie, & Campbell, 1999).

In order to maximize their utility, Deno (1985) initially proposed that CBMs must share a number of basic attributes. First, CBMs must be reliable and valid, to ensure evidence of student achievement as a foundation for instructional decisions, yet sensitive to skill changes. Basic psychometric properties are essential components of well-made CBMs, distinguishing them from the often teacher-made CBAs (Deno, 2003; Good & Jefferson, 1998; Shinn, 1989). The National Center on Response to Intervention (NCRTI) maintains a database and reviews properties of CBMs in order to provide comparisons of CBMs for educators. In addition, Deno (1985) called for an increase in the sensitivity of the measures to model student growth and changes. Thus, beyond basic measures of reliability and validity, there is now an emphasis on more advanced psychometric components such as the reliability of the slope and disaggregated norms as reported by NCRTI.

CBMs need to be simple and efficient in order to frequently monitor student achievement (Deno, 1985). As a result of reliance on repeated measures to monitor student progress, CBMs use different but equivalent forms (Hosp & Hosp, 2003), that are based on the curriculum (Kelley, Hosp, & Hollow, 2008). These equivalent forms ensure that students are measured on the same task and level of difficulty to draw conclusions about student proficiency while keeping the items unfamiliar to add to the generalizability (Deno, 2003) and decrease the possibility of practice effects. While the exact number of equivalent forms varies, the NCRTI considers 9 alternate forms as the criterion for an adequate number of measures; those that show stronger reliability and equivalence are rated higher.

In addition, with frequent monitoring, the need for efficiency is increasingly important.

As Wesson, King, and Deno (1984) found, nearly 50% of teachers identified time as a barrier to implementation of frequent measurements. Thus, to maintain time efficiency, administration time for CBMs is usually between 1 to 3 minutes, depending on which skills are being assessed (Deno, 2003). Although, the need for efficiency is evident, especially for educators, preservice teachers appeared somewhat skeptical of the validity of a brief, one-minute measure as a general indicator of performance (Foegen, Espin, Allinder, & Markell, 2001).

Deno (1985) also specified the importance of measures that could be easily understood in order to communicate results to educators, students, and parents. As a result, CBMs have tended to be uncomplicated to ensure educators and parents alike can learn the procedures while maintaining the integrity of the measurement (Deno, 2003). When preservice teachers were given presentations about the basic properties of CBMs, they rated the utility of CBMs to evaluate and modify instruction positively, regardless of the presentation type (Foregen et al., 2001). Thus, it appears that the nature of CBMs enable teachers to understand the usefulness of these measures and provides support for the uncomplicated design of CBMs.

Given the repeated measurements of CBMs, they need to be inexpensive. With the increasing number of CBMs, the NCRTI displays prices for each set of measures so that educators can choose those that can meet their needs and be cost-effective. In addition, there are a number of CBMs that can be accessed for free. Thus, the increasing usage of CBMs has given rise to measures that can fulfill assessment needs while remaining inexpensive.

Benefits and Applications of CBMs. The benefit of the standardized procedures of CBMs was discovered soon after their inception. For example, Fuchs, Deno, and Mirkin (1984) found that teachers who were randomly assigned to measure student progress and were systematically using repeated measures on an ongoing basis were more likely to have greater student gains, more realistic views of student progress, and have students who were more knowledgeable about their own learning when compared to teachers randomly assigned to less systematic evaluation procedures. Furthermore, teachers who implemented systematic evaluation techniques were more likely to adapt their instruction based on the CBM results. Thus, systematically monitoring progress with an outcome measure had positive consequences for teachers and students.

Compared to other CBAs, CBMs are able to evaluate instructional methods (Fuchs & Deno, 1991), and are commonly used to evaluate individual instructional programs, classroom interventions, and instructional placement (Fuchs, 2003). Consequently, the evaluative capabilities of CBMs have consistently been shown to increase the likelihood that educators adapt instruction and increase the quality of instruction (e.g., Fuchs et al., 1984; Fuchs, Fuchs, Hamlett, & Stecker, 1991; Fuchs, Fuchs, Hosp, & Hamlett 2003). Apparently, the consistent evaluation of instruction has not only led to the identification of effective and ineffective strategies, but it appears to have prompted educators to adapt their strategies.

By ensuring that CBMs have established basic psychometric qualities, educators have been afforded the opportunity to interpret data in multiple ways. CBMs are capable of monitoring students' skills in reference to school curriculum while allowing for comparison of each student's individual progress and a comparison of skills across peers (Deno, 1985). Specifically, schools can use data from CBMs to establish norms to compare student performance and growth to peers within a similar environment and establish targets for performance (Deno, 2003). Therefore these norms can allow for a norm-referenced comparison to peers and a criterion-referenced standard for academics. With an established set of norms, districts are able to identify students who have academic weaknesses and are at-risk for academic failure (Deno, 2003). Schools can then identify the lowest performing students and provide additional intervention and monitoring of their progress (Deno, Reschly-Anderson, Lembke, Zorka, and Callende, 2002). As a result of these flexible applications, CBMs have become a prominent part of education, especially given recent legislative changes that require progress monitoring within RTI.

Response to Intervention. The Individuals with Disabilities Education Improvement Act of 2004 (IDEIA 2004) allowed school districts options for special education determination of learning disabilities. Schools are no longer required to use an achievement-ability discrepancy to determine special education eligibility; instead, they are permitted to monitor students' response to research-based interventions. This process, commonly known as Response to Intervention (RTI), can vary in the specific details of implementation since the legislation did not include specific guidelines nor endorse a specific model (Bradley, Danielson, & Doolittle, 2007). Generally, most models ensure that students are given adequate instruction, that progress is monitored, and that students not responding to instruction are given additional intervention with continued progress monitoring. Finally, students who consistently fail to respond are referred for special education determination (Fuchs, Mock, Morgan, & Young, 2003). More specifically, students are screened for initial skills deficits and move through tiers (Ardoin, Witt, Connell, & Koenig, 2005), receiving increasingly individualized, intensive instruction and interventions as they increase tiers (Hollenbeck, 2007) with consistent monitoring of progress (Hosp & Hosp, 2003). Models usually incorporate three or more tiers and include at least a primary, secondary, and tertiary tier (Bradley et al., 2007); the three tier model (Ardoin et al., 2005; Fuchs & Fuchs, 2007; Vaughn & Fuchs, 2003) is most commonly implemented. Given the numerous applications of CBMs, and the procedures of RTI, CBMs have become an important component of all phases of the process (Hopkins, 2011).

Within the first tier of RTI, the primary tier, students are universally screened with CBMs to identify at-risk students with skill deficits (Ardoin et al., 2005). Using local norms, schools identify students scoring at or below a specified percentile (Hopkins, 2011). Schools have flexibility as to what is designated as a skill deficit, and researchers have studied or proposed different cut-offs most commonly ranging from the lowest 10% to lowest 20% (Deno et al., 2002; Hopkins, 2011; VanDerHeyden, Witt, & Barnett, 2005). Universal screening typically occurs three times during the school year (Hopkins, 2011), although some models only use screening once per year (Fuchs & Fuchs, 2007).

Students who are identified as at-risk, move to tier 2 (Bradley et al., 2007) and usually receive additional, more intensive group-instruction (Fuchs & Fuchs, 2007; Hollenback, 2007). This typically lasts eight to 12 weeks in order to determine students' rate of progress (Bradley et al., 2007), although some models recommend a longer time period (Fuchs & Fuchs, 2007). Typically, CBMs are given on a weekly or bi-weekly schedule to monitor the progress of students (Hopkins, 2011). This growth is used to determine if students fail to respond to the additional instructional support (Deno, 2003). Similar to universal screening, students' rate of growth is usually determined by local norms to determine if they are adequately responding (Hopkins, 2011). Students who fail to respond to the additional instruction, move to tier 3, the tertiary tier (Bradley et al., 2007). Within this phase, students are given even more intensive, individualized research-based instruction (Fuchs & Fuchs, 2007; Hopkins 2011) and progress monitoring with CBMs is usually given on a similar schedule as the second tier (Bradley et al., 2007). Students who fail to respond during this phase often receive a special education referral

(Ardoin et al., 2005). In sum, CBMs are used for initial screening of potential skill deficits in tier 1; CBMs are used to progress monitor skills as students receive additional support in tier 2 and can be used to identify students not responding to additional support; and finally, CBMs are used to progress monitor intensive interventions in tier 3 and to identify students who are not responding who may need referral for special education eligibility. Although CBMs can assess reading and math, the majority of research has focused on the assessment of reading.

Curriculum-Based Measurement of Math. Despite nearly 30 years of research with CBMs and the increased application of CBMs because of RTI, the majority of research and use of CBMs has focused on reading-based measures. For example, the NCRTI lists data for 65 measures of reading with 32 measures for universal screening and 33 measures for progress monitoring; however, the NCRTI only lists data for 42 measures of mathematics with 17 measures for universal screening and 25 measures for progress monitoring. In review of the literature on mathematical curriculum-based measurement (M-CBM), Foegen, Jiban, and Deno (2007) found 32 out of 163 empirical studies on CBMs (20%) were with M-CBMs. Contrast these results with those of Recshly, Busch, Betts, Deno, and Long (2009) who performed a meta-analysis on reading curriculum-based measurement (R-CBM) studies that used a specific type of R-CBM measure (oral reading fluency), analysis (predicting norm-referenced reading achievement), and did not possess other exclusionary criteria (e.g., above grade 6, modified procedures, etc.). Even with the strict criteria, Rechsly et al. found 41 studies for inclusion. Thus with fewer measures and far fewer studies, M-CBMs are still in the relatively early stages of development compared to the R-CBMs.

One possible reason that math measurement has lagged behind reading measurement are the differences in the development between the two skills. For example, Juel (1988) discovered that nine out of 10 students, who had established strong basic reading skills by the end of first grade, were also strong readers by the end of fourth grade. This overwhelming consistency in reading performance highlights the nature of reading. As Foegen et al. (2007) pointed out, the goals of reading are simpler and clearer than the goals of math; moreover, while each process is complicated, the nature of reading is relatively consistent while math is not.

Clarke, Baker, and Chard (2008) identified three major difficulties that arise in math measurement. Unlike reading development, students may be strong in a particular skill at one level but struggle within the next. For example, a student may perform well on measures of basic multiplication facts but have difficulty when multiplying by multiple digits that require multiple math concepts and procedures. Therefore, the proficiency of a student on one math skill may not be generalizable to another math skill, especially when compared to reading skills. The second difficulty Clark et al. acknowledged, was the multiple domains of math skills that, again, may not be completely generalizable. For example, a student may perform well within the broad domain of math calculation skills but have difficulty within the domain of measurement. Thus, not only can it be difficult generalizing across skills levels, it can also be difficult generalizing across math skill types and domains. Interrelatedness of skills may cause difficulty establishing concurrent validity (Polignano & Hojnoski, 2012) as relationships between theoretically similar and dissimilar constructs may not be as predictable as expected. The final issue identified by Clark et al. is the more limited experimental research with math instruction compared to reading. In sum, although these difficulties likely make math measurement and educational decision-making more complicated, there is an expanding foundation of instruments and research of M-CBMs.

Mathematics assessment generally measures either narrow skills using a single-skill probe, which are aligned with subskill mastery measurement, or broad skills using a multiple-skill probes, which are most aligned with general outcome measurement (Christ & Vining, 2006). For example, a single-skill probe may measure multiplication facts through nine while a multiple-skill probe may measure multiplication and division facts. Further specifying the types of mathematics curriculum-based measurement (M-CBM), Kelley, Hosp, and Hollow (2008) explained that it is difficult to generalize outcome measures since there is no single task that can be measured that generalizes math skills across domains. Instead, according to Kelley et al., skill-based measures, which measure multiple skills across a domain (e.g., computation), have been used to fulfill the needs of general outcome measures. Regardless of whether skills-based measures can be categorized as true forms of outcome measures, they have become the predominant instruments for M-CBM.

There is little consensus regarding the best practices in developing M-CBMs (Foegen et al., 2007). Despite this lack of consensus, most M-CBMs are developed using either curriculum sampling, a representative sampling of a specific year's curriculum, or robust indicators, skills that have been empirically linked as indicators of math proficiency (Fuchs, 2004). The principal advantage of curriculum sampling is the link to skills and instruction within a classroom that can give teachers information to modify and adapt instruction (Foegen et al.). On the other hand, the principal advantages of robust indicators are that they are predictive of achievement and they are easier to measure across grade levels (Foegen et al.). Foegen et al. found that M-CBMs for early math skills focused exclusively on robust indicators while M-CBMs for later skills included a combination of curriculum sampling and robust indicators.

While the development of math skills and math measures are complex, Thurber, Shinn and Smolkowski (2002) hypothesized that M-CBM can be broadly placed into two categories: computation or operation, and applications or problem solving. Through confirmatory factor analysis (CFA), Thurber et al., established construct-validity for their hypothesis of two-factor model of M-CBM (computation and applications), as it had considerably better fit than a one-factor model. That being said, while they found strong support for two separate constructs, these constructs were still highly correlated (*r*=.83) and both constructs were highly correlated with reading as well (i.e., .76 for applications, .69 for computation).

Despite the expanding foundation of research establishing the basic psychometric properties of M-CBMs, the research base is still rather limited. Furthermore, there are known limitations, such as the large variation of scores in M-CBMs that limits the reliability of their decision-making proficiency (Fuchs, Fuchs, & Zumeta, 2008; Hopkins, 2011; VanDerHeyden et al., 2005). Fortunately, there are a couple of research areas that may alleviate some of the uncertainty related to M-CBM. In particular, two important goals have shown promise: (a) establishing the predictive potential of M-CBMs with high-stakes assessment; and (b) modeling student growth on M-CBMs.

Predictive Potential. Fuchs (2004) proposed three research paradigms that are important to establishing the validity and utility of CBM: Stage 1 focuses on the technical components of the static score, or the score of a single CBM administrations; Stage 2 focuses on the technical components of the slope of student growth on CBMs; and Stage 3 focuses on the instructional utility of CBMs. Consistent with the first stage, one major line of research has demonstrated the validity of CBMs with standardized, norm-referenced constructs by establishing their concurrent and/or the predictive validity. Because one component of curriculum-based measurement is outcome measurement, predictive validity has become a popular line of research. In addition, since CBMs are meant to be efficient, the short time limit of these measures means that students' scores are a function of the students' speed, or fluency, on a task. In order to ensure that the efficiency of this constrained time does not offset its utility, it is especially important to establish the predictive potential of these measures. Overall, the utility and validity of reading fluency measures have yielded positive results, with moderate to strong correlations with norm-referenced, standardized assessments of reading (Neddenriep, Skinner, Hale, Oliver, & Winn, 2007; Skinner, Neddenriep, Bradley-Klug & Ziemann, 2002; Skinner, Williams, Morrow, Hale; Neddenriep, & Hawkins, 2009; Williams, Skinner, Floyd, Hale, Neddenriep, & Kirk, 2011) and math (Allinder, Fuchs, Fuchs, & Hamlett, 1992; Clarke & Shinn, 2004; Fuchs, Fuchs, & Hamlett, 1989; Phillips, Hamlett, Fuchs & Fuchs, 1993; Thurber et al., 2002; VanDerHeyden & Burns, 2005).

Research has just begun to focus on the predictive power of CBMs on statewide end-of-the-year, high-stakes assessment. Unfortunately, not only is this a relatively new line of research, but it appears that the majority of the research has primarily focused on the predictive power of reading CBMs (Crawford, Tindal, & Stieber, 2001; Good, Simmons, & Kame'enui, 2001; Helwig, Anderson, & Tindal, 2002; Hintze & Silberglitt, 2005; Miller, 2012; McGlinchey & Hixon, 2004; Richardson, Hawken, & Kircher, 2011; Shapiro, Keller, Lutz, Santoro, & Hintze, 2006; Silberglitt & Hintze, 2005; Silberglitt, Burns, Madyun & Lail, 2006; Stage & Jacobsen, 2001; Wiley & Deno, 2005; Wood, 2006).

While the research base is limited, a select number of studies have used M-CBMs to predict these assessments. For example, Keller-Margulius, Shapiro, and Hintze (2008) studied relationship of math computation and math concepts and applications from an M-CBM, across grades 1 through 3, with math achievement on state-wide achievement tests (grades 1 and 3) and a nationally standardized test of achievement (grades 2). They discovered that M-CBM scores were significantly related to the outcome measures taken from both one year and two years after measurement. Across the three measurements, the median correlation was .50 for grade 1, .57 for grade 2, and .46 for grade 3. The researchers noted that math concepts and applications CBMs had a nearly identical pattern of relationships with the outcome measures. While this research established the predictive potential of single delineated measures, the predictive potential of the combined measures (e.g., slope) was not established.

Jiban and Deno (2007) attempted to determine if two, one-minute CBM measures could significantly relate to a statewide achievement test given approximately two weeks later for grades 3 and 5.A measure that modified basic math facts and required students to complete the missing portion of a basic computation equation with the missing number varying in its position (i.e., before the equal sign, on either side of the operator, or after the equal sign) was a better predictor than a measure of basic math facts with the missing number always following the equal sign. Individually, the modified math facts explained 17% of the variance in third grade and 34% of the variance in fifth grade math performance while the basic math facts explained 4% of the variance in third grade and 31% of the variance in fifth grade math performance. The authors concluded that the combination of the two measures provides the most predictive potential of the statewide test of achievement. Altogether, these limited studies have established a link between M-CBMs and high-stakes testing. An additional research area that is likely to strengthen this link and increase the practical utility of CBMs is modeling student growth.

Student Growth with CBMs

Fuchs (2004) hypothesized that establishing the basic parameters of the slope would be consistent with "stage two" of the research process establishing the utility of CBMs since slope-based data could determine the extent to which growth is associated with improvement in an academic domain. Student growth on CBMs has been conceptualized and modeled in a number of ways. Recently though, modeling individual student growth has become a prominent and flexible technique. In particular, this process allows for great flexibility for testing research hypotheses, conceptualizing different initial skills and growth parameters, and identifying exogenous variables that may impact growth. Furthermore, these methods provide models that are more consistent with identifying individual growth that is consistent with RTI goals. Therefore, these models provide a useful and practical link between research and practice.

While modeling growth has become increasingly popular, the methods and procedures have varied across studies. Initial modeling of growth used ordinary least squares (OLS) regression and modeled scores by time (Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993; Hintze & Christ, 2004; Hintze & Shapiro, 1997; Keller-Margulius et. al 2008). For example, Fuchs et al. (1993) ran an OLS regression with calendar days of each administration regressed on the scores of the CBM to obtain the slope. This growth estimate becomes a function of the average growth over time across students. More specifically, coefficients are contingent on the mean entering skill levels (intercept) and the mean growth of all students, and therefore, the model represents the expected growth of students within that sample. However, this approach is inconsistent with the typical use of CBMs which are often used to make decisions on the individual level regardless of the average growth of all students (Silberglitt & Hintze, 2007). While an efficient approach to modeling that may be sufficient for general research and evaluation purposes, it lacks practical applicability. Essentially, unless schools need to use the entering skills and growth of the average student, this method is likely not an effective approach to modeling growth.

Given the shortcomings of using OLS to model student growth, and the advancements in statistical analysis for change, a number of advanced models have emerged as more effective means to capture growth. In addition, these models allow for other applications to model the unique properties of CBM data that allow for additional flexibility not available in traditional analyses. In particular, Latent Growth Modeling, an extension of Structural Equation Modeling (SEM), and Growth Curve Modeling (GCM), an extension of Hierarchical Linear Modeling (HLM), have gained favor in academic research.

Latent Growth Modeling. Latent Growth Modeling (LGM) within Structural Equation Modeling (Meredith & Tisak, 1990) is an increasingly popular method of assessing growth. As a result, this method is more commonly used to model growth of CBMs (Chard et al., 2008; Clarke, Baker, Smolowski, & Chard, 2008; Costa, Hooper, McBee, Anderson, & Yerby, 2012; Yeo, Fearington, & Christ, 2011; Yeo, Kim, Branum-Martin, Wayman, & Espin, 2012). The broader method of SEM provides some distinct advantages to other analytic techniques. SEM is generally a confirmatory methodology that tests pre-specified models by simultaneously modeling the covariances of regression coefficients or the means and intercepts of basic parameters (Byrne, 2010). Furthermore, SEM can model latent (unobserved) constructs that cannot be directly measured by using observed indicators that are representative of the latent construct (Singer & Willet, 2003). Since analysis of SEM provides an estimate of a model's fit, it allows for the flexibility of modeling options which include: the ability to model the error and reliability of the observed indicators; model and test relationships of the constructs; and test complex models and hypotheses (Chin, 1998; Lei & Wu, 2007).

Tomarken and Waller (2005) identified a number of benefits of using LGM to model longitudinal data compared to traditional measures (e.g., repeated measures ANOVA, OLS). Many of the benefits of LGM lead to useful applications with CBM modeling. These useful applications include: (a) flexibility of modeling time and inclusion of time-variant covariates and the relationship of constructs over time (e.g., Curran & Hussong, 2003; Hussong, Curran, Moffit, Caspi, & Carrig, 2004); (b) ability to test data that are structured within a hierarchy (e.g.,Curran & Hussong, 2003; Duncan et al., 2002); (c) ability to model multivariate changes across measures (e.g., Sayer & Willet, 1998); and (d) ability to model and account for missing data (e.g., Allison, 2003; Duncan et al. 1999).

A useful component of LGM is the ability to model time flexibly and include both time-invariant and time-variant covariates. When modeling growth, accurately capturing the trajectory of this growth across a period of time is a central issue. When CBMs model student growth as a result of instruction, the amount of instruction that occurs between measurements should contribute to growth. Therefore, if students have two weeks of instruction between all measurements, the growth would be expected to remain relatively consistent between the measurements. On the other hand, if the time between measurements is not constant, then the amount of growth would be expected to vary between these measurements. Given the structure of an academic school year, equally spaced measurements may not be possible. LGM can handle unequally spaced time between measurements, although it is assumed that all individuals are measured at the same time (Byrne, 2010). In addition, LGM can account for time-invariant covariates (e.g., gender) (Byrne, 2010) as well as time-variant covariates (e.g., environmental changes) that may impact the model beyond the traditional growth trajectory (Curran & Hussong, 2003). For example, Gottfried, Marcoulides, Gottfried, Oliver, and Guerin (2007) found that over time, as students' math achievement declined so did their math motivation. Therefore, variation of math achievement explained a significant amount of the variation in math motivation beyond the time component of the model.

Another useful application of LGM is the ability to model random intercepts and slopes for each participant (Curran & Hussong, 2003). Essentially it provides a separate model of initial status and growth for each individual. Also LGM can model the heterogeneity of growth as a function of the initial status (Klein & Muthen, 2006). In other words, LGM can analyze the initial status of a variable and its impact on the variability of growth parameters. For example, modeling the growth of math achievement of students in grades 7 through 10, Klein and Muthen (2006) found a better fit when the initial skills in seventh grade were compared to an ordinary LGM since students with lower initial skills varied far greater in their later measurements than students with higher initial skills. These applications can be useful with CBMs since the slope and intercept of each individual can be modeled and the growth, as well as the variability of growth, may be a function of differences in initial skills. This provides a great deal of flexibility since initial performance and rate of growth on CBMs are both important components of RTI.

While SEM has provided a number of flexible applications to model measurement error and obtain reliability estimates, LGM can more flexibly model the reliability of repeated measures than traditional analyses. LGM can provide flexible specification of the correlation between errors and does not assume that error variance stays constant over time (Stoel, ven den Wittenboer, & Hox, 2003). In addition, conventional analyses of reliability (e.g., test-retest, alternate-forms) do not account for measurement error and therefore only model the observed error (Conroy, Metzler, & Hofer, 2003). Furthermore, since LGM can allow different spacing of measurement, it can account for differences in time that may impact the reliability (Yeo et al., 2012). By accounting for measurement error and time, it is possible to partition and model the stability of three distinct components of measurement error over time: (a) structural stability, the multidimensional properties of the measure; (b) differential stability, or the correlation of two measurement occasions over a period of time and from the rank ordering of individuals over time; and (c) mean stability, the overall changes of scores over time (Conroy et al., 2003; Marsh, 1993). While the structural stability may not be testable with CBMs that use unidimensional measurements modeling, the differential and mean stabilities could provide useful information about the reliability of the measures.

Traditional analyses of longitudinal data often call for a listwise deletion of missing data with only individuals with all data points kept for analysis (Curran & Hussong, 2003). Therefore, any individual with missing data is deleted from the analysis, whether the data is useful or not. This can lead to biased estimates of parameters (Jones, 1996). Fortunately, LGM has the ability to account for missing data. According to Curran and Hussong (2003), there are a number of powerful approaches within LGM to account for missing data. Furthermore, when compared to conventional approaches to handling missing data, the methods within LGM were less likely to bias the estimators (Jones, 1996). Given the fluid and idiosyncratic nature of data collection within the school setting, in addition to the regular absences that may occur, it is not uncommon for a number of students to have missing data points. In addition, as the number of probes increases, the probability of a student missing an administration also increases, thereby offsetting the benefits of the repeated measurements.

Applications with CBMs. The applicability of LGM in general, is related to its flexibility. For example, Yeo et al. (2011) used LGM to test the equivalence of the growth curves between two measures and determine if the growth was correlated. Since LGM allowed the researchers to control for the initial status of skills, Yeo et al. found unexpected results because there was little equivalence between the two measures. This finding would not have been apparent from more traditional analyses. Chard et al. (2008) produced a similar growth model, using a reading fluency CBM, but instead of comparing three measurements across one year, they examined three measurements across three years. They then identified which variables were significantly related to both students' initial skills and their growth. In particular, Chard et al. found that basic reading skills, such as letter-naming, were significantly related to students' initial skills while students' growth was significantly related to comprehension, an advanced reading skill. While these results are consistent with the theoretical development of skills, LGM's ability to delineate between students' initial skills and growth confirmed the expected relationship.

Using four measurements of early numeracy with students in kindergarten and first grade, Missall, Mercer, Martinez, and Casebeer (2012) used a complex application of LGM to identify latent clusters of students based on their skills at each level of measurement and their growth trajectories. They were able to determine the probability that students within a cluster at measurement would perform within a different cluster taken from a different level of measurement. In addition, they found that students generally stayed in their initial clusters and this was predictive of math achievement in third grade; however this analysis could allow researchers to identify students who grew to a higher level cluster to determine which variable contributed to this change.

In addition, Yeo et al. (2012) applied LGM to measure the alternate-forms reliability of a common CBM by accounting for time between measurements and measurement error of six measurements. In particular, this study showed that reliability is not a constant coefficient, and it can vary across time and measures. Therefore, rather than assuming that the reliability is fixed, this study showed that some measurements were more reliable than others, which can impact the interpretation of the results of a single measurement and its impact on the overall slope.

Pianta, Belsky, Vandergrift, Houts, and Morrison (2008) found that time-varying covariates of student-teacher emotional interactions and exposure to math activities had a positive relationship with math achievement. In addition, the Pianta et al. study highlights the variation in environmental conditions within RTI that may explain growth in CBMs. Within an RTI framework, instructional changes are made for low-performing students (e.g., adding additional instructional time, implementing an intervention.). Presumably, these changes will increase a student’s growth beyond what would be expected within the prior instructional environment. Therefore, LGM can account for these changes by modeling its impact while still modeling the growth parameters.

Growth Curve Modeling. Another flexible group of models that has become increasingly popular for measuring change are growth curve models within Hierarchical Linear Modeling (Bryk & Raudenbush, 1987). As with LGM, the flexibility of growth curve modeling, and the increasing use of CBMs within the hierarchical nature of school structures has resulted in more frequent application of this technique (Codding et al., 2007; Graney & Shinn, 2005; Kamata, Nese, Patarapichayatham, & Lai, 2012; Miller, 2012; Stage, 2001; Stage & Jacobsen, 2001; Silberglitt, Appleton, Burns, & Jimerson, 2006).

Having taken on many different labels such as *multilevel modeling*, *mixed modeling*, *random coefficient modeling*, *mixed methods modeling*, *multilevel linear modeling*, among others, HLM refers to the nested and hierarchical relationships among units and allows for coefficients of the units to vary within each level of the respective units (Graves & Frowherk, 2008). More specifically, HLM allows for variables to be nested within a structure (e.g., classrooms within a school, students within a classroom, measurements within a student) and allows regression intercepts and slopes to randomly vary within these nested units which make up different levels of the model (Raudenbush & Bryk, 2002). For example, if students' scores are going to be impacted by the particular classroom of instruction, then it is important to model the effects of the classroom. These varying effects, called random effects, derived from the need to model equations for particular components of a study (Raudenbush, 1998).

Complementing the random effects, Raudenbush (1988) explained the need to model the macro-parameters of a study, called the fixed effects. According to Littell, Miliken, Stroup, Wolfinger, and Schabenberger (2006), the fixed effects of a model can be conceptualized in a few general ways: (a) as the treatment effect of a study and the primary variable to test; (b) as a variable whose levels have all been included in the model (e.g., gender); and (c) as a variable with many levels but the generalization of a variable is not intended to extend beyond the levels that are included in the model. Conversely, Littell et al. explained that one of the primary reasons to treat an effect as random is when the levels of a variable are believed to be drawn from a larger population of that same variable. For example, selected classrooms in a study would be a sample from the larger population of classrooms that could have been selected. Extending from fixed and random effects, HLM and mixed models allow for more complex applications. For example, it is possible to control for the random effects then test the fixed effects, or identify the variation in random effects that may explain variation in other variables. In addition, Raudenbush (1988) identified two core components that help explain the rise of these models within educational data structures; they can: (a) test models occurring between and within educational units, which can lead to a whole new class of testable hypotheses; and (b) model random intercepts and coefficients in order to identify proper error structures.

Extending these applications to model change, growth curve models treat the measurements as an additional nested level of the model. In particular, the measurements are nested within the individuals; therefore, the parameters of the measures are allowed to vary within each individual (Bryk & Raudenbush, 1992). Many of the advantages of latent growth modeling in SEM are also present in growth curve analysis in multilevel models, and specialized applications within each have been combined in many instances. According to Stoel et al. (2003), the distinction between the two methods is becoming less clear, and they hypothesized that they may one day merge into one set of procedures. Moreover, under certain conditions, LGM and growth curve analyses produced equivalent results, but produced discrepant results under different conditions (Curran & Hussong, 2003). Given these differences growth curve analysis may offer distinct advantages over LGM counterparts in some situations. In particular, according to Stoel et al. (2003), analyzing longitudinal data with multilevel models is better at handling missing data and varying measurement occasions. In addition, Shin, Espin, Deno, and McConnell (2004) found additional advantages in multilevel models compared to SEM and LGM in that multilevel models can handle smaller sample sizes and growth curve analysis can apply larger weights to growth rates that are more reliable (Raudenbush & Bryk, 1992).

While one of the benefits of LGM is its ability to model varying length between measurements, all participants are assumed to have been measured on the same occasions. In many situations, this is a reasonable assumption for data collection. With growth curve analysis, individuals can have their own data collection schedule, and this flexibility can extend predictors as they can be treated as time-invariant or time-variant; these changes in analyses need few adjustments to implement (Singer & Willet, 2003). To illustrate this flexibility, Biesanz, Papadakis, Deeb-Sosa, Bollen, and Curran (2004) used data from the principal author's previous research (Biesanz, West, & Graziano, 1998; Biesanz & West, 2000) and modeled the growth of self-esteem, across three measurements, for college students throughout a semester. Biesanz et al. described the data as full and unbalanced, since data were not available for all students and each student had a different data collection schedule. Modeling the measurements by day originating from the beginning of the semester, and conversely, originating from the end of the semester, Biesanz showed that the results were equivalent to six decimal places across all parameters. Within a school setting, the assumptions of consistent measurement occasions may be difficult to adhere to, especially given the number of measurements. It is not always possible for schools, classrooms, and students to get the data collected at the same point in time. These differences may be exacerbated by the fact that CBMs are meant to be sensitive to changes and not accounting for small variations in measurement occasions may not capture this sensitivity.

Applications with CBMs. A number of studies have used growth curve analysis to test an assortment of distinct hypotheses. Similar to many LGM studies, growth curve analyses have been used to examine growth rate as a function of each student's initial skills and slope but with the additional benefits of treating time flexibly (Silberglitt & Hintze, 2007; Stage & Jacobsen, 2001). For example, Stage and Jacobsen centered the intercept based on students' scores on the last probe administration since their analyses focused on students' scores at the end of the year.

Shin et al. (2004) used growth curves with CBMs to test whether student participation as part of a computer-based instructional system for math impacted achievement while controlling for specific demographic characteristics. In addition, they were able to model initial skills and determine whether growth rates were linear or quadratic, as they hypothesized that students with a learning disability would have a lag in their growth compared to their peers as growth would be slower initially but become more rapid over time. Shin et al. found that students with learning disabilities had similar growth trajectories as their peers and those students with greater participation had higher initial skills and higher growth rates. Therefore, in this study, researchers were able to determine whether an instructional program was effective while testing the hypothesis about the differences in learning between students with and without learning disabilities.

Growth as a Predictor of High-Stakes Assessment

One of the most far-reaching implications of IDEIA and RTI is the frequent measurement of student performance using CBMs. Within RTI models, instructional decisions and specific interventions are often a result of a student’s performance level on a CBM and growth across time. Therefore, schools are inundated with data of students’ performance and growth to make such decisions. In addition, since the passage of NCLB, educators spend the school year preparing students for high-stakes, statewide assessment and the broad implications of student performance on these assessments. Therefore, schools use frequent measurements to monitor academic progress throughout the year with RTI while concurrently preparing for end-of-year assessments. Naturally, with an abundance of RTI academic performance data, it may be in schools’ best interest to maximize the utility of these data to predict statewide test scores.

As noted earlier, researchers have begun to integrate information and use CBMs to predict performance; for example, in a number of studies researchers have used the scores derived from a single administration of CBM to predict achievement. While this method provides a relatively good estimation of students' performance level, it disregards the growth of students' skills over time. Moreover, this strategy does not maximize the use of the data that a school district will have collected throughout the year. As a result, researchers have begun to use students' rate of growth as a predictor of achievement (Baker et al., 2008; Chard et al., 2008; Keller-Margulis et al., 2008; Miller, 2012; Hinkle, 2011; Yeo, 2010; Yeo et al., 2011; Stage and Jacobsen, 2001).

Unfortunately, growth-rate studies suffer the same problem as the single administration studies; the research on the predictive potential of student growth on M-CBMs is much more limited than the R-CBM research. However, the initial research has yielded promising results. For example, Keller-Margulis et al. (2008) modeled the slopes of three administrations of math computation and math applications CBMs and found the correlation of the slopes for math computation with achievement tests two years post-measurement to be .42 for grade 1, .43 for grade 2, and .45 for grade 3.

Some of these studies (e.g., Keller-Margulis et al., 2008) appear to use the more traditional modeling of growth (OLS), which may inhibit the modeling of important parameters and limit the testing and comparison of alternative hypotheses. Results from studies using more advanced modeling techniques have tested multiple hypotheses and found somewhat different results than those from the more traditional analyses. Stage and Jacobsen (2001) used growth curves to predict end-of-year test scores and create cut-scores to predict students who passed or failed the test. Using three measurements, Stage and Jacobson found that the students' initial skills and growth were both significant predictors of the tests and could correctly identify the passage and failure of 74% of the sample using the cut-scores.

Yeo et al. (2011), using LGM with two different reading CBMs, found that the student growth did not significantly add to the prediction of an end-of-the-year assessment beyond students' initial performance level. In other words, modeling students' initial academic skills essentially accounted for the variability between academic growth and the end-of-the-year assessment. On the other hand, when Chard et al. (2008) performed a similar analysis using the results of growth from grade 1 through 3 to predict a norm-referenced standardized test, they found that growth was a far better predictor of the outcome variable than the initial first grade score. The similarities in the analytic techniques, but the differences in the methodology and results of these two studies, complicate the simple hypothesis that within-year growth is a significant predictor of high-stakes assessment.

Another limiting component of most studies that use student growth to predict high-stakes measures is that they generally only use three measurements to establish the slope. This limited number of measurements is problematic for two reasons. First, CBMs traditionally have high degrees of variability between alternate forms with large standard errors of measurement (Ardoin & Christ, 2009; Christ, 2006; Brown-Chidsey, Davis, & Maya, 2003; Christ & Ardoin, 2009; Hintze & Christ, 2004; Hintze, Christ, & Keller, 2002; Hintze, 2001; Hintze, Owen, Shaprio, & Daly, 2000; Hopkins, 2011; Poncy, Skinner, & Axtell, 2005). For example, Poncy et al. (2005) used Generalizability Theory (Cronbach, Nageswri, & Gleser, 1963) to partition and identify the sources of variation within sets of R-CBM probes of oral reading fluency. Since all 20 measures were given within 5 days, growth would not be expected to impact the variability. As a result, researchers found that 10% of the variance could be attributed to non-equivalence in the probes; probe variability could potentially account for a half a year's worth of growth. Furthermore, Ardoin and Christ (2004) found that growth with three measurement occasions was significantly impacted by the variability between probes. The other primary disadvantage in the previous research is the limited measurement occasions. In particular, growth models typically do better when there are more measurement occasions (Byrne, 2010). While these advanced modeling techniques can account for these sources of variation, the limited number of measurements and the expansive time between administrations may limit the extent to which growth can be modeled. As Yeo et al. (2011) concluded, the non-significant prediction using student growth may be a result of the instability of the slope. The restricted number of measurements is a limiting factor, regardless of the analytical method that is employed.

Miller (2012) attempted to overcome these limitations by modeling the slope of a full years' worth of measurements (12 measures), using growth curves to predict end-of-the-year high-stakes scores (TCAP) from R-CBMs that yields comprehension and reading rate scores. Miller found that the static score of comprehension on the midpoint administration (measure 6) was most highly correlated with the criterion measures (*r* = .31) while the slope across all 12 probes was also significantly correlated (*r* = .22) and significantly added to the overall prediction above the comprehension score. Miller compared the slopes of all 12 measurements to slopes with fewer measurements (i.e., every odd measurement, every even measurement, and every third measurement) and found statistically significant differences between the 12-measure slope and the slopes with fewer measurements; the 12-measure slope scores were the most predictive. In a follow up study, Miller et al. (2013) performed a multiple regression on the three alternative slopes predicting the end-of-the-year scores and found that using every even measurement did not add to the prediction above the other two slopes. These studies highlight the differences in slopes when using fewer measures and varying measurement time-points. Given these differences, it would appear that using all measurement components is the most effective and accurate method to model growth and to predict high-stakes assessment.

Growth and High-Stakes Predictive Potential of a Multidimensional M-CBM

Traditional mono-operational M-CBMs have shortcomings that may limit their predictive potential for high-stakes assessment; multidimensional M-CBMs may eliminate some of these limitations but research in this area is sparse. Recent developments in multidimensional M-CBMs have now made it possible to evaluate the growth and predictive potential of these instruments.

As Hopkins (2011) noted, most traditional M-CBMs generally target only one (mono-operational) set of skills and are prone to significant measurement error. Within the context of an RTI paradigm, this limitation raises concerns over the utility of these traditional measures. Considering the multiple mathematical skills assessed by high-stakes assessments, the utility of traditional M-CBM measures is further diminished. In order to compensate for the limited efficacy of a single-skill probe, a group of researchers created the Monitoring Instructional Responsiveness: Math (MIR:M) (Hopkins, McCallum, Bell, & Mounger, 2010) which maintains the efficiency of traditional M-CBMs but assesses multiple mathematical skill in grades K-5. Rather than relying exclusively on computation skills, MIR:M targets additional problem-solving elements, which may provide an important link between CBMs and high-stakes assessment given the multidimensional skills that are measured in both assessments.

The MIR:M was designed to target basic math skills and math problem-solving skills using a three minute, group administered format, and the difficulty level and measurement of these skills vary by grade level. For example, whereas 3rd grade students are required to solve addition/subtraction problems, fourth and fifth grade students solve multiplication problems. For the most part, the MIR:M assesses three general domains across grades 1 through 5 and one domain that changes between grades 3 and 4. Specifically, grades 1 through 5 maintain three problem types: Computation, Number Sentence-Quantity Discrimination, Shape Patterns, and Number Patterns. In grades 1 through 3, students complete a problem type called Shape Patterns, and this is substituted for Equation Completion for grade 4 and 5.

Hopkins (2011) established concurrent validity with the Monitoring Basic Skills Progress (MBSP) and found median correlations of .66 for grade 1, .41 for grade 2, and .52 for grade 3; however, the significantly lower reliability the MBSP for grade 2 may have impacted these results. Hopkins found very small standard errors of slopes when using all 12 instruments (.04 to .06) and larger standard errors when using only three measurements (.29 to .48). Finally, Hopkins found that the MIR:M was more predictive of end-of-the-year tests scores measured by Star Math (Renaissance Learning, 2002) than the MBSP.

Expanding on the initial research of Hopkins (2011), Coles, McCallum, and Bell (2013) established basic construct validity and determined that the best fitting psychometric model was consistent with the construction of the subtests and the two constructs of Math Reasoning (MR) and Math Computation (MC). These constructs were consistent with the research by Thurber et al (2002), who also found two factors of computation and application. In addition, MIR:M and its reading counterpart, the Monitoring Instructional Responsiveness: Reading (MIR:R; Hilton-Prillhart, Bell, McCallum, & Hopkins, 2009), have been used to create new applications of CBMs by using a discrepancy-based model to find students with academic strengths and weaknesses who may have a learning disability but may also be gifted (Coles, McCallum, & Bell, 2012; McCallum et al., 2013). This model was found to be predictive of students with significant weaknesses on end-of-the-year tests (Taylor, Hayes, Coles, McCallum, & Bell, 2014). In sum, the MIR:M for grades 1 through 3 have moderate to strong psychometric properties. This evidence is especially promising since general outcomes measures for M-CBMs have traditionally been difficult to establish.

The development of the MIR:M met the strict criterion for general outcome measurement as specified by Kelley et al. (2008); it assesses multiple domains and multiple skills within each domain. Furthermore, this format is also consistent with robust indicator CBMs (Fuchs, 2004) that measure general math proficiency. MIR:M provides a reliable estimate of a high-stakes assessment of general math proficiency, but additional psychometric data are needed to establish its utility.

Statement of the Problem

Within the current environment of accountability, efficiency, and high-stakes testing, inclusion of existing CBM data into an RTI framework is critical, as is examination of the relationship between efficient CBMs and high-stakes measures. While studies have investigated the predictive validity of CBMs and high-stakes measures, there are four major limitations in the current literature base. First, the overwhelming majority of studies have focused on the predictive validity of reading CBMs, while only a handful of studies have focused on math CBMs and even fewer investigated measures of multidimensional skills. Second, most predictive studies have used static scores from a single administration; very few studies have established the predictive potential of student growth. Third, most studies of growth have used only three measurements, called universal screeners within the RTI Model, given months apart. Limited measurements may be susceptible to measurement error and slope instability compared to bi-weekly measurements common in RTI progress monitoring. Finally, the majority of studies have used traditional analyses to obtain the student growth. These analyses lack a number of analytical options of more advanced methods like latent growth modeling and growth curve modeling.

Accounting for some of the methodological weaknesses of the current research mentioned above, some researchers employed a multi-dimensional M-CBM (MIR:R) to investigate then predict end of the year reading performance (Miller, 2012); however, the evidence for the predictive potential of the MIR:M is limited. The predictive validity data that are available (Hopkins, 2011; Taylor et al., 2014) were obtained primarily from scores from the third grade measures. Moreover, these analyses focused primarily on the prediction from the MIR:M composite score taken from one probe administration, rather than scores derived from multiple MIR:M components and multiple probe administrations. Additional predictive validity data are needed given the complexity of the MIR:M (i.e., the multiple skills assessed and represented by component scores, including Math Calculation, consisting of two subtests, Computation and Number Sentence-Quantity Discrimination and Math Reasoning, consisting of Equation Completion and Number Patterns, and the administration pattern. Typically, 12 administrations occur over the course of an academic year. Therefore, the primary purpose of the study is to determine the predictive power of various MIR:M measures (e.g., slope and intercept and using all academic-year administrations, various component scores, and impact of variable instructional time between probe administrations) when a high stakes, end-of-year measure, the Tennessee Comprehensive Assessment Program (TCAP) composite score (Tennessee Department of Education), is the criterion. A secondary purpose is to determine the best fit of scores of 12 MIR:M probe parameters as a function of modeling specific parameters (e.g., individual intercepts, slopes). Specific research questions are:

Research Questions

1. To what extent do the various parameters (e.g., individual intercepts, slopes, and differences in probe administration dates) of the three MIR:M composites account for variation of scores and provide the best model fit using all 12 MIR:M administration?

2. What is the relative power of MIR:M parameters (i.e., intercept and slope) obtained from all administrations within one academic year to predict the TCAP Math composite score (e.g., raw or scaled) when modeling each MIR:M composite individually?

3. What is the best predictive model using all parameters of the MIR:M global scores (i.e., Math Reasoning and Math Calculation) when the TCAP Math composite score is the criterion?

4. What is the best predictive modeling using all parameters of the three MIR:M composite scores when the TCAP Math composite score is the criterion?

CHAPTER II

METHOD

Participants

Participants from this study were tested within classrooms from a sample of fifth-grade students within one school district, comprised of eight elementary schools in East Tennessee. That is, personnel from each school were asked to choose at least one classroom that provided the most representative sample of skills and demographics from the school as a whole. Simply put, personnel from each school attempted to pick classrooms characterized by a diverse subset of students that best exemplified skills and characteristics of the larger student body. The two largest schools picked three classrooms. Out of 604 fifth-grade students, 262 (43%) were included in the sample. Students missing more than 25% of the probes administered (< 9 probes) were deleted, leaving 223 (37%) of the total fifth grade population in the study. Students were primarily Caucasian (92%) and 51% were Males; nearly 60% of students in this district are considered disadvantaged; that is, they are eligible for either a free or reduced price lunch (Miller, 2012).

Instruments

Monitoring Intervention Responsiveness: Math As noted by Hopkins, 2011, the probes were designed to assess academic subtests and objectives for each specific grade level, as set forth by the National Council of Teacher Mathematics Curriculum Standards (National Council of Teachers of Mathematics, 2000) and the State of Tennessee’s scope and sequence of mathematics (Tennessee Department of Education, 2004). Therefore, as students progress through the year, they should have exposure to the skills assessed by the MIR: M.

Students were administered the Monitoring Instructional Responsiveness: Math (MIR:M) as part of implementation of the RTI model within the system. MIR:M is a brief three minute curriculum-based measure (CBM) of basic math skills, and consists of three universal screeners and nine progress monitoring probes administered bi-weekly throughout the school year. Typically, Universal Screeners are administered to all students to assess their skills at the beginning, middle, and end of a given school year and serve as a screener for skill deficits that need further monitoring. Progress Monitoring probes are designed to monitor student progress as instruction changes within the RTI framework.

Using scripted instructions, students are requested to silently complete as many MIR:M math problems as possible within three minutes. Teachers were trained to administer the probes by university-based authors (of the probes) with a set of standardized instructions detailing the procedures for completing each type of math problem. After scoring the probes, teachers entered students' total scores into a district-wide database. Each student can earn two global scores, Math Calculation (MC) and Math Reasoning (MR). Each global score is comprised of even numbers of two item types. The MC global score is comprised of Computation and Number Sentence-Quantity Discrimination and the MR global score is comprised of Number Patterns and Equation Completion. Scores can be derived, but typically are not differentiated and made available at the subtest level. Students could achieve a maximum score of 98 points on the total MIR:M probe. Support for use of the component scores were obtained by Coles et al. (2013), who found that the hypothesized factor structure of Math Calculation and Math Reasoning provided the best fit (Root Mean Square Error of Approximation; RMSEA = .012) when compared to a one factor solution (RMSEA = .055) for the third grade MIR: M probes. Furthermore, this model provided a nearly perfect fit when analyzing students that scored at least one point on each of the four subtests. The third-grade measures that were analyzed in the Coles et al. study included item-types that directly parallel item-types from the current study's measures.

Thus far, MIR:M authors and colleagues have focused on obtaining psychometric properties for grades 1 through 3 (Hopkins, 2011; Coles, McCallum, & Bell, 2012; McCallum, Bell, & Coles, 2012; McCallum et al., 2013; Coles, McCallum, & Bell, 2013; Taylor, Coles, Hayes, McCallum, & Bell, 2013). Hopkins (2011) found an average alternate-forms reliability to be .73 for grade 1, .66 for grade 2, and .72 for grade 3. Since administrations occurred throughout the year, the reliability was expected to decrease as the spacing between measurements increased. Accounting for the time effect, average reliability coefficients of subsequent administrations of the MIR:M, given two weeks apart, were .78 for grade 1, .73 for grade 2, and .77 for grade 3. To further establish reliability, Hopkins used a Generalizability Study and found that the generalizability coefficients increased to .85 for grade 1, .80 for grade 2, and .85 if two probes were administered instead of 1, thus establishing an alternative application to obtain more reliable results.

Math Calculation. Math calculation (MC) consists of 16 items, eight items from two subtests, Computation (COMP) and Number Sentence-Quantity Discrimination (NSQD). For the COMP items, students are required to solve 2 X 2 multiplication and 3 X 3 multiplications problems and division problems with three digit numbers divided by a single digit. Probes are scored based on the number of digits correctly placed. For the NSQD items, students are required to solve a horizontally presented number sentence then solve a quantity discrimination task by comparing the number they supplied to a randomly supplied number by circling the requisite sign (i.e., <,>,=) (Hopkins, 2011). Specifically, NSQD items are structured as addition and subtraction with fractions, incorporating single digit numerators and denominators. Probes are scored based on digits correctly placed and the discrimination sign correctly circled. If students incorrectly solve the number sentence, their quantity discrimination answer is scored based on the number they supplied in relation to the random number given. If students do not supply an answer to the number sentence, the quantity discrimination score is automatically scored as incorrect, whether they circle a discrimination sign or not. Students can score a maximum of 12 points on the NSQD problems.

Math Reasoning. Math reasoning (MR), consists of 16 items, eight items from each of two subtests, Equation Completion (EC) and Number Patterns (NP). For the EC items, students are given an arithmetic equation with an addition problem on one side and a subtraction problem on the other. One side of the equation is completed while the other side has a number missing designated by a blank line. Students are required to supply a number in the blank space that ensures that both sides are equal. For the NP items, students are required to correctly identify and place missing numbers to complete a numerical sequence ordered from least to greatest. The placement of the missing numbers, pattern of the sequence, and numbers, are randomly assigned based on grade specific parameters (Hopkins, 2011).

Tennessee Comprehensive Assessment Program Test. At the end of each school year, students are given the Tennessee Comprehensive Assessment Program (TCAP), a criterion-referenced, statewide assessment of academic skills and achievement in math, reading, social studies, and science given to students in grades 3 through 8 (Tennessee Department of Education). Each test is designed to assess a student's performance in state content standards. For each academic area, students receive a scaled scored and achievement composite from the scaled score.

Although psychometric properties of the current TCAP tests do not appear to be widely available, Ciczek (2005) reviewed the Second Edition of the TerraNova, a precursor to the TCAP. Ciczek reported internal consistency coefficients of .95 to .96 but no test-resest reliability. Cizek noted adequate content validity and limited concurrent validity (Miller, DeLapp, and Driscoll, 2007). Harwell's (2010) review of Third Edition of the TerraNova, the overall technical properties appeared consistent with the Second Edition.

In a memo addressing issues related to the TCAP, Dan Long, Executive Director of the Office of Assessment and Evaluation at the Tennessee Department of Education, and Marcy Tidwell, Associate Director of Assessment Literacy, explained the process for creating test-items (Long and Tidwell, n.d.). They explained that Educational Testing services (ETS) creates items and Tennessee educators participate in a review of the items prior to field-testing. ETS then field-tests items and includes only those that meet acceptable statistical standards. Specific psychometric properties were not presented.

Specifically, for fifth grade math, students are given 64 items with scaled scores ranging from 600 to 900. Scaled scores are then used to determine each student's four achievement composites based on proficiency levels: Below Basic, Basic, Proficient, and Advanced. For fifth grade mathematics, scores from 600 to 727 are considered Below Basic; scores from 728 to 763 are considered Basic; scores from 764 to 794 are considered Proficient; and scores from 795 to 900 are considered Advanced.

Each academic area is also comprised of subscales that address specific content standards within that subject area with scores ranging from 0 to 100. This score, called the Reporting Categories Performance Index (RCPI) is defined as "an estimate of the number of items the student would be expected to answer correctly to achieve basic, proficient and advanced designation if there had been 100 such items for each category" (Tennessee Department of Education). The RCPI scores are also reported in three composites, Basic, Proficient, and Advanced, to determine a student's relative achievement within the content standards assessed. Math skill areas are: Mathematical Processes, Number and Operations, Algebra, Geometry and Measurement, and Data Analysis, Statistics, and Probability.

For Mathematical Processes, students are expected to estimate decimals and fractions; draw a conclusion about a figure; draw conclusions about a geometric figure from given statements; identify missing information in contextual problems; and recognize unit in remainder from division problems of the fractional part of a whole. For Numbers and Operations, students are expected to understand numbers from millions to millionths; write the prime factorization of numbers; identify the reasonable conclusion to applied division problems with remainders; solve 2 and 3 digit by 1 and 2 digit problems; add and subtract decimals, proper fractions, improper fractions, and mixed numbers; identify equivalent representation of numbers; convert decimals to fractions; compared whole numbers, fractions, and decimals.

For Algebra, students are expected to solve expressions with fractions, decimals, and multi-step numerical problems using order of operations; find a variable in single-step fraction and mixed-number problems; and identify values that make an inequality true. For Geometry and Measurement, students are expected to solve contextual problems by calculating the area of triangles and parallelograms; find the perimeter and area of irregular shapes; identify three dimension objects from two dimension representations and identify a two dimensional object from three dimensional representations; find the surface area and volume of rectangular prisms and polyhedrons; find the length of lines within coordinate system; and record measurements using decimals and fractions. Finally, for Data Analysis, Statistics, and Probability students are expected to describe data; make predictions using graphs and other visual representations; and analyze dating by computing the central tendency.

Procedures

All procedures were implemented within a school district in East Tennessee. One to three classrooms were selected from each school, as previously described. TCAP data were collected by educators from each school during the spring of the year in which this study took place (fall 2010-spring 2011).

Before MIR:M data collection, written permission was obtained from the University's review board, the district level superintendent, and building level administrators. Teachers and school personnel were trained to administer and score the MIR:M by the probe authors, using a script. The MIR:M probes were administered within the classrooms approximately every two weeks for a total of 12 administrations. The first, sixth, and 12th administrations were designated as Universal Screeners, while the remaining probes were progress monitoring probes. The Universal Screeners were given to all students within the district while the progress monitoring probes were only given to the pilot sample. Although the administrations were intended to be given every two weeks, the time between administrations was not always constant because of vacations, snow days, and other unforeseen circumstances. In addition, the dates of administrations also varied by the school. However, the administration dates between schools tended to be within one to two weeks of each other and order of administrations was consistent.

Probe Administration. MIR:M probes were administered to fifth- grade students in a group format by their regular classroom teachers. Prior to administration but after training, teachers completed a practice test themselves. This allowed teachers to have an opportunity to get acquainted with the practice, administration, and scoring procedures of the test to raise any questions.

After the initial training, teachers began administering the probes to their students. Each teacher read to the students a set of standardized directions. In addition, students were given a practice administration of each subscale and were given time to ask questions to ensure they had an understanding of the procedures. Teachers gave the students three minutes to answer as many math problems as possible within the allotted time and immediately collected the probes. These exact procedures were followed for each administration. Although probes were designed to be given every two weeks, there was significant variation between probe administrations between schools and probes. Specifically, within each school, students were given the probes at the same time, although there was variation between each schools' administration dates. In addition, there was significant variation between probes, from as short as one week to longer than a month. TCAP achievement tests were administered by classroom teachers in accordance with the Tennessee Department of Education's administration guidelines and procedures.

Scoring. Initial scoring and data entry was completed by school personnel. Initial scoring of the MIR: M yielded a Math Calculation composite, a Math Reasoning composite, and a Total Score composite. In order to obtain each composite score, each column of problems had boxes that mark the number of total points a student can receive on each problem. A vertical line separates the Math Calculation problems from the Math Reasoning problems. At the bottom of each column, there is a space to sum the points for each composite. Once the total for each column of problems has been calculated, these four totals are combined to obtain a total MC and total MR composites score. These two score are summed to create the total score for that administration.

In order to obtain the scores for each subscale, this author and four graduate research assistants from the research team supervised by the probe authors, tallied scores from each probe and separated the scores by subscale. This process allowed for initial data cleaning since each score was checked for any discrepancies between the initial scoring and the graduate students' scoring. When a discrepancy was present, the data were reevaluated to determine the cause of the discrepancy and the correct score.

The research team extracted students' TCAP Math scaled score, Math Achievement composite, and the RCPI score from the five subscales. Given the number of problems reportedly given from each subscale, the graduate assistants were able to calculate the total number of problems correct for each of the subscales and the total.

Data Cleaning

Although the flexible statistical modeling techniques of this study were capable of handling missing data, some students had missing data from multiple MIR:M administrations.. Since a limitation of prior research was the infrequent administrations of CBMs, it was imperative to collect as much data for each student as possible. Modeling students with significantly fewer than all 12 probes is counter to this goal. In addition, since the cause of missing data may be known to school personnel, but not the researchers, it may not fit the definition of "acceptable" missing data (e.g., missing at random; Little & Rubin, 1987) the "missing data" effect has the potential to bias the results. Furthermore, although the number of instructional days was the primary time variable, there was no information about student absences (e.g., why they were absent, how many days absent). A reasonable number of absences should not significantly impact the results; conversely, numerous absences may impact the results due to lack of instruction. Because of these concerns, students with less than 75% of all MIR:M scores (i.e., < 9 probes) were excluded from the analyses.

Because the second focus of this study was to predict TCAP scores, students without a TCAP score were excluded in the analyses. In addition, school districts may allow two-percent of students with a disability to take the Modified Academic Achievement Standards (MAAS; Tennessee Department of Education). Although the student's taking the MAAS were assessed on curriculum content areas consistent with the TCAP, the administration and questions of the assessment were different and the scores were on a different scale. These students were excluded from the analyses. Following the data cleaning, a Total of 223 students were included in the final analyses.

Data Analyses

The research questions of this study required two distinct modeling procedures. The initial set of modeling procedures identified various growth parameters for each composite and determined the extent of the variation that can be attributed to these parameters. The predictive power of these parameters was determined using TCAP scores as the criterion for the second set of modeling. To differentiate the distinct parameters of the two procedures, the following sections provide a framework of the modeling and the specific information about the parameters and their notation.

Modeling MIR:M Growth. In order to provide the most useful analysis of the MIR: M data, LGM and growth curve analysis were considered. Although under certain assumptions, the results are often equitable, certain conditions can be better handled by one technique or the other. In particular, the growth curve analysis has the ability to model variable measurement occasions (times) and can handle missing data better than LGM. Given the varying measurement occasions and the missing data, growth curve analysis was deemed the most appropriate technique for these data.

Using the Total score from the MIR:M, each instructional day was modeled as the time parameter along with the probe administered. Time was conceptualized as a function of instructional days with the first day administration of the MIR:M considered day one and an increase occurred only when students were in school. This strategy seemed to be the most accurate representation of time to model growth for a number of reasons. First, the MIR:M was designed to model instructional responsiveness. Therefore, the opportunity for growth should be greatest when the time between administrations has maximum instructional days. For example, suppose one set of probes was given two weeks apart and another set of probes was given a month apart because of a school vacation. The number of calendar days that passed would be 14 and 30 days respectively. Now if students were only in school for 10 days for both intervals, then the opportunity to respond and progress as a result of in-school instruction would be equal. As a result, the doubling in time between the intervals in the first time variable may not be an accurate representation of growth. In addition, a secondary parameter of time, conceptualized as a function of the total number of days (i.e., school days, weekends, and breaks), was used to determine if long breaks from school contributed to the variation of scores.

MIR: M Growth Parameters Modeling. The first analysis involved modeling the student growth on the MIR:M MR, MC, and Total scores separately. The procedures for Growth Curve Modeling were consistent with the model building specified in Singer and Willet (2003); however, the notation of the models was consistent with the notation specified in Raudenbush and Bryk (2002) to better differentiate the variance components. The first model, the unconditional means model, partitioned the variance across individuals while disregarding the impact of time. This model was used to determine if there was significant within-student variation within the MIR:M scores to warrant further analysis. It also quantified the amount of variation within- and between-students. This model provided two basic levels, the level-1 model of the random, individual parameters, and the level-2 model of the fixed, population level parameters. The level-1 unconditional means model was written as $Y_{ij} = \pi_{0i} + \varepsilon_{ij}$ and the within-student variance denoted by σ^2. The level-2 unconditional means model was written as $\pi_{0i} = \beta_{00} + r_{0i}$ and the between-student population level variance denoted by τ_{π}.

The next tested model, the unconditional growth model, built upon the unconditional means model by adding the time variable as a level one predictor. This model helped determine the overall amount of change that occurred as well as the extent of variation that occurred within and between individuals due to changes in probe scores. Furthermore, this model provided a baseline model for comparison of later models. The level-1 unconditional growth model was written as $Y_{ij} = \pi_{0i} + \pi_{1i}TIME_{ij} + \varepsilon_{ij}$. More specifically, π_{0i} represents the intercept, $\pi_{1i}TIME_{ij}$ represents the rate of change of individual i from the population, and ε_{ij} represents the error term. The level-2 unconditional growth model parameter was written as $\pi_{1i} = \beta_{00} + \beta_{10}TIME_{ij} + r_{0i}$. In particular, β_{00} represents the intercept, $\beta_{10}TIME_{ij}$ represents the average rate of change for all students, and r_{0i} represents the error term.

To test the impact of student's classroom (and teacher), an additional classroom level nested model was analyzed. Thus, this level-3 between-classroom unconditional means model was written as $\beta_{0jk} = \gamma_{00k} + \mu_{00k,}$ with the variance denoted by τ_{β}. The between-classroom unconditional growth model was written as $\beta_{10k} = \gamma_{10k} + \gamma_{10k}TIME_{ijk} + \mu_{1jk}$. In particular,

γ_{10k} represents the intercept, $\gamma_{10k}TIME_{ijk}$ represents the average rate of change for classroom k, and μ_{1jk} represents the error term.

Additional predictors were analyzed by adding each separately. If more than one predictor was significant, then later models would test the combination of predictors. These modeling procedures were used separately for the MIR:M Total composite score, MIR:M Math Reasoning Global composite score, and MIR:M Math Calculation Global composite score. Restricted Maximum Likelihood Estimation (RMLE) was used since it provided less biased estimation of variance components than Maximum Likelihood Estimation (MLE); however, RMLE allows only model comparisons of the variance components with identical fixed effects (Singer & Willet, 2003). MLE was used only to compare the model fit when the fixed effects were changed. RMLE was used for all other model comparisons and the estimation of coefficients and variance components.

The daily time component was transformed to a bi-weekly component. This only affected the scaling of time and provided two advantages. First, the estimates of growth were easier to interpret, as daily change would not have much practical significance as the estimates would be very small; instead, two weeks provided more time for growth and hence a more practical estimate of trend. In addition, these scores were consistent with the approximate bi-weekly administration schedule of MIR probes.

TCAP Predictive Modeling. TCAP prediction modeling was based upon the parameters from the MIR:M growth modeling procedures. Once the growth analyses were complete, each student's individual parameters (intercept and slope) were derived and transformed into new variables; these were the primary predictive variables. To maintain consistency, the same parameters were derived from each of the three composites. An additional variable for each composite was derived using the prediction of the MIR:M score from the growth models; the prediction was set to the day of the TCAP math administration. The exact date of the Math administration was not known as each of the TCAP tests (i.e., Reading, Math, Social Studies, and Science) was given in one of four consecutive days (i.e., school days 117-120 of study). Since the TCAP math components comprised the second set given for each student, it was assumed that it was given on the second day of TCAP administrations (i.e., school day 118).

Modeling Predictive Parameters. Since high-stakes tests (i.e., TCAP) scores are a primary tool for evaluating students *and* teachers, the predictive parameters of the MIR:M differentiates student and teacher level predictions. This process partitions the sources of TCAP variation at different levels and fits an unconditional means model, a one way random effects ANOVA, with the classroom (i.e., teacher) intercept as the only predictor to identify the between-student and between-teacher variation effects. Therefore, the student level model is defined as $\text{TCAP\_Raw}_{ij} = \beta_{0j} + \varepsilon_{ij}$ with the level two, teacher model as $\beta_{0j} = \gamma_{00} + \mu_{0j}$.

Centering of the component's values was used to delineate the predictive potential of components at both levels and provide a clearer interpretation of the coefficients. The primary centering method involves group-mean centering by classroom. This allows for separate modeling of between-group (i.e., teacher) and within-group (i.e., students within classroom) estimates, which can identify complex cross-level interactions (Bryk & Raudenbush, 1992, Hoffman & Gavin, 1998).

The first step of the group-mean centering involved computing the mean score of each MIR:M component (e.g., intercept) by class; this was the class average for each teacher. This approach tests the effect of each of the MIR:M components between teachers. Next, the grand mean, the overall mean of the entire sample for each component, was subtracted from the teacher mean. Therefore, a classroom that, on average, had higher Total intercept scores than the overall sample would have a positive value; a classroom with lower Total intercept scores would have a negative value. Centering on the grand mean made the TCAP intercept the expected TCAP value when the intercept, linear slope, and quadratic slope were each at the grand mean (i.e., 0 for each value).

Next, students' scores for the components were centered on their respective teacher's mean. In other words, students' values represented their score within their respective classrooms. Therefore, for the student-level data, the intercept was the expected TCAP value for the average student's score within each classroom. While a student's raw score could be higher than a student in a different class, the higher-performing student's value may be less than the lower-performing student lower if the class average is higher; however, the class average was accounted for at the teacher level. Altogether, this permitted the examination of the classroom effects as well as the individual effects within each classroom. Take for example the prediction of TCAP using the MC intercept. The level-1 student level model would then be TCAP_Raw$_{ij}$= $\beta_{0j} + \beta_{1j}MC\_Intercept - MC\_TeacherIntercept_j + \varepsilon_{ij}$. The level-2 teacher level models would then be $\beta_{0j}= \gamma_{00}+\mu_{0j}$, for the model intercept and $\beta_{1j}= \gamma_{01}MC\_TeacherIntercept - MC\_GrandMean_j+\mu_{1j}$ for the slope of the MC intercept variable.

In addition, grand mean centering at the student level was also completed for each of the variables. This was necessary since the modeling involved highly correlated predictors. Moreover, it is possible that level-1 variable (i.e., student level) would be significant while the level-2 variable (i.e., teacher level) would not be significant. Bryk and Raudenbush (1992) recommend that when there are significant individual-level predictors in a model, group means should be included to estimate the correct between-groups effect. That being said, including non-significant variables at the teacher level may complicate the model. In particular, teacher sample size was small (i.e., 12 teachers) which limited the degrees of freedom available; this procedure also adds a risk of multicollinearity in the model. Thus, grand mean centering ensured that both levels were accounted for, which can alleviate the multicollinearity problem without using valuable degrees of freedom. Unfortunately, inclusion of grand mean centered variables for each student can affect both levels of the model, making it difficult to interpret the impact at each level (Hoffman & Gavin, 1998). Even so, because this procedure followed the initial centering method, a general understanding of the variable's impact on each level was possible to obtain. Since grand mean centering was used with non-significant teacher-level variables, this decreased the likelihood that these variables would have a significant impact at the teacher level.

Because the two-level model had medium to high correlations between variables, an initial test of multicollinearity with all 12 MC and MR group mean centered variables was performed. Although multicollinearity did not appear problematic for the group centered individual scores (i.e., Variance Inflation Factors < 3), the teacher level means revealed multicollinearity, most notably the Slope and quadratic components (i.e., Variance Inflation Factors > 5). An analysis of the variance inflation using the 12 variable model with grand mean centered variables indicated that this procedure alleviated the concern for multicollinearity.

CHAPTER III

RESULTS

Descriptive Statistics for MIR:M Scores

Descriptive data for the MIR:M Total composite scores, Math Calculation global composite scores, and Math Reasoning global composite scores are presented in Tables 1, 2, and 3 respectively. Table 4 provides correlations of MIR:M Total by probe administration; Table 5 provides correlations of MIR:M Math Reasoning by probe administration; and Table 6 provides the correlations for the MIR:M Math Calculation by probe administration. Because MIR:M probes were designed to be given every two weeks, consistent administration of the probes could still take advantage of considerable time between measurements. However, the actual time between probe administrations was variable and there were differences between the dates of administrations between schools. The administration variability could contribute negatively by impacting magnitude of these coefficients, and they should be interpreted with caution.

As expected, Total composite and the Global composite scores correlation coefficients strengthened as the time between administrations decreased. In addition, this relation became more pronounced as the year progressed. In other words, the correlation coefficients between adjacent probes were weaker at the beginning of the year than at the end of the year. For example, the median correlation of the MIR:M Total global score of the first administration with the following 11 administrations was .47 (.45 to .56); conversely, the median correlation for the last administration with the preceding 11 administrations was .66 (.51 to .82). This pattern of relationships was present in the MR and MC Global Scores as well. The first administration had the weakest relation with the other 11 administrations.

Modeling Parameters of MIR:M

To address research question one, the first modeling of MIR:M growth identified the variation that can be attributed to the specific MIR:M parameters and probe differences. A graphical depiction of the MIR:M Total mean score at each administration shows that there may have been a leveling of scores as the year progressed. Figure 1 represents a graph of these scores as well as a linear and quadratic trend (also referred to as growth or slope) line. Both the linear growth model, and the quadratic growth model were fitted to the data to determine whether linear or non-linear trend lines represent the best fit to the MIR:M scores. Tables 7, 8, and 9 represent the respective growth models for the MIR:M Total, MR, and MC composites. Each model represents the addition of a parameter, with fixed parameters (e.g., fixed linear trend) added first followed by random parameters (e.g., random linear trend). The most substantive growth models (discussed below) involved the modeling of the same fixed and random trend (e.g., both fixed and random) parameters. The other models provided a stepwise comparison of each fixed and random parameter, independent of the other parameters. Therefore, in Tables 7, 8, and 9, models A, C, and F represent the most important models.

MIR:M Total Composite Growth. Table 7 presents the various models fitted for the MIR: M Total composite. Model A represents the initial unconditional means model. This model indicated that 42% of the variation in the MIR:M Total can be attributed to within-student variability (σ^2 = 45.49) while the remaining 58% was attributed to differences between students (τ_π = 62.62). Model D represents the unconditional growth model of MIR:M scores. The growth model indicated that 24% of within-person variation of the MIR:M Total was attributed to the linear growth of students; it accounted for an additional 29% of the between-student differences

(τ_π = 46.39). Therefore, the average intercept across students was 15.82 (*SE*= .50) and students gained an average of .48 (*SE*= .05) Total points every two weeks.

Fitting a model employing a quadratic trend decreased the within-student residual (σ^2 = 33.35) by 4% from the unconditional growth model; the between student residual (τ_π = 46.01) increased slightly. The overall model fit (*-2LL* = 17035) was better and indicated that the quadratic growth model provided a better representation of the actual data. Thus, the average intercept across students was 15.23 (*SE* = .55) with an average linear growth .763 (*SE* = .13) and quadratic trend of -.02 (*SE* = .01).

Teacher Nesting and Probe Variability. Students were nested within teacher classrooms and classrooms (and teachers) may be considered a substantive contributor to variation. 16% of the variation of the overall model (τ_β = 17.67) was a result of teacher level differences; furthermore, this accounted for 27% of the variation (τ_π = 45.15) between students compared to the non-nested unconditional means model (i.e., Table 7, Model A). The linear growth of this nesting decreased the between-teacher variance (τ_β = 8.67) by 51% from the unconditional growth model (i.e., Table 7, Model C). The between-student variance (τ_π = 35.83) decreased by 21% from the three-level unconditional means model and 19% from the non-nested unconditional growth model. Adding an additional quadratic trend increased the between-teacher variance (τ_β = 11.27) but decreased the between-student variance (τ_π = 32.78) 9% from the previous model and 29% from the non-nested quadratic growth model. The within-student residual of the nested models were similar to the non-nested models.

Although the 12 probes were designed to be equivalent, it was important to test whether this assumption was met, and the extent to which probe differences could account for variation in the data. The probes were treated as a fixed effect given that all 12 of the available probes were available and not as a random variable since these probes were not considered a subset of possible probes. However, since Generalizability Theory (Cronbach, Nageswari, & Gleser, 1963) studies often treat it as a random parameter (e.g., Poncy et al. 2005), ML estimation was used to compare the overall model fit when treating probes as a fixed effect compared to a random effect. Comparison of the overall fit of the fixed probe effect (*-2LL* = 16857.9) to the random probe effect (*-2LL* = 16911.3) indicated a better fit for the fixed effect, $X^2(10) = 53.4$, $p < .0001$; this validates the fixed probe assumption for subsequent analyses. Adding the probe variable to the non-nested quadratic model decreased the between-student variance ($\tau_\pi = 45.03$) by 2% and the within-student variance ($\sigma^2 = 30.84$) by 8%. When added to the nested quadratic model, it decreased the between-teacher variance ($\tau_\beta = 9.60$) by 15% but increased the between-student variance slightly ($\tau_\pi = 33.38$).

MIR:M Math Calculation Global Growth. Table 8 represents the various models for the MC global score. Model A, the unconditional means model, indicated that 48% of the variation was attributed to the within-student variation while 52% was attributed to between student variation ($\tau_\pi = 25.65$). Model D, the unconditional growth model, indicated that the within-student variance ($\sigma^2 = 16.06$) decreased from the unconditional means model ($\sigma^2 = 23.47$). Therefore, 32% of the within-student variation of MC scores was attributed to individual linear growth. The individual growth modeling decreased between-student variance ($\tau_\pi = 14.56$) by 43% from the unconditional means model. The level-2 model indicated that the average student's initial MC score was 9.136 (*SE*= .30). Across the year, the average student's MC score increased by .51 points in two school weeks (*SE*= 0.03).

The addition of the quadratic trend resulted in a slight increase in the between-student variance ($\tau_\pi = 15.46$) but a 4% decrease in the within-student variance ($\sigma^2 = 15.34$). Although the level-2 linear trend was significant, β = .63, t(2307) = 6.95, $p < .0001$, the level-2 quadratic trend was not significant β = -.009, t(2307) = -1.42, p = .16. ML estimation of the overall fit with the level-2 quadratic trend (*-2LL* = 14972.5) compared to the exclusion of the quadratic trend (*-2LL* = 14974.5) indicated that it did not significantly improve the model $X^2(1) = 2.0$, p = .16. The better fitting model slightly decreased the between-student variance (τ_π = 15.37) and slightly increased the within-student variance (σ^2= 15.345); it provided a marginal decrease in variance across the two levels.

Teacher Nesting and Probe Variability. The three level unconditional means model nested within teachers, indicated that between-teacher variance (τ_β = 10.50) accounted for 22% of the total variation in the model. Furthermore, application of this model decreased the between-student variation (τ_π = 14.84) by 42% from the two level unconditional means model. Adding the linear trend decreases the between-teacher variance (τ_β = 5.47) by 48% from the previous model. Furthermore, this decreased between-student variance (τ_π = 8.99) by 40% from the previous model and 39% from the unconditional growth model (i.e., Table 8, Model C). The three level quadratic model, without the fixed quadratic trend, increased the between-teacher variance (τ_β = 8.83) but decreased the between-student variance (τ_π = 6.44) by 29% from the previous model; this resulted in a 58% decrease in variance from the two-level quadratic model.

Adding the fixed-probe effect on the random quadratic model with a fixed linear trend, resulted in a 4% reduction in the between-student variance (τ_π = 14.70) and a 4% decrease in the within-student variance (σ^2= 14.69) from the comparable model without this probe effect. Across the two levels, this decreased the overall variation by 4%.

MIR:M Math Reasoning Global Composite Growth. Table 9 represents the unconditional means and growth models for the MR global score. Model A, the unconditional means model, indicated that 45% of the variation can be attributed to within-student differences. Model D, the unconditional growth model, indicated that the within-student residual (σ^2= 20.78) decreased from the unconditional means model (σ^2= 25.65). Thus, 19% of the within-student variation in MR scores was attributed to the individual linear growth. The between-student variance (τ_π = 26.35) decreased from the previous model (τ_π = 31.33) with a 16% reduction of the within-student variance. The level-2 model indicated that the average student's initial MR score was 6.69 (*SE*= .38). Across the year, student's MR decreased slightly -.04 points (*SE*= .04) every two weeks; this decrease was non-significant $t(2307) = -0.9$, $p = .37$, indicating that the average student's MR score was relatively consistent across a school year. Compared to the unconditional means model (*-2LL* = 16000.7), the unconditional growth model (*-2LL* = 15766.3) provided significantly better fit, $X^2(2) = 234.4$, $p < .0001$.

Adding the quadratic trend to the model decreased the between-student variance (τ_π = 20.90) by 21% and the within-student variance (σ^2= 19.91) by 4%. Neither level-2 linear trend, β = .12, $t(2307) = 1.15$, $p = .25$, nor quadratic trend, β = -.011, $t(2307) = 1.15$, $p = .25$ were significant. The overall fit (*-2LL* = 15734.0) was significantly better than the unconditional growth model, $X^2(4) = 32.3$, $p < .0001$; however, fit estimates were obtained through RMLE and only the random parameters were tested. ML estimation indicated that the level-2 quadratic trend (*-2LL* = 15721.0) did not improve the fit above the level-2 linear trend model (*-2LL* = 15723.8), $X^2(1) = 2.8$, $p = .09$ nor the unconditional level-2 model (*-2LL* = 15725.4), $X^2(2) = 4.1$, $p = .12$. RML estimation without any level-2 predictors resulted in small decreases in the between-student variance (τ_π = 20.86) and the within-student variance (σ^2= 19.90).

Teacher Nesting and Probe Variability. The three level unconditional means model with students nested within classrooms indicated that the between-teacher variance (τ_β = 1.26) accounted for 2% of the variance in the model. In addition, compared to the non-nested unconditional means model, the between-student variance (τ_{π} = 30.18) decreased by 4%. Compared to the two-level unconditional means model (*-2LL* = 16000.7), the overall fit of the three level nested unconditional means model (*-2LL* = 15998.9) did not significantly improve the fit $X^2(1) = 1.8$, $p = .18$. This indicated that the three level, nested model was not necessary and no further tests of this model were used. The final test of the MR growth model was to determine the impact of the probes on the best model. Modeling the fixed effect of the probe with the quadratic model and no level-2 time predictors resulted in an increase in the between-student variance (τ_{π} = 21.52) but a 5% decrease in the within-student variance (σ^2= 18.81).

Summary of Composite Growth Modeling. For the two-level growth modeling, quadratic models were found to provide the best representation of change across the Total, MC, and MR composites. Compared to the initial unconditional means model, these models accounted for significant variance within-students (i.e., level-1) and between-students (i.e., level-2). In particular, the within-student variance reduction was 27% for the Total composite, for the 35% MC composite, and 22% MR composite. The between-student variance reduction was 27% for the Total composite, for the 40% MC composite, and 33% MR composite. Across both levels, the overall variance reduction was 27% for the Total composite, 38% for the MC composite, and 28% for the MR composite.

An additional teacher level accounted for substantial variance in the Total composite (23%) and MC Composite (16%). quadratic trends provided the best fit for the three-level Total and MC growth models. The teacher level did not account for substantial variation in the MR composite (2%). The use of alternative probes was found to have a significant impact on the variation across the three composites. Specifically, modeling the alternative probe decreased the within-student variation by 8% for the Total composite, 4% for the MC composite, and 5% for the MR composite.

Alternative Modeling of Time. A secondary set of analyses was conducted to determine the most appropriate conceptualization of time; these were the last analyses associated with research question one. In particular, these analyses tested whether modeling every day, regardless if students were in school, provided a better fit than modeling school days. Using the unconditional growth model, MLE was used to compare the fit of the MIR:M Total, MR, and MC scores. For the MIR:M Total, the school day time component provided a slightly better fit (-*2LL*= 17034.5) than the Total day time component (-*2LL* = 17043.7). In addition, there was a small increase from the school day residual variance (σ^2= 33.35) to the Total day residual variance (σ^2= 33.40). The MC score followed a similar trend as the Total score with a better fit and smaller residual variance (-*2LL*= 14986.7, σ^2= 15.34) with the school day model compared to the Total day model (-*2LL*= 14994.9, σ^2= *1*5.40). MR, differed from the Total score as the school day model provided poorer fit and larger residual variance (-*2LL*= 15734.0, σ^2= *1*9.91) compared to the Total day model (-*2LL*= 15732.7, σ^2= *1*9.79). Overall, the differences were negligible across models. To maintain consistency across scales, the school day time models were retained.

Predicting TCAP Scores

Table 10 provides descriptive statistics of the TCAP scaled score, TCAP raw scores, and the five TCAP subscales. Table 11 provides the correlation coefficients between these same scores. As is apparent from this table, TCAP raw score and TCAP scaled score show a very strong relation, $r(221) = .96$, $p < .0001$; this was expected since one is a transformation of the other. However, this association was less than expected. In addition, when comparing TCAP

Raw Scores and TCAP scaled score correlations with the five subscales, the TCAP raw score had a stronger correlation with each subscale than the TCAP scaled score. These TCAP raw score and TCAP scaled score discrepancies warranted further analysis.

Analysis of the TCAP scaled scores showed an unequal distribution of scores and extreme outliers (i.e., Z-Scores > $|4|$); on the other hand, the TCAP raw score analysis did not reveal any outliers. Visual analysis of the relationship between TCAP raw score and TCAP scaled score indicated that the relation diverged at the extreme values (i.e., scaled score outliers). Closer examination of this relationship indicates that a one point change in the raw score did not equate to a consistent change in the scaled score. For example, a one point raw score change from 41 to 42 resulted in a two point scaled score change from 747 to 749; on the other hand, a one point raw score change from 63 to 64 resulted in a 58 point scaled score change from 842 to 900. Thus, the outliers of the TCAP scaled scores on the extremes likely resulted from this unequal distribution of scores. Given the clearer distribution of scores and the stronger relationship with the subscales, the TCAP raw scores were used as the dependent variable.

MIR:M Composite Prediction. To address research question three, each of the three composites were compared with the TCAP scores. The predictive models examined the extent to which the composites independently predicted TCAP scores. Overall, these analyses provided comprehensive analysis about the relative predictive power of the parameters within each composite. However, since these analyses modeled each composite separately, they did not provide a relative predictive power comparison of the composites, only the components within each composite.

Since the composite predictive modeling established the predictive potential at the teacher and student level, a two-level unconditional model was fitted to the data to partition the variance between these levels. This model identified the between-teacher variance (τ_π= 58.08) as well as the student-level variance (σ^2= 99.14). Overall, this unconditional model found that 37% of the total variance in the model occurs at the teacher level; the remaining variation occurs at the student level.

MIR: M Growth Model Correlates of TCAP. Zero-order correlation coefficients provided initial information about the relation between the TCAP and the three composites. The average correlation coefficient between the TCAP Total and MIR:M composites across all 12 probes was .41 (range of coefficients .31 to . 54) between the TCAP Total and Total composite, .35 (.21 to .41) between the TCAP Total and MC composite, and .25 (.09 to .39) between the TCAP Total and MR composite. These data do not take into account any of the individual growth parameters; instead this indicates the basic predictive potential of the three composites at any one measurement occasion before modeling of the growth parameters.

To maintain consistency, only the three components (i.e., intercept, linear slope, and quadratic slope) from the individual growth models were used in the prediction of TCAP scores. Table 12 presents the correlations for the TCAP raw scores, TCAP scaled scores, and the subscale scores with the intercept, linear slope, and quadratic slope from the Total, MR, and MC models. The intercept of the Total and MC was the highest correlated component with the all TCAP scores. In particular, the Total intercept was significantly correlated with the TCAP raw scores, $r(221) = .45$, $p < .0001$; the Total intercept provided a stronger relation than the Total linear slope, $r(221) = .29$, $p < .0001$ and Total quadratic slope, $r(221) = -.12$, $p = .07$. Similarly, the MC intercept was significantly correlated with the TCAP raw score, $r(221) = .46$, $p < .0001$; the MC intercept provided a stronger relation than the MC linear slope $r(221) = .16$, $p < .05$ and MC quadratic slope $r(221) = .15$, $p < .05$. The MR intercept was significantly correlated with the

TCAP raw score, *r*(221) = .23, *p* < .001; however, the MR linear slope provided a slightly stronger relation with the TCAP Total, *r*(221) = .26, *p* < .001. The MR quadratic slope was also significant *r*(221) = -.20, *p* < .01.

Additional comparisons of the Total, MC, and MR models were determined by taking the predicted scores of the models on the day of the TCAP administration. In other words, from this analysis, the three components of each model (i.e., intercept, linear slope, and quadratic slope) produced an expected value of the MIR:M scores on the day of the TCAP math administration. Compared to the previous modeling, the components were not disaggregated; instead, these scores revealed the relative predictive power of the entire model for each composite.

Table 13 provides the correlations of predicted values with the TCAP components. All of the predicted values were significantly related to the TCAP raw score; specifically, the Total predicted was most significant, *r*(221) = .53, *p* < . 0001, followed the MC predicted, *r*(221) = .43, *p* < .0001, and the MR predicted, *r*(221) = .35, *p* < .0001. Relationships with the TCAP subscales revealed that the MIR:M Total composite, *r*(221) = .54., *p* < .0001 had its strongest correlation with the Numbers and Operations TCAP subscale. In addition, MIR:M MR composite, *r*(221) = .38, *p* < .0001 and the MIR:M MC composite, *r*(221) = .47, *p* < .0001 had their strongest relation with the Geometry and Measurement TCAP subscale. The Mathematical Processes subscale provided the weakest relation with the MIR:M Total composite, *r*(221) = .46, *p* < .0001, MIR:M MC composite, *r*(221) = .37, *p* < .0001,and MIR:M MR composite, *r*(221) = .32, *p* < .0001. The MIR:M composites' weak relationship with the Mathematical Processes scale may have weakened their correlation with the TCAP raw and composite scores.

MIR:M Total Composite Prediction. The initial MIR:M Total model included the centered intercept, linear trend, and quadratic trend at both levels. Application of this model resulted in a 21% decrease in the between-student variance ($\sigma^2 = 79.77$) and a 14% decrease in the teacher level variance ($\tau_\pi = 43.95$). All individual level variables were significant, but none of the teacher level variables were significant. Given the reduction in the teacher-level variance, Total intercept showed some evidence of contribution to the model ($p = .12$) while the other teacher level variables provided little contribution. Further analyses examined the teacher level Total intercept variable. The student level grand mean centered linear and quadratic trend variables were used in these applications.

Next, only the student level and teacher level Total intercept variable were entered into the model. Entering the Total intercept alone resulted in a significant decrease in variance at both the student level ($\sigma^2 = 88.10$) and the teacher level ($\tau_\pi = 34.68$). Adding the grand mean centered Total Slope variable resulted in an increase in the teacher-level variance ($\tau_\pi = 36.61$) and a 9% decrease in the student level variance ($\sigma^2 = 80.56$) from the previous model. The Total quadratic resulted in a slight increase in the variance at the teacher level ($\tau_\pi = 36.74$) and a 3% decrease in the variance at the student level ($\sigma^2 = 78.13$) from the prior model. Allowing the student-level Total intercept to vary randomly within classrooms resulted in a slight increase in the teacher level variance ($\tau_\pi = 36.92$) but a 7% decrease in the student level variance ($\sigma^2 = 72.64$) from the previous fixed effects only model. Although this variable was not significant ($\pi = .28689$, $p = .11$), the overall fit ($-2LL = 1611.2$) improved from the previous model ($-2LL = 1615.6$) and the improvement was significant, $X^2(1) = 4.4$, $p < .05$. Therefore, the variance reduction and the improved fit indicated that it should be included in the overall model. Compared to the unconditional model, the final model accounted for 36% of the variance at the teacher level, 27% of the variance at the student level, and 30% of the variance across levels (i.e., combining teacher and student level variance reduction).

MIR:M Math Calculation Global Composite Prediction. Modeling all MC components concurrently resulted in a significant reduction (34%) of the teacher level variance (τ_{π} = 38.14) from the unconditional model. In addition, the model also reduced the student-level variance (σ^2 = 92.36) by 7% from the unconditional model. Of all the parameters, only the student-level MC intercept was significant π = 0.79, t(208) = 3.02, p < .001 and provided greatest predictive potential. This model provided a reduction of 17% of the overall variance across levels.

The next analysis was designed to investigate the significant MC components that reduces the variance and improves fit. Because MC intercept values produced the most predictive potential at both levels, they were entered in the model first. This model significantly reduced the student- (σ^2 = 92.44) and teacher-level (τ_{π} = 31.10) variance in the model. This resulted in a reduction of variance of 46% on the teacher level but only 7% at the student level.

Since the linear and quadratic slopes were not significant at either level, the grand-mean centered Slope variables were individually added to the previous model. The linear slope was not significant, π = 1.74 t(209) = 1.30, p = .19. In addition, including the linear slope resulted in a small decrease in variance at the student level (σ^2 = 92.02) and a small increase at the teacher level (τ_{π} = 32.06). The quadratic slope was not significant, π = -25.47 t(209) = -1.01, p = .31. This model also slightly decreased the student-level variance (σ^2 = 92.35) but slightly increased the teacher-level variance (τ_{π} = 31.75). Therefore, the linear slope and quadratic slope did not add substantive predictive potential at either level of the model.

The final MC model determined if the individual MC intercept randomly varying within classroom provided an improvement fixed-effects MC intercept model. The random MC intercept was not significant, π = .86, p = .11; however, it resulted in a 6% reduction in the student level variance (σ^2 = 86.69) but a small increase in teacher-level variance (τ_{π} = 31.36) in the fixed-effects only model. Across levels, the random effect decreased the variance by 4% from the fixed-effects only model. Compared to the fixed-effects model (*-2LL* = *1*659.9), the addition of the random effect (*-2LL* = 1655.3) resulted in significantly better fit, $X^2(1) = 4.6$, $p < .05$. The reduction of variance and improved fit indicated that this was this was the best predictive MC model. Altogether, this model reduced the student level variance by 12%, the teacher-level variance by 46%, and the overall variance across levels by 25% compared to the unconditional model.

MIR:M Math Reasoning Global Composite Prediction. The initial modeling of MR components simultaneously resulted in an increase of the teacher-level variance ($\tau_\pi = 68.02$) though the increase did not reach the level of statistical significance. The increased variance may have been a result of multicollinearity between the teacher level variables identified in the variance inflation testing (i.e., Variance Inflation Factor > 5). This model did reduce the student-level variance ($\sigma^2 = 85.84$) by 13%. The most significant teacher-level variable was the MR intercept. In addition, the individual MR intercept, $\pi = 0.65$, $t(208) = 3.63$, $p < .001$, MR Slope slope, $\pi = 7.91$, $t(208) = 4.37$, $p < .0001$, and MR quadratic slope $\pi = 117.70$, $t(208) = 2.62$, $p < .01$ were all significant. Because of the increase in variance of the teacher-level variable, the model only accounted for 2% of the total variance across models.

Since modeling all teacher level variables negatively affected the fit, each of the three teacher variable were considered separately to assess their impact on the model. Although none of the variables were significant, the teacher MR linear slope ($\tau_\pi = 61.52$) and quadratic slope ($\tau_\pi = 62.83$) both increased the variance; conversely, the teacher MR intercept decreased the variance ($\tau_\pi = 54.64$) by 6%. Thus, the MR intercept was used in subsequent teacher-level analyses for the MR; the MR linear slope and quadratic slope were centered on the grand mean rather than the teacher mean.

Because the MR intercept was a significant student-level variable, and the only substantive teacher-level variable, it was modeled first. In addition to the reduction of variance on the teacher level, the MR intercept reduced the student-level variance (σ^2 = 94.93) by 4%. Introducing the grand mean centered MR Slope (σ^2 = 88.24) decreased the variance by an additional 7%. Adding the grand mean centered MR quadratic (σ^2 = 85.83) decreased the student-level variance by 3%. Altogether this model decreased the variance by 13% relative to the unconditional model at the student level. All variables but the teacher MR intercept were significant (i.e., $p < .001$); however, because the teacher-level MR intercept provided a substantial reduction in variance (τ_π = 54.34), it was retained. Allowing the student-level MR intercept to randomly vary within classrooms did not significantly improve the overall fit (*-2LL* = 1635.1) from the previous model fixed-effects only model (*-2LL = 1*635.6), $X^2(1) = .5$, $p = .48$. This model decreased the student-level variance (σ^2 = 84.42) by 2%; there was a slight increase in the teacher-level variance (τ_π = 54.41). In sum, the fixed-effects only model accounted for 13% of the variance at the student level, 6% of the variance at the teacher level, and 11% across levels compared to the unconditional model. In addition, the mixed-effects model accounted for 15% of the variance at the individual level, 6% of the variance at the teacher level, and 12% across levels compared to the unconditional model.

Summary Comparison of the Composite Predictive Models. Table 14 provides the best predictive model for each composite; these models show the relative predictive potential of TCAP scores at the student level, teacher level, and across the two levels. The Total composite model had the best overall prediction, accounting for 30% of the variation across the two levels;

the MC composite had the second best overall prediction, accounting for 25% of the variation across the two levels; and the MR composite model had the worst overall prediction, accounting for 12% of the variation across the two levels.

The student-level and teacher-level prediction deviated in their overall prediction across the composites. In particular, the student-level prediction indicates that the Total composite provided the best prediction, accounting for 27% of the student-level variation; the MR composite provided the second best prediction, accounted for 16% of the student-level variation, and the MC composite provided the worst prediction, accounting for 13% of the student-variation. The teacher-level prediction indicates that the MC composite had the best prediction, accounting for 46% teacher-level variation; the Total composite had the second best prediction, accounting for 37% of the teacher-level variation; and the MR composite provided for worst teacher-level prediction, accounting for only 6% of the teacher-level variation.

Predictive Modeling of MR and MC Global Scores. To determine the best fitting model taking into account both the MC and MR components, the significant fixed variables from the MR and MC predictive models were entered into the equation (research question three). This analysis included the teacher level MC and MR intercept, the individual MC and MR intercept, and the grand mean centered individual MR linear slope and the quadratic slope. Compared to the unconditional model, this configuration resulted in an 18% reduction in the student level variance (σ^2 = 80.86) and 40% reduction in the teacher level variance (τ_π = 35.09). Only the teacher level MR intercept was not significant, β = -.48, t(9) = -.24, p = .81. Since the individual level MR intercept was significant, the grand mean centered individual intercept was then tested. This slightly decreased the student level variance (σ^2 = 80.85) but substantially decreased the teacher level variance (τ_{π} = 32.38). In particular, this resulted in a 8% decrease in the teacher level variance from the previous model and a 44% decrease from the unconditional model.

The next model allowed the MC intercept to vary within each classroom. Consistent with the previous modeling, this variable was not significant (p = .13) but it reduced the variance by 7% at the student level (σ^2 = 75.55) and slightly increased the variance at the teacher level (τ_{π} = 32.601). Moreover, the overall fit of this model (*-2LL = 1*615.2), compared to the previous model (*-2LL* = 1620.2), indicated significant improvement, $X^2(1) = 5.0$, $p < .05$. The grand mean centered individual MC linear slope and quadratic slope were both entered individually, but neither the linear slope, π = 2.44 $t(206)$ = 1.94, p = .054 nor quadratic slope, π = -33.50 $t(206)$ = -1.41, p = .16 were significant. As these variables do not improve the model, the previous model provided the best overall fit. Overall, this model decreased the student level variance by 24%, the teacher level variance by 44%, and the variance across level by 31%.

Best Predictive Model. To address research question four, the final MIR:M growth modeling determined the best model from all of the MIR:M components. The modeling of the Total MIR:M scores showed the greatest reduction in variance on the individual level. At the teacher level, it appeared that the MC intercept provided the most significant reduction in variance. The components from the final Total model were entered first into the model, with the Total intercept allowed to vary within classes. All variables were entered with the grand mean centered variables since it was likely that the teacher level Total intercept was highly correlated with the teacher-level MC intercept. It was assumed that this configuration would produce results similar to the final model of the MIR:M Total components. As expected this strategy resulted in teacher level-variance (τ_{π} = 36.60) and student-level variance that were consistent with the Total final model (σ^2 = 72.64). Adding the teacher-level MC intercept, resulted in a similar student-level variance (σ^2 = 72.82) but a 7% decrease in the teacher-level variance (τ_π = 34.17). Although it accounted for a substantial reduction in variance, the MC intercept was not significant, β = .99 t(10) = 1.30, p = .22, likely a result of a strong correlation with the Total intercept variable, r(221) = .51, p < .0001.

Since the purpose of this modeling was to identify the best fit at both levels, the same components were modeled with the group mean centered Total components; this strategy was considered necessary to decrease the correlation between the intercept variables. Using the same procedures as the previous model, the student-level variance (σ^2 = 72.64) decreased slightly, but the teacher-level variance (τ_π = 31.98) decreased by an additional 6%. Furthermore, the teacher-level MC intercept was significant β = 2.19 t(10) = 3.07, p < .05. Allowing the Total quadratic slope to vary within classrooms resulted in a very small increase in the teacher-level variance (τ_π = 32.04) but a 2% decrease in the student-level variance (σ^2 = 71.36). Although the fit (*-2LL* = 1608.6) did not significantly improve from the prior model (*-2LL* = 1609.3), X^2(1) = .7, p = .40, it reduced the variance across levels by 1%. Overall, when the Total quadratic slope was free to vary within classrooms, the model accounted for 45% of the variance at the teacher level, 28% at the student level, and 34% across levels from the unconditional model. On the other hand, with its exclusion, the model accounted for 44% of the variance at the teacher level, 27% at the student level, and 34% across levels.

CHAPTER IV

DISCUSSION

This study was designed to investigate the predictive power of probes (MIR:M) developed within an RTI Model when a high-stakes end-of-year test (TCAP)is the criterion. In addition, the data allowed tests of fit of several probe parameters across the academic year (e.g., intercept, slope, variation in administration time). The model testing strategies employed provided insight into the fit characteristics of several probe-score influences (e.g., students, teacher/classroom).

The basic correlational coefficients of the MIR:M composites provide data on the general relationship between the 12 administrations without regard for the impact of time. Not surprisingly, the greater the lag between administrations, the weaker the relationship between the probes. Therefore, probes generally had the strongest relationship with administrations that immediately preceded or followed them. These data are consistent with previously reported MIR:M data for other grades (Hopkins, 2011) and were expected given that students' skill level changes as a function of time, and the longer the time between administrations, the more discrepant these skills become. This pattern was consistent across the three composites. Of note, data revealed additional insight in the probe administration pattern regarding the relationship of the first probe administrations and the following administrations. While the decreasing correlation pattern is still present, the first-administered probe consistently showed the weakest relationship across all lags (i.e., correlation with the closest administration, correlation with next closest administration, etc.). The correlation between the first administration and the second administration was only .45; conversely, the relationship of the other probes with the following administration (e.g., second with third, third with fourth, etc.) revealed an average correlation coefficient of .72. This presents a markedly weaker relationship that is consistent across the three composites.

Apparently there is something qualitatively different about the first administration from the remaining administrations. Although two practice administrations were given prior to the first administration using scripted directions, this was the first opportunity students had to complete the MIR:M and also the first opportunity that teachers had to administer it under actual testing conditions. The novel nature of the measures may have impacted the results in spite of the efforts to make both students and teachers comfortable with the format by requiring the two practice administrations. Students may have used the first administration to identify their individual problem-solving methods within the three-minute time constraint. Teachers may have been less capable of adhering to the standardized procedures during the first attempt. Educators need to ensure that students have a strong understanding of the different items and their task demands prior to take the MIR:M to decrease the novelty of the items; furthermore, comprehensive preparation of administration procedures for teachers will help to ensure proper standardization. Even though teachers were trained to use a script with practice administrations, these precautions may be inadequate.

Modeling of Initial Skills and Growth

The growth models highlight individual differences in student growth for the MIR: M. across an academic school year. Furthermore, these data suggest the growth of students can be modeled as a Total composite score or as MR and MC Global composite scores. Within an RTI model, this gives educators three unique ways to measure students' skills and their growth across the year. That is, the intercept, linear slope, and quadratic slope offer conceptual and applied implications related to the nature of these academic skills and the growth across a school year.

The fixed-level coefficients give robust estimates about the average, or expected value of the three components across students. These coefficients provide a general understanding of the students' entering skills and their growth across the year. The random parameters can explain the variation of scores and provide information about differences within- and between- components, students, and teachers that affect the model.

Intercept. The intercept provides information related to the initial skills of the students. Within an RTI framework, this may be the first quantifiable data about these specific skills. In particular, since the first data are collected with a Universal Screener, these data are typically used within RTI to screen for potential skill deficits. Based on the final quadratic growth model the average Total intercept was 15.22, the average MC intercept was 8.88, and the average MR intercept was 6.37. According to these results about 58% of students' Total initial skills are a result of their MC skills while 42% are a result of their MC skills. In addition, this indicates that the average MC score will be 2.5 points, or 39%, higher than the average MR score.

Although educators can identify initial skill deficits by using each intercept or the combination of the intercepts, they should be aware of the differences between MC and MR composites when making decisions. If using the Total score, they should be aware of the heavier MC weighting. In addition, if interpreting either of the composites, they should know that the MC will be about 2.5 points higher than the MR score. Awareness of these initial differences is paramount to making rational decisions about skill deficits when identifying at-risk students.

Linear Slope. The linear growth skills show the changes of scores across the year. Overall, the linear (within-student) growth accounts for 24% of students’ Total score, 32% of their Math Calculation score, and 19% of their Math Reasoning Score. In addition, linear growth accounts for 29% of the total score, 43% of the MC scores, and 16% of the MR scores between-students. Therefore, the MC linear growth provides more reduction in variance across the levels of the model than the MR linear growth. The linear growth across the year was 10.91 points for the Total score, 9.02 for the MC score, and 1.73 for the MR score. Thus, the linear growth of students is about 5.2 times greater on the MC than the MR. In other words, the MC accounts for about 84% of the linear growth while the MR accounts for 16%. Educators should be aware of more rapid MC skill growth when comparing the global composites; MC skills contribute more growth to the Total composite growth as well.

Quadratic Slope. Although not a primary topic of interest before the analyses, the significant quadratic trend presents meaningful implications about the growth of skills on CBMs. The quadratic trend indicates more rapid growth early in the year with less pronounced growth late in the year; this decrease in growth reflects a plateauing of skills. This is consistent with previous studies that found similar nonlinear growth trends on R-CBMs (Ardoin & Christ, 2008; Christ, Silberglitt, Yeo, & Cormier, 2010; Kamata et al., 2012) and M-CBMs (Graney, Missall, Martinez, & Bergstrom, 2009; Keller-Margulis, Mercer, & Shapiro, 2012) from tri-annual administrations. While results from this study are consistent with previous research, the 12 probe administrations provide additional data not available from three administrations, which was typical of previous research designs. In particular, 12 administrations provide more reliable estimates of the composites' slopes. Moreover, more administrations allow for a better understanding of the changes of the trends throughout the year. For example, data from the tri-annual administrations established less growth from winter to spring than fall to winter showing only that growth is different in the second half of the year. Results from this study more precisely estimated when the changes occurred and the progression of the changes (i.e., whether the changes are rapid or gradual).

Overall, the quadratic slopes account for significant variation in the scores. In particular, this parameter accounts for 4% of the Total score, 4% of MC score, and 4% MR scores. Although the quadratic slope did not significantly affect the Total and MC scores between-students, it accounted for 21% of the between-student variability of the MR. Hence, the quadratic growth accounts for similar variation within-students across the three composites; the between-student differences produced more divergent results across the composites.

The quadratic slope can be compared to the linear slope in other ways as well; while the linear slope increases positively, the quadratic coefficients are negative. Across the school year, this resulted in a decreasing quadratic trend of 4.29 points on the Total composite, 1.84 points on the MC composite, and 2.45 on the MR composite. Thus, 57% the Total quadratic slope is a result of the MR quadratic slope. Although a heavier MR weighting presents a marked difference from the other components, this indicates a downward trend, and is consistent with the overall composite trends.

Overall Growth. The best growth models included both a linear slope and quadratic slope. Although each has unique impact, the general growth of students across the year can be determined by the combination of the two coefficients. The overall growth can be obtained by adding these results together. Therefore, on average, students will experience 6.62 points growth on their Total score, 7.18 points on their MC score, and -.72 points on their MR score across the year. In general, growth on the Total score reflects an increase in MC skills but a decrease in MR skills. Thus, the Total score increases 44% from the initial score, the MC score increases 81% from the initial score, and the MR score decreases 11% from the initial score. While the MC scores were initially higher than the MR scores, MC scores also experienced more growth. This has important implications when interpreting the growth of students.

First, educators need to be aware that the Total score is a combination of a large increase in MC scores growth and a slight decrease in MR scores. Students' overall skill growth may appear slower than expected because of the MR scores limiting this growth. Educators need to carefully consider whether to interpret growth as one large construct or two distinct constructs. Even when interpreting the Total growth, it may be prudent to evaluate the MC and MR growth and identify discrepancies that could be impactful.

Global Composite Differences in Growth

Since the total score is the sum of the MR and MC composite score, the separation of these scores can identify nuances which would not be known otherwise. The significant differences in growth between the MR and MC can have significant impact on the interpretation of the Total score. Initially the MC composite scores account for about 57% of the total score with the remaining 43% a result of the MR score. By the end the year (i.e., 143 days of the study), the MC accounted for about 74% of the score with the MR accounting for 26%.

From an interpretive standpoint, the initial total score is fundamentally different than the scores as the year progresses. While the MC score has a higher weighting on the Total throughout the year, this weighting becomes increasingly significant throughout the year. Thus the Total score is not a consistent representation of skills across the year. Within an RTI framework, these interpretive differences highlight the importance of separating the two composites when evaluating growth. A major component of RTI is to evaluate individual growth of specific skills. Given the differences in the weighting of the Total score, educators who only consider this score may not obtain the clearest picture of skill growth. There may be both natural and artificial causes for these differences; educators should at least be aware of these possibilities and their implications when making decisions.

Since skill development and measurement are nuanced issues, it is difficult to pinpoint a singular cause for the differences between the MR and MC. There are a number of possibilities that may impact these scores. While each possibility may a have unique interpretive implication, the most salient implications depend on whether the differences are natural or artificial. Natural processes would refer to differences in the development of skills or conceptual differences that impact the nature of the measurement; artificial processes would refer to differences that are unrelated to the skills and their measurement (i.e., differences in student responding). Identifying and understanding these possibilities may be useful in educational decision-making within an RTI framework.

Natural Processes. The simplest explanation for the score differences is that MC skills improve more rapidly than MR scores. This growth could stem from some unique differences such as the curriculum or math skill development. Regardless of the precise reason, the point is that differences are organic and related to the constructs themselves. The results of this study provide evidence that the differences are, at least in part, consistent with this possibility.

The impact of natural growth differences appears most evident in the comparison of the variance reduction between the fixed and random parameters. For example, by modeling a random intercept but fixed slopes, the within-student variation of the MC Scores decreased by 23% and resulted in no error reduction between-students from the unconditional means model. The MC within-student error reduction indicates that it is an important component; however, the inability to differentiate between students indicates this growth is consistent. This fixed slope resulted in no reduction of error for the MR. When the random slopes were added, the MR within-student variation decreased. This addition resulted in a significant decrease in both composites' between-student variation. The random slopes can differentiate between students for both composites.

Within an RTI framework, between-student comparisons are necessary to identify students with relative deficits. These comparisons are possible after establishing baseline data through the individual growth trends and the decrease in the within-student variation. Thus, the baseline trend is relatively nonexistent for the MR but increases for the MC. The consistency of scores across the sample provides tentative support that these reflect natural trends.

The consistency of the changes between the composites when comparing the final model to these natural baselines provides additional support. In particular, Model F in Table 8 represents the final MC model and Model F in Table 9 represents the final MR model. These can be compared to the baseline models for each: the fixed growth trend for the MC (i.e., Table 8, Model D) and the unconditional means model for the MR (i.e., Table 9, Model A). The final MC model resulted in a 42% reduction in the between-student variance, 16% reduction in the within-student variance, 31% reduction of the total variance from the baseline model. The final MR model resulted in a 33% reduction in the between-student variance, a 22% reduction in the within-student variance, 28% reduction of the total variance from the baseline model. This consistency indicates that these are baselines that represent the natural changes in development between the two composites.

Another natural explanation for growth differences between the two composites is the conceptual difference between Math Calculation and Math Reasoning. Math Calculation can be conceptualized as a measurement of mathematical skills while Math Reasoning may be more consistent with a measurement of cognitive abilities that are relatively stable. The Cattell-Horn-Carroll (CHC; Flanagan & McGrew, 1997; Flanagan, Ortiz, & Alfonso, 2006; McGrew &

Woodcock, 2001; McGrew, 2005) provides an empirically-supported theory of cognitive abilities and posits that there are three hierarchical stratums. Stratum III is represented by one general ability (G); G is comprised of number of broad abilities (stratum II); the broad abilities are comprised of narrow abilities (stratum III). The broad and narrow abilities are most relevant to the conceptualization of the MIR:M.

One broad cognitive ability conceptualized within the CHC model is Quantitative Knowledge (Gq). Flanagan et al. explained that Gq is the acquisition and storage of quantitative information used to manipulate numeric symbols and is most often measured by achievement tests. From a conceptual standpoint, both MR and MC fall subsumed within the Gq broad ability. Gq consists of Quantitative Reasoning (RQ), Mathematical Knowledge (MK), and Mathematical Achievement (MA), all narrow Stratum I abilities. MK and MA rely strongly on formal instruction in mathematics; thus, MC is highly dependent on MK and MA. On the other hand MR is probably more influenced by RQ. According to Flanagan et al., RQ relies considerably on Fluid Reasoning (Gf), but not as much on formal instruction and classroom experiences. Furthermore, RQ requires some basic mathematical knowledge and understanding of concepts, but is more related to inductive and deductive reasoning. According to Flanagan et al., an example of an RQ task is a number series, a specific skill measured within the MR composite. On the other hand, MC items require straight-forward calculations.

Taub, Keith, Floyd, and Reynolds (2008) performed a CFA and found that Gf had a significant factor loading (.58) on Gq; although this was lower than a math calculation subtest loading (.70). In addition, an RQ (i.e., numerical reasoning) subtest had a .51 factor loading on Gf. Multiplying the RQ factor loading on Gf with Gf factor loading on Gq can provide the indirect factor loading of RQ on Gq (Bodin, Pardini, Burns, & Stevens, 2009; Keith, Fine, Taub, Reynolds, & Kranzler, 2006); this results in a .30 factor loading for RQ on Gq. The heavier weight of math calculation compared to the RQ is consistent with the MIR:M scores which showed consistently higher MC than MR weighting on the Total score.

The nature of the MIR:M growth data offers support for the hypothesis that MR requires both Gq and Gf and is consistent with the assumption that MR requires rules of logic to solve each problem. In particular, the lack of fast-paced growth across students is consistent with the more stable MR ability attribute. The fact that the MR composite changes as much as it does is consistent with the notion that it reflects both ability *and* skill-based components, but perhaps more reflective of the slower-growing *ability* component.

Evaluating the classroom nesting showed additional conceptual differences between the MC and MR. Based on the basic unconditional means models, 42% of between student MC scores can be attributed to the nesting within their classrooms while 4% of the MR scores can be attributed to this nesting. The differences between individual students can impact variables within their respective classrooms, but these classroom variables have minimal impact on MR scores. About 60% of the differences between students' MC scores are accounted for by this nesting. Thus, within-classroom differences become more pronounced when accounting for growth. Since within-student differences were already modeled, these class wide variables occur outside the student but are shared by students within a given classroom; this pattern of influence would most likely impact changeable skills (MC) but not more invariant abilities (MR).

Artificial Processes. Changes in the items may also explain differences between the MR and MC composites. Coles et al. (2013) found that third-grade students had distinct responding styles to the four item-types on the MIR:M. The differences in fit indicated that some students appeared to respond disproportionately to certain item types. This pattern appeared to increase the scores on the items they choose while decreasing the score on remaining items. It is likely that some students responded in a similar fashion for this study as well. Since students initially had higher scores on the MC, students may be more inclined to try items that they believe they are more likely to answer correctly. The time constraints may exacerbate this problem as students have limited time to attempt other items. This provides a “rich-get-richer” effect to skill acquisition (i.e., Matthew Effect; Merton, 1968) as MC scores continue to increase substantially while MR scores change much more gradually over the year.

Interpretation Issues. Given the reasonableness of each of the above explanations, it is unlikely that a singular cause accounts for the differences of scores. Thus, without identifying a singular cause, it may be best to interpret these differences as the combination of the explanations. To be exact, the MC represents skills that show more rapid improvement while MR scores are more resistant to quick change; in addition, some students are more likely to spend their time responding the MC items than the MR items.

Teachers may find it useful to identify students with differential scoring patterns as they are likely to impact their growth parameters. An item analysis of all students may be too time-consuming; however, MR and MC trends could offer enough student detail without taking up significant time. This may be achieved in a few steps.

Using RTI data, teachers could first compare the trend of each composite to the average trend across students and identify students with discrepant trends. This would account for the natural differences in MC and MR scores and can be used to compare the two trends to find any within-student discrepancies. For example, two students could be identified for a below average MR trend, with one student having a below average MC trend and the other having an above average MC trend. Since the student with the below average MC trend consistently shows less growth in both constructs, inconsistent responding would be less likely; the other student's normative difference from peers and ipsative difference between composites, may indicate a greater chance of inconsistent responding. An item-level analysis could be used to verify these results.

When students have consistent responding throughout the year, educators have the flexibility to interpret all composites. They can identify students with specific deficits in their reasoning, calculation, or overall math skills. When students show inconsistent responding, the interpretation is much less clear. In fact, none of the composites may represent accurate skill development. The Total growth would likely be the most reliable because given the time constraints, the discrepancies between composites should counter one another to an extent. Basically, the artificial increase of one score should result in an artificial decrease of the other; this trend would be captured in the Total score. Overall, it is important to carefully evaluate the composites before making educational decisions, and the results provide some support for the use of assessing both.

Variables Impacting Common Applied Modeling Procedures

Although a primary purpose of modeling was to identify the sources of variation, some of the variables that affected the modeling may not be accounted for within a school system. It is prudent to compare the common modeling procedures used in schools with the more complex models of this study to understand the practical limitations of the applied use of CBM data.

In general, practical use of CBMs focuses on individual variables; there is often no consideration for teacher or probe effects. From the conditional means and growth model, these components account for additional error across all 12 probes. While these variables did not account for as much error as the individual components, they are significant enough to have an impact on the decision-making.

Probe Variability. CBM probes are designed to have alternate, but equivalent forms. A number of studies (e.g., Ardoin & Christ, 2004; Poncy, Skinner, & Axtell, 2005) have shown that the significant variation can be attributed to differences (e.g., difficulty level, item content differences) between the probes; these differences were also present in the MIR:M probes for third grade (Hopkins, 2011). In addition, these differences were identified from probes within proximity of time (e.g., adjacent probes) when students' skills should be consistent; previous studies did not establish the impact of probe differences as they relate to student growth. Thus, measurement error from probe variability may have a significant impact and should be investigated. In this study, differences in probes accounted for variation within and between students. Although this error is not substantial (i.e., less than 10%), it supports previous research on measurement error imbedded within CBM measures, even when using 12 administrations. For example, Miller (2011) found significant differences between slope estimates when decreasing the number of measurements (i.e., half the administrations) or comparing measurements (i.e., every even vs. every odd administration). It is likely that this error would increase if fewer administrations were used and decrease the confidence in the estimations of slope. Educators need to be aware that fewer administrations decrease the reliability of these estimates (Miller, 2011).

Teacher Differences. The nature of academic skill development within a classroom is likely to impact students' scores. Since teachers have been found to impact student math achievement by .17 standard deviations per year (Hanushek & Rivkin, 2012), it was important to model the impact of the teacher. Between-classroom differences are likely in the growth of any skill; however, the group format of the MIR:M may have additional impact that is not present in individualized testing. In particular, since the teacher is also a test proctor, any administrative differences (e.g., clarity of directions) and time-specific contextual differences within the classroom (e.g., class disruption, time of day) may impact student scores (Forbes, 2013; Volpe, McConaughy, & Hintze, 2009). Thus, modeling this nesting of students within their classrooms provides information on the impact of skill development and administration variables within the classroom and teacher.

Accounting for the random intercept and slope of the teacher did not result in the reduction of error within students; however, the between-student variation on the Total and MC composites was affected to some degree by this additional level. The student growth on these scales is in part a function of their classroom setting. Within this modeling, the teacher has two important roles: the teacher and the probe proctor. Students' initial skills and growth would be affected by the teacher role as students learn the specific skills and strategies that could impact their skills and growth of skills. Therefore, if comparing students from different classrooms, it may be important to consider the impact of their respective teachers on their scores.

The addition of the probe variable to the nested Total model indicated that about 16% of the between-teacher variation on the total score can be attributed to probe differences. In addition, the marginal impact between students indicates that this is something embedded within the classroom. As such, this may account for teacher's role as a test proctor. While the procedures are standardized, it is difficult to account for environmental variables that exist outside the standardization procedures. A loud noise interrupting students for a few seconds, administration at different times of the day, or what was learned just before administration may result in a small but significant impact on student scores. In addition, although the MIR:M procedures attempt to maintain standardized administrations, there may be subtle differences between administrations. For example, changing a few words (i.e., do your best vs. do your quickest) in the administration of CBMs were found to impact scores (Forbes, 2013). From a practical perspective, ensuring administration fidelity by strictly adhering to the scripted procedures is important. In addition, noting any environmental conditions that may impact students' scores could also give useful interpretive information when making educational decisions.

Predictive Validity

The predictive modeling of the MIR:M involved unique analyses across levels (i.e., student, teacher), MIR:M composites, and individual components (i.e., intercept, linear slope, quadratic slope). In addition, these analyses address a variety of methodological shortcomings of the previous literature. To address the results and their implications, the organization of the discussion will focus on three broad issues. First, the focus will be on the prediction at the student level, teacher level, and across levels to understand the major implications of the results. Next, the discussion will focus on the prediction of the three composites, and the extent to which each composite's components predicts TCAP scores at the various levels. Finally, the discussion will focus on the predictive potential of the three components.

Prediction by Level. When evaluating the prediction of the MIR:M across students, the best predictive MIR:M model (i.e., correlation analysis) accounted for a little over a quarter of the variation in the TCAP scores; however, these basic predictive models fail to account for variation between and within classrooms. By accounting for the hierarchical nature of the data, the prediction accounted for a third of the variation. Overall, this modeling indicated that, across the two levels, over a third of variation in TCAP scores across the two levels can be accounted for by MIR:M scores. This provides substantive data on these basic skills, irrespective of the high-stakes assessment; however, the unique predictive potential at the different levels has both applied and conceptual implications of CBM and high-stakes testing.

Teacher Level. A major goal of high-stakes testing is to evaluate the contribution of teacher and school differences on the students' scores. Results of this study show that over a third of the variation in the TCAP scores can be attributed to the mean differences between teachers. Data were obtained from 12 teachers across eight schools, and six of the schools had data from one teacher. Some differences are most likely a result of school impact as well as teacher influences. Because data came from one district, certain variables (e.g., curriculum, general procedures) should be consistent across this district; however, school-level differences (e.g., principal leadership, interaction of faculty) may affect the results. Students within a school are expected to be more homogeneous (e.g., SES, community characteristics) than students across schools. These variables may be more attributable to the school rather than a particular teacher. Given the inability of the design employed in this study to account for these variables, the between-teacher modeling may be a combination of teacher and school level differences. Accounting for school differences may decrease the overall error further.

The mean MIR:M predicted scores by classroom accounts for about half of the variation of TCAP scores between teachers. This means that teachers who on average, have higher performing students on the MIR:M will also have higher performing students on the TCAP. Thus, the predictive potential of the MIR:M extends beyond the student level. Teacher effects provide some ability to predict the MIR:M scores of individual students, as well as TCAP scores. In fact, the prediction was more accurate at the teacher level than the student level.

These results address some of the major methodological weaknesses of teacher modeling in the literature exploring high-stakes assessment. One of the primary shortcomings of Value Added Models (VAMs) that measure teacher effectiveness within high-stakes testing has been the inability to identify *what* makes an effective teacher (Aaronson et al., 2007); however, in this research the teacher variables (e.g., experience, educational degrees) are conceptualized as the *what,* rather than teaching itself. The MIR:M's curriculum-based measurement of *basic* skills may better capture data related to teaching skills and behaviors. Essentially, the repeated measurement of basic skills provides consistent data on the day-to-day processes that contribute to skill acquisition within the classroom. Thus the conceptualization of *what* makes an effective teacher can be more directly linked to the act of teaching, which may also make the effects of those other teacher variables clearer.

The two-level analysis used in this study also highlights the methodological shortcoming in the earlier research on the CBM prediction of high-stakes testing. Previous research fails to identify the predictive potential of CBMs beyond the student level. Some studies (Graney et al., 2009; Keller-Margulis et al., 2012) have evaluated CBM prediction and included the teacher level variables in the prediction of high-stakes assessment; however, the teacher-level variables were from prior high-stakes assessment scores, not CBMs. While this study highlighted the importance of teacher-level variables, with the use of CBMs, the teacher-level prediction with CBMs is still lacking though it is addressed to a limited degree in this study.

The overall lack of substantive research is interesting from both a conceptual and applied standpoint. For example, CBMs are designed to measure global outcomes of skills, which offer the ability to evaluate instructional methods (Fuchs & Deno, 1991). The basic conceptual link is clear between CBMs and high-stakes measurement—both provide a crude evaluation of instruction, which can be considered a teacher-level variable. This study highlights the strength of this relationship and provides important practical implications.

From a practical standpoint, predicting the aggregate score of students within a class may be especially useful to educators. Since teachers are evaluated based on the performance of students as a group, the reduction of variance of the mean teacher score demonstrates the usefulness of these data. Thus, a teacher can evaluate the performance of students within his or her class on the MIR:M to determine the possible range of mean scores within that class. Without this level of modeling, the likely value of mean scores (i.e., 95% CI) would center around the entire mean of scores and could vary by 15 points on either side of the mean. In other words, this interval extends across 21% of the Below Basic achievement scores, 100% of the Basic achievement scores, and 80% of the Proficient achievement scores. The addition of teacher level modeling has the potential to decrease the interval of the mean scores down to 11 points on either side. In other words, this interval extends across 8% of the Below Basic achievement scores, 100% of the Basic achievement scores, and 46% of the Proficient achievement scores. Altogether, the mean score range for each teacher would decrease from 46% of the total scores to 34%.

Student Level. By establishing the between-teacher differences, it was possible to identify the between-student differences. Because whenever possible, students' scores were centered on their respective classroom mean, the resulting models show the variation of scores between students but within the classroom. This decreases the correlation between the teacher variables and individual variables, allowing for a clearer understanding of the individual effects without impacting the between-teacher effects. As a result, the addition of the individual score decreases the between-student variation by over 28%. Although the MIR:M provides significant predictive potential, the student-level predictive potential was limited compared to the teacher-level modeling; differences in measurement procedures and task demands of the MIR:M and TCAP may account for the limits at the student level.

Although the MIR:M and TCAP measure similar skills, they have vastly different measurement procedures. To maintain efficiency, the MIR:M data collection for each probe was 3 minutes. Across all 12 MIR:M administrations, students were tested for a total of 36 minutes; students were allotted 75 minutes for the math part of the TCAP. The efficiency of the data collection of the MIR:M may decrease the amount of information that can be obtained from a single student. Furthermore, the three minute administration of the MIR:M results in data on the calculation and reasoning fluency of students. Students have 75 minutes to answer 64 questions on the math TCAP. Although scores reflecting strong math fluency skills would be useful on the TCAP, other math skills (i.e., overall Math Calculation and Reasoning skills) are useful as well.

The items on the MIR:M and TCAP place different task demands on the students. Appendix A provides a sample of the MIR:M test while Appendix B provides samples from the TCAP test. One salient difference in these task demands is the amount of reading required. Whereas MIR:M items required no reading, most of the TCAP items require students to read at least one sentence. Some items have as many as four or five sentences in the question and a sentence for each answer. Therefore, students' reading skills may have a significant impact on the predictive potential of the MIR:M.

A recent study provides preliminary evidence that reading skills impact TCAP math scores. Taylor et al. (2014) used the MIR:M and the MIR:R to identify third-grade students with strong math skills but significantly weaker reading skills. These students' TCAP scores were compared to their peers with strong math skills but without a reading weakness. They found that students with reading weakness scored significantly lower on the TCAP math than their peers. In addition, across the entire sample, MIR:R provided a slightly stronger relationship with the TCAP math than the MIR:M. Thus, students' reading skills provided a significant predictor of the TCAP math. At the very least, reading skills can moderate the relationship between the MIR:M and the TCAP.

Although the MIR:R was not used in this study to determine the impact of reading, the relationship between the MIR:M and the specific TCAP scales may provide evidence for the impact of reading. In particular, of all the TCAP scales, the TCAP Mathematical Process scale produced the weakest relationship with the three MIR:M predicted scores. A review of six item samples (Tennessee Department of Education) showed that four of the items had seven or more sentences, one problem had three sentences, and one problem had a single sentence. Thus, reading skills appear to have an important role in solving these problems. The amount of reading required to solve these problems, combined with the limited predictive potential across all three MIR:M scales, provides preliminary evidence that the Mathematical Processes scale may be moderated by students' reading skills, decreasing the overall prediction for the MIR:M as a whole, given that reading is a component of the criterion variable (TCAP).

The MIR:M was designed to efficiently measure basic math skills across the academic year; this required brief measurements and isolation of specific math skills by eliminating reading demands. The TCAP was designed as a broad measurement of math skills and required reading demands. Since both instruments target robust math skills, the MIR:M significantly predicted the TCAP at the student level; however, this prediction may be limited due to the unique design and measurement of each measure.

Prediction by Skill. By identifying the best fitting predictive model of the Total, MR, and MC components, it was possible to identify the predictive potential differences for each composite between teachers, between students, and across levels.

MIR:M Total Composite Prediction: The most powerful predictive model was created using the Total MIR:M score; it accounted for the most variance between students. In particular, within a classroom 27% of the variance in TCAP scores can be attributed to Total MIR:M. Thus, if teachers intend to predict their students' individual TCAP scores, the Total MIR:M would be most useful. In addition, at the individual level, all components improved the model. Therefore, the joint total of students' initial skills and their skill growth contributes to this prediction.

Between teachers, the Total MIR:M scores accounted for the second most variation. Unlike the student-level components, only the Total intercept provides substantial predictive potential at the teacher level. Thus, the class wide average of initial skills can predict the teacher's average TCAP score. The best fitting Total model accounted for the most variance across the two levels. This indicates multidimensional measurement of the Total score has significant predictive potential at both the student and teacher level, a finding of use to educators who are interested in determining the most accurate prediction.

MIR:M Math Reasoning Prediction. The Math Reasoning components provided the second best fitting model at the student level. All student components significantly contributed to the prediction. Similar to the Total model, students' entering Math Reasoning skills score and their skill growth across the year are significant predictors. Since MR scores, on average, remained consistent across the year, it is noteworthy that the growth parameters were significant predictors. Regardless of the specific trajectory of growth, modeling these parameters can differentiate scores on the TCAP. While CBM scores are intended to capture subtle changes in skill development, these data show that these changes are not required to differentiate between students.

Math Reasoning provided the worst fitting model at the teacher level. As with the Total prediction, the MR intercept is the only teacher level variable that accounts for any of the between-teacher variance, although this estimate is non-significant. Overall, this pattern is consistent with the growth modeling of MR. That is, MR differences are primarily evident at the student level; it provides very little information to differentiate between teachers. This would appear to support the notion that the MR scale is more reflective of stable abilities; on the other hand, the prediction by the growth parameters shows individual differences in skill development. Because of the small reduction in variance at the teacher level, the MR provides the worst prediction across levels. Therefore, based on these data it is not necessary for educators to model teacher MR parameter; modeling of the student parameters would suffice.

MIR:M Math Calculation Prediction. The Math Calculation components provide the least predictive potential at the student-level. In particular, only the MC intercept provides any substantive prediction; the linear and quadratic slopes do not contribute beyond the MC intercept prediction. Only students' initial calculation skills are predictive. Although, the MC experienced the most growth throughout the year, this growth does little to differentiate between students on the TCAP.

While the student-level MC components are the least predictive, these components provide the best prediction at the teacher level. This is consistent with the growth modeling that identified significant differences at the teacher level on the MC. The MC intercept is the only teacher-level variable that provides any substantive prediction, although it is the single most predictive teacher-level component across models. This indicates that the class wide average of the entering MC skills can best differentiate between teachers. Because of this teacher-level variance reduction, the MC components provide the second best prediction across the levels. In addition, this provides the best prediction between the two global composites. Educators may find it most useful to model the MC teacher parameters

Practical Application of Skill Predictions. Since the estimation of the MIR:M growth was revealed most by modeling the Total composite and the Two Global composites, educators have the flexibility of using these data for predictive purposes. If a quick predictive estimate of TCAP scores is needed, then the Total parameters would be most beneficial. That is, within an RTI framework, educators can have confidence that only using the Total parameters will still provide a robust prediction.

In situations where both the MR and MC are modeled, educators do not need to use all parameters. The best fitting model included both the teacher and student MC intercept, and the grand mean centered MR intercept, linear slope, and quadratic slope. From the separate predictive modeling of the composites, the MC provided the most predictive potential at the teacher level while the MR provided the most predictive potential at the student level. The differences in the composites' predictive potential by level provide further evidence of two distinct constructs.

The growth modeling indicated modeling all three scales could give educators more comprehensive information and flexible use of data. In particular, the most predictive model included the all three student level Total components and the teacher-level MC intercept. Therefore, by identifying the class average of students' initial MC skills and all within-class Total components, educators can make a more accurate prediction. The distinct nature of the composites and their components provides useful data for growth modeling and predictive modeling. Gathering all available MIR:M data is most useful.

Prediction by Components. Both similarities and differences are evident in the predictive potential of the MIR:M Total, MR, and MC components across the intercept, linear slope, and quadratic slope.

Intercept. Across the three composites, the intercept is consistently the most predictive component at both the student and teacher level. These findings are consistent with earlier research using within-year growth on CBMs to predict high-stakes achievement (Baker et al., 2008; Stage & Jacobson, 2001; Yeo et al., 2011). Although the results are similar, some important distinctions arise between the current study and previous research.

First, the majority of research that has identified entering skills as the strongest predictor has been based on R-CBMs. Although one study, (Keller-Margulis et al., 2008), examined the relationship of entering skills of M-CBMs and high-stakes assessment, it did not show the unique contribution of these skills when accounting for growth. The current study provides preliminary evidence of the unique predictive potential of the entering skills on M-CBMs on high-stakes assessment while accounting for student growth.

Second, earlier research primarily used growth based on three measurement occasions. The limited number of measurements likely impacts the estimation of growth as a result of measurement error; this makes it difficult to determine if the intercept is truly the strongest predictor or rather a result of poor slope estimation. By decreasing variability using 12 measurements, results from this study can more accurately determine initial skills' predictive potential.

Last, as noted earlier, this study provides novel evidence of the predictive potential of CBM data at the teacher level. Surprisingly, this predictive potential is exclusive to the intercept. This indicates that 45% of teachers' class wide TCAP scores can be explained by the entering Math Calculation skills of their students. Teachers can review their students' MIR:M scores from early in the year to gain a sense of the average end-of-the-year TCAP scores.

The relative predictive potential of initial skills appears to directly challenge the basic ideals of the current educational environment which stresses student growth. The intercept is a reflection of skills gained prior to the initial administration and likely measured in previous high-stakes assessments. Nonetheless, since the first probe was not administered until the 36th school day, the intercept also includes skills gained while in these teachers' classrooms. Thus, these entering skills represent skill acquisition before the school year of this study in addition to the 36 days with the teacher. Because the combination of tendering skills and 36 days of instruction represents substantially more skill development than the 140 days of this study, it should not be surprising that this leads to a more accurate prediction.

Linear Slope. The linear slope also provides significant predictive potential at the student level. Research has been mixed as to the predictive potential of the slope; some studies have found it to be predictive (Chard et al., 2008; Miller, 2012; Keller-Margulis et al., 2008) while others have not (Yeo et al., 2011). This study controlled some of the methodological issues that were present in prior research. Most obvious (Chard et al.; Keller-Margulis et al.; Yeo et al.) used only three probe administrations, which may make it difficult to estimate the slope because of measurement error. In addition, the predictive potential using three probes appears more evident when the data are collected across years (Chard et al.) rather than within a year (Yeo et al.). Perhaps estimation of the slope across years is more robust to this measurement error than within a single year. By using 12 administrations, measurement error is less of a concern, and the within-year slope can be estimated with more confidence; this is consistent with Miller's (2012) finding that the reading version of the MIR provides a more accurate prediction of high-stakes assessment using all administrations throughout the year.

A prevailing issue across previous studies is that most research examined R-CBM growth; one study (Keller-Margulis et al., 2008) evaluated M-CBM growth. Their study found that high-stakes tests had strong correlations with both a math computation measure and math reasoning linear slopes; however, this predictive potential above the intercept was not established nor was there a comparison of the predictive potential between computation and reasoning. In addition, they used only three probe administrations and the math computation and math reasoning probes were separate measures altogether. This current study builds upon those results and provides information about the predictive potential above the intercept. Moreover, comparisons of the predictive potential between math reasoning and math calculation were performed.

Consistent with the Miller (2012) study of the MIR:R, the multiple components of the MIR:M, may explain the predictive potential of the slope compared to some other studies (e.g., Yeo et al., 2011) that used instruments with only one component. For example, the MR linear Slope was found to be predictive while the MC linear Slope was not. If the MIR:M only measured MC, then there may not have been predictive potential in the slope. This provides support that multidimensional CBM measurement may provide the most predictive potential, and this measurement can be obtained in one, not two separate instruments (e.g., Keller-Margulis et al., 2008).

The differences in the MC and MR Slope predictions are especially interesting when considering the differences in their overall growth parameters. In other words, although on average, students did not gain consistently throughout the year on the MR, it was more predictive than the MC growth, which was sensitive to changes. As CBMs are designed to be sensitive to change (Deno, 1985) this has implications on both the modeling of CBM growth and the prediction components of this growth.

The lack of MR growth indicated that it was not very sensitive to change; consequently, these results show that a measure's sensitivity to change may not be as important as previously thought. Instead, subtle changes in the MR are more predictive than larger changes in the MC. As MR may be more robust to change, when change does occur, the implications are clear. In addition, since on average, students already begin with higher MC skills, the growth does not have as much impact; rather, the intercept of the MC provides the predictive potential of the MC skills.

Although MC skills experience greater change, the growth modeling indicated that the individual differences in growth on the MC did not account for the same amount of variation as the individual differences in growth on the MR. In other words, students tend to experience more growth on the MC as a whole, but individual differences account for more variation within students on the MR. Therefore, the amount of change is not necessarily the most important predictive coefficient; rather, the coefficient that can capture the most variability at the student level is the most essential, regardless of whether or not the average growth is significant.

Quadratic Slope. Despite the number of studies that have identified quadratic slopes in CBM growth (Ardoin & Christ, 2008; Christ et al., 2010; Graney et al., 2009; Keller-Margulis et al., 2012), the predictive potential of quadratic slopes have not been established. As a result, this study provides preliminary evidence in the predictive potential of the quadratic slope. In particular, this study found that the quadratic slope was predictive above the intercept and linear slope for the Total composite and the MR composite. Similar to its linear Slope, the MC quadratic slope was not predictive.

These results are consistent with studies supporting the value of growth modeling as the MR quadratic model has the most extensive reduction in variance when adding its quadratic slope; the MC quadratic slope has the least. Therefore, the more variation a quadratic slope captured for a particular skill, the greater the predictive potential. Although a more significant quadratic trend is not the preferred trajectory since it indicates a more dramatic leveling and decreasing of student scores, the predictive potential of that data is valuable nonetheless. Thus, focusing on how well the trajectory can differentiate between students, both on MIR:M growth modeling and the predictive modeling, may be more useful than what an overall trajectory represents.

Summary

The above discussion provides perspectives on the overall measurement, growth, and predictive potential of the MIR:M. The growth modeling of the MIR:M provided evidence of two unique global constructs in Math Calculation (MC) and Math Reasoning (MR). This was most apparent in the overall change of scores throughout the year. More specifically, on average, MC scores increased significantly throughout the year while MR scores remained consistent. While the growth trajectories of the global composites were unique, both trajectories provided differentiation of scores between-students. In other words, educators could interpret each trajectory and make comparisons and decisions about the status of student skill development throughout the year; this is important within an RTI framework.

The modeling patterns provided data to show that modeling one Total composite, instead of two global composites, could offer valuable evidence about overall skill development and significantly differentiate between students. Educators should be aware of the differences in the overall growth trajectories of the two global constructs that comprise the Total score. More specifically, as the year progresses the total score becomes more heavily weighted toward the MC. Therefore, even when making decisions based on the growth trajectory of Total score, it may be useful to evaluate the trajectories of the global constructs to obtain more comprehensive information about skill growth.

The growth modeling showed that all three scales were better modeled with a quadratic slope in addition to the linear slope; this is consistent with earlier research on within-year CBM growth. This indicates that as the year progresses, the trajectories show less growth. Therefore, students experience a greater gain in scores initially, but these scores become less pronounced towards the end of the year. This is important to consider when using this growth trajectories to make educational decisions. In particular, educators should be aware that a linear trend may not provide the most accurate representation of growth as student scores begin to level. As such, they need to consider that scores falling below the trend line may be expected and not necessarily indicative of a lack of growth.

The predictive modeling provided validation for the constructs themselves and also provided evidence that the growth parameters (i.e., intercept, linear slope, quadratic slope) have significant predictive potential. In addition, the predictive potential is evident at both the student and teacher level. This means that educators can use the MIR:M scores to predict individual student scores on the TCAP, but they can also use the class wide MIR:M scores to predict the average TCAP score for the class. This provides a flexible use of the scores to meet to their individual predictive needs.

Overall, there was remarkable consistency between the growth modeling and the predictive modeling. For example, the intercept accounted for the most variation in scores across scales in both the growth modeling and the predictive modeling. In addition, the growth modeling indicated that student's MC and Total scores were impacted by the nesting within their respective classroom; conversely, the MR growth modeling was not affected by this nesting. The predictive modeling provides similar results as the Total and MC scores were predictive at the teacher level while the MR was not. Taken altogether, this consistency provides validation of the constructs and highlights the flexible utility of the MIR:M data.

Recommendations

These results provide information that educators can use to determine the best administration procedures of the MIR:M and the best modeling procedures for growth and TCAP prediction. Across the three composites, the initial MIR:M administration produced the weakest correlation with the other administrations. This is believed to be partially related to the novel nature of the first administration under testing conditions. Although practice administrations were provided, perhaps they did not provide enough exposure nor adequately simulate the testing administration format. Therefore, it is recommended that an additional practice administration occur, and that examiners ensure adequate fidelity during these practice sessions to the actual administration conditions and procedures.

In addition, the initial administration was not given until the 36th day of school. One-fifth of the school year had passed before the MIR:M data could be used to identify skills deficits and inform instruction. Therefore, it is recommended that the initial MIR:M administration be given much earlier in the school year to provide better information about entering skill level.

Recent research (e.g., Christ, Zopluoglu, Long, & Monaghen, 2012; Thornblad & Christ, 2014) indicates that six weeks of progress monitoring may not be enough time to reliably estimate CBM growth. Because of the delay in initial data collection in this study, nearly 40% of the school year passed before a reliable estimate of MIR:M growth could be obtained to inform any instructional changes; evaluating the impact of these changes would thus require even more time. Obviously, MIR:M administration should begin earlier in the year to provide time to obtain reliable growth estimates; nonetheless, the current format does provide evidence in support of the two-week administration format. This schedule appears ideal for obtaining enough administrations to reduce error while providing enough time to model change.

The predictive modeling data provide further evidence in support of the argument for beginning initial screening earlier in the school year. Specifically, the intercept was the most predictive component across composites and levels. Therefore, the intercept represented the combination of entering skills and the first 36 days of skill development. The earlier the initial screening, the more representative the intercept is of the entering skills of students, which may also provide a more accurate representation of change.

Given the differences between the three MIR:M composites in the growth modeling as well as the predictive modeling, it is recommended that all three composites be modeled. For the growth modeling, this would highlight any changes in two distinct global math composites and the overall composite. Because the MC and MR composites have distinct growth trajectories, comparison of the skills would provide information characterizing the specific changes in student skills and ensure more accurate interpretation of the Total composite changes. As noted in the

TCAP prediction, all three Total composite student level components and the MC intercept for the teacher-level components provided the most accurate prediction. Therefore, modeling all composites to obtain these components for prediction appears optimal.

Limitations

Population and Sampling. There are several limitations that exist within the current study. The relatively homogenous sample limits the generalizability of the results; although treating the various components of the study (i.e., teachers and students) as random variables within HLM should mitigate this issue. The cause and impact of missing data is another limitation of this study. Although individuals with significant missing data (i.e., > 3 missing probes) were deleted from analyses, the cause of missing data is not known. In addition, while HLM is adept at handling missing data, assumptions about the missing data must be met; whether these assumptions were met or not is still unclear without more comprehensive information.

Randomization (or lack of) is another concern. The group administration procedures make it difficult to select individual students at random. As a result, each school identified the most representative classroom(s) for each school; the methods of identifying these classrooms in not entirely known. The veracity of choosing classrooms that are generalizable to their respective schools could not be determined; it may not be possible to find a classroom that provides a true representation of the school. Thus, even the most representative classroom may have qualitative differences from the rest of the school and lack true generalization. A within-school comparison of the students in the classrooms selected compared to the remaining sample on the first probe revealed no significant difference between students on any of the three composites; this result provides evidence that students' math skills were initially similar. In addition, the combination of the modeling procedures and schools' attempt to generalize should minimize this concern.

Analytical Limitations. Although this study provides evidence for the relative predictive power of the MIR:M, these results should be seen as a preliminary evidence for all levels of the model. In particular, the teacher level estimates, while substantially reducing variation in the model, should be seen as tentative given the small sample size of teachers. As noted Bryk and Raudenbaush (2002), RMLE provides more unbiased estimates of the covariance parameters, and in particular the level-2 estimates of small samples; however, RMLE is likely to overestimate these effects when the sample size is particularly small. It is not known whether these effects are overestimated, but they should be interpreted with caution until validated. The strong predictive potential is clear; the extent of this power and the overall generalizability are less clear. Furthermore, additional classroom variables were not accounted for in the prediction at the teacher level.

TCAP Data Limitations. The TCAP data are characterized by limitations. One major limitation is the lack of psychometric data for these TCAP tests. Basic components, such as reliability and validity, are available for the TerraNova, the test that preceded the TCAP (Ciczek, 2005); however, these components for the version of the TCAP used in this study are not readily available. More specifically, technical properties are not publically available; moreover, information reported to school districts, teachers, and parents does not provide technical information. The limited information obtained is concerning from a research perspective; since the test is used statewide, the practical implications for this study may be unaffected by this lack of data. We assume that the TCAP test used in this study meets basic reliability and validity standards; however, we cannot verify the validity of this assumption without published information.

Another limitation from the TCAP data is the inclusion of two separate tests, the TCAP and the MAAS. Since these are different tests and yield different scores, it is difficult to make a comparison without further information about the data from each test. More specifically, the MAAS is designed for students in special education, and the test is meant to provide an easier way to measure the same constructs of the TCAP; however, each test uses a different measurement scale. Without more information, these tests were considered fundamentally different in the context of this study. What makes this problematic is that students taking the MAAS had to be deleted from the analyses. Since MAAS students were students in special education, the exclusions of their data resulted in an exclusion of students with distinct academic profiles compared to their peers. Fortunately, this should have limited impact on the results since only 2% of students are eligible to take the MAAS (Tennessee Department of Education). In addition, some of these students were already excluded from analyses because of missing MIR:M data.

Another troubling aspect of the TCAP data is the unknown derivation of the TCAP scaled scores. The inconsistent intervals of spacing between scores resulted in outliers during initial data screening. The analysis of raw scores indicated that a raw score point difference resulted in a 2 point score difference at the center of the distribution but as many as 58 point difference at the extremes. Thus, the outliers were more of a result of this unequal distribution. In addition, a near perfect relationship between the raw score and scaled score was expected, but less than 93% of the variance in scaled scores can be attributed to the raw scores. While there may be sound logic to the derivation of scores, the lack of psychometric data and explanations of the transformations eliminated the usefulness of the scaled scores.

MIR:M Data Limitations. Missing data at the item level is also a possible shortcoming of this study. It was clear that students skipped problems throughout the testing, but the impact of this was not quantified within this study. In addition, some students may have changed their responding style (e.g., focused on certain item types) throughout the duration of the data collection. While it extended beyond the scope of this research, it may be useful to account for the number of items attempted and skipped by item-type to obtain the most comprehensive data of responding. This may help identify students that disproportionally attempted or skipped different item types throughout the study. Differentiating between students who incorrectly answered but completed certain item from students who chose not to complete an item may impact the findings of the research and have implications for educators using these instruments.

Future Research

In the future, researchers should address the primary limitations of this study. First, additional research should include a more representative sample across more districts, schools, and more teachers within schools, which will increase the generalizability of the results and provide a more complete modeling of the components. Obtaining data from more districts can account for confounding variables (e.g., districts' curriculum). In addition, all but two of the schools in the present study have data from a single teacher. This makes it difficult to model the effect of nesting teachers within schools, which may be an important hierarchical level. Finally, this will allow for a more heterogeneous sample of individuals students to identify more distinct differences that may not have been evident in this study.

Additional research may examine the impact of each of the four item types since there were important findings and differences compared to the total score as a function of breaking down the data into the MR and MC composites. Differences in item scores may identify trends that affect the composites and the total score as a whole. In addition, there may be additional predictive validity data that are specific to the individual item types.

The impact of exploring additional teacher and student demographic characteristics could provide useful information about the factors that impact growth and predictive potential. Student level variables (e.g., attendance, free and reduced lunch status, disciplinary data, etc.) and teacher level variable (e.g., experience, education level, administrator evaluation, etc.) may account for additional variability between teachers and within classrooms. Although value-added modeling studies have looked at these teacher-level variables, it has not been done within the context of skill growth.

Finally, modeling reading skills could also offer valuable information about the growth of skills. Given the impact that reading could have on the MIR:M's predictive potential, modeling these skills could quantify their effects on the prediction of TCAP and other high-stakes testing. Modeling reading may also increase the predictive potential, which would be especially useful for educators.

REFERENCES

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25*(1), 95–135.

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

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*(4), 545–57. doi:10.1037/0021-843X.112.4.545\

American Institute for Research, National Center for Response to Intervention. (2013). Retrieved from http://www.rti4success.org/

Ardoin, S., & Christ, T. (2009). Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental passage set. *School Psychology Review*, *38*(2), 266–283.

Ardoin, S. P., Witt, J. C., Connell, J. E., & Koenig, J. L. (2005). Application of a three-tiered Response To Intervention model for instructional planning, decision making, and the identification of children in need of services. *Journal of Psychoeducational Assessment*, *23*(4), 362–380. doi:10.1177/073428290502300405

Ardoin, Scott P, & Christ, T. (2008). Evaluating curriculum-based measurement slope estimates using data from triannual universal screenings. *School Psychology Review*, *37*(1), 109–125.

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Kame'enui, E. J., & Beck, C. T. (2008). Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. *School Psychology Review*, *37*(1), 18–37.

Bell, S. M., Hilton-Prillhart, A. McCallum, R. S., & Hopkins, M. B. (2009). *Monitoring Instructional Responsiveness: Reading (MIR:R)*. Unpublished test. University of Tennessee.

Biesanz, J. C, Deeb-Sossa, N., Papadakis, A. A, Bollen, K. A, & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods*, *9*(1), 30–52. doi:10.1037/1082-989X.9.1.30

Biesanz, J. C., & West, S. G. (2000). Personality coherence: Moderating self-other profile agreement and profile consensus. *Journal of Personality and Social Psychology*, *79*(3), 425–437. doi:10.1037//0022-3514.79.3.425

Biesanz, J. C., West, S. G., & Graziano, W. G. (1998). Moderators of self-other agreement: reconsidering temporal stability in personality. *Journal of Personality and Social Psychology*, *75*(2), 467–77.

Bodin, D., Pardini, D. A., Burns, T. G., & Stevens, A. B. (2009). Higher order factor structure of the WISC-IV in a clinical neuropsychological sample. *Clinical Neuropsychological, 15*(5), 417–424.

Braden, J. P., & Schroeder, J. L. (2004). High-stakes testing and No Child Left Behind: Information and strategies for educators. *Helping children at home and school II: Handouts for families and educators*, 73-77. Retrieved from http://www.nasponline.org/communications/spawareness/highstakes.pdf

Bradley, R., Danielson, L., & Doolittle, J. (2007). Responsiveness to Intervention: 1997 to 2007. *Teaching Exceptional Children*, *39*(5), 8–12.

Byrne, B.M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York, TN: Taylor and Francis.

Brown-Chidsey, R., Davis, L., & Maya, C. (2003). Sources of variance in curriculum-based measures of silent reading. *Psychology in the Schools*, *40*(4), 363–377. doi:10.1002/pits.10095

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*(1), 147–158. doi:10.1037//0033-2909.101.1.147

Bryk, A. S. and Raudenbush, S. W. (1992) *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Newbury Park, CA.

Burns, M. K., MacQuarrie, L. L., & Campbell, D. T. (1999). The difference between curriculum-based assessment and curriculum- based measurement : A focus on purpose and result. *Communiqué*, *27*(6).

Chin, W. (1998). Issues and opinion on structural equation modeling. *MIS quarterly*, *22*(1), vii–xvi.

Chard, D. J., Stoolmiller, M., Harn, B. A., Wanzek, J., Vaughn, S., Linan-Thompson, S., & Kame'enui, E. J. (2008). Predicting reading success in a multilevel schoolwide reading model. *Journal of Learning Disabilities*, *41*(2), 174–188.

Chingos, M. M., & West, M. R. (2011). Promotion and reassignment in public school districts: How do schools respond to differences in teacher effectiveness? *Economics of Education Review,* 30(3), 419-433.

Christ, T. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating standard error of the slope to construct confidence intervals. *School Psychology Review*, *35*(1), 128–133.

Christ, T. J., & Ardoin, S. P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology, 47*,55−75.

Christ, T. J., Zopluoglu, C., Long, J., & Monaghen, B., (2012). Curriculum-based measurement of oral reading: Quality of progress monitoring outcomes. *Exceptional Children*, *78*(3), 356–373.

Christ, T. J., Silberglitt, B., Yeo, S., & Cormier, D. (2010). Curriculum-based measurement of oral reading: An evalu- ation of growth rates and seasonal effects among students served in general and special education. *School Psychology Review, 39*, 447–462.

Christ, T. J, & Vining, O. (2006). Curriculum-based measurement procedures to develop multiple-skill mathematics computation probes: Evaluation of random and stratified stimulus-set arrangements. *School Psychology Review*, *35*(3), 387–400.

Christenson, S. L., Decker, D. M., Triezenberg, H. L., Ysseldyke, J. E., & Reschly, A. (2007). Consequences of high-stakes assessment for students with and without disabilities. *Educational Policy*, *21*(4), 662–690.

Cizek, G. (2005). Review of TerraNova, The Second Edition. In R. A. Spies & B. S. Plake (Eds.), The sixteenth mental measurements yearbook (pp. 1025-1030). Lincoln, NE: Buros Institute of Mental Measurements.

Clarke, B., Baker, S. K., & Chard, D. (2008). Best practices in mathematics assessment and intervention with elementary students. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (Vol. V), (pp.453-464). Bethesda, MD: National Association of School Psychologists.

Clarke, B., Baker, S., Smolkowski, K., & Chard, D. J. (2008). An analysis of early numeracy curriculum-based measurement: Examining the role of growth in student outcomes. *Remedial and Special Education*, *29*(1), 46–57. doi:10.1177/0741932507309694

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

Codding, R., Shikyo, M., Russo, M., Birch,S., Fanning, E., & Jaspen, D. (2007). Comparing mathematics interventions: Does initial level of fluency predict intervention effectiveness? *Journal of School Psychology, 45*, 603–617.

Coles, J.T., Bell, S.M., & McCallum, R.S. (2012, August) *Using RTI to identify academic-discrepant students*. Poster presented at the national convention for the American Psychological Association, Orlando, FL.

Coles, J.T., McCallum, R.S., & Bell, S.M. (2013, February). *Factor structure of the monitoring intervention progress: Math*. Poster presented at the national convention for the National Association of School Psychologists, Seattle, WA.

Conroy, D. E., Metzler, J. N., & Hofer, M. (2003). Factorial invariance and latent mean stability of performance failure appraisals. *Structural Equation Modeling*, *10*(3), 401–422.

Cooper, B., & Sureau, J. (2008). Teacher unions and the politics of fear in labor relations. *Educational Policy 22*(1), 86–105.

Costa, L. J., Hooper, S.R., McBee, M., Anderson, K.L., & Yerby, D.C. (2012). The use of curriculum-based measures in young at-risk writers: Measuring change over time and potential moderators of change. *Exceptionality, 20*, 1-19.

Cowen, J. M., Fleming, D. J., Witte, J. F., Wolf, P. J., & Kisida, B. (2013). School vouchers and student attainment: Evidence from a state-mandated study of Milwaukee's Parental Choice Program. *Policy Studies Journal*, *41*(1), 147–168. doi:10.1111/psj.12006

Crawford, L., Tindal, G., & Stieber, S. (2001). Using Oral Reading Rate to Predict Student Performance on Statewide Achievement Tests. *Educational Assessment*, *7*(4), 303–323. doi:10.1207/S15326977EA0704_04

Cronbach, L.J., Nageswari, R., & Gleser, G.C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology, 16*, 137-163.

Cullen, J. B., Jacob, B. A., & Levitt, S. D. (2005). The impact of school choice on student outcomes: An analysis of the Chicago Public Schools. *Journal of Public Economics*, *89*(5-6), 729–760. doi:10.1016/j.jpubeco.2004.05.001

Cullen, J. B., Jacob, B. A. & Levitt, S., (2006). The effect of school choice on participants: Evidence from randomized lotteries. *Econometrica*, *74*(5), 1191–1230.

Curran, P. J., & Hussong, A. M. (2003). The use of latent trajectory models in psychopathology research. *Journal of abnormal psychology*, *112*(4), 526–44. doi:10.1037/0021-843X.112.4.526

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, *125*(6), 627–668..

Dee, T., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*(3), 418–446. doi:10.1002/pam

Deere, D., & Strayer, W. (2001). *Putting schools to the test: School accountability, incentives. and behavior*. Unpublished Manuscript.

Deno, S L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*(3), 219–32.

Deno, S. (1992). The nature and development of curriculum-based measurement. *Preventing School Failure*, *36*(2), 5–10.

Deno, S. L. (2003). Developments in Curriculum-Based Measurement. *The Journal of Special Education*, *37*(3), 184–192. doi:10.1177/00224669030370030801

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

Deno, S. L., Reschly-Anderson, A., Lembke, E., Zorka, H., & Callender, S. (2002, March). *A model for school wide implementation: A case example.* Paper presented at the annual meeting of the National Association of School Psychology, Chicago, IL.

Dobbie, W., & Fryer, R. G., Jr. (2011). Are high-quality schools enough to increase achievement among the poor? Evidence from the Harlem Children's Zone. *American Economic Journal: Applied Economics, 3*, 158–187.

Duncan, T. E., Duncan, S. C., Okut, H., Strycker , L. A., Li, F. (2002).Anextension of the general latent variable growth modeling framework to four levels of the hierarchy. *Structural Equation Modeling: A Multidisciplinary Jounral, 9*(3), 303–326.

Duncan, T. E., Duncan, S. C., Strycker, L. A., Li, F.,& Alpert, A. (1999). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Espin, C., Wallace, T., Lembke, E., Campbell, H., & Long, J. D. (2010). Creating a Progress-Monitoring System in Reading for Middle-School Students: Tracking Progress Toward Meeting High-Stakes Standards. *Learning Disabilities Research & Practice*, *25*(2), 60–75. doi:10.1111/j.1540-5826.2010.00304.x

Flanagan, D. P., & McGrew, K. S. (1997). A cross-battery approach to assessing and interpeting cognitive abilities: Narrowing the gap between practice and cognitive science. In D. P. Flanagan D. P., Genshaft, J. L., & Harrison, P. L. (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 314-325). New York: Guilford

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2006). *Essentials of cross-battery assessment* (2nd ed.). New York: Wiley

Figlio, D. N., & Rouse, C. E. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics*, *90*(1-2), 239–255. doi:10.1016/j.jpubeco.2005.08.005

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education*, *41*(2), 121–139. doi:10.1177/00224669070410020101

Foegen, A., Espin, C. A., Allinder, R. M., & Markell, M. A. (2001). Translating research into practice: Preservice teachers' beliefs about curriculum-based measurement. *The Journal of Special Education*, *34*(4), 226–236.

Forbes, B. (2013). Effects of prompts and comprehension assessment on oral reading; Moderating effect of reading skills, *Unpublished Dissertation.*

Fuchs, L. (2003). Assessing intervention responsiveness: Conceptual and technical issues. Learning Disabilities: *Research & Practice, 18*, 172-186.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children*, *57*(6), 488–500.

Fuchs, L. S., Fuchs, D. (2007). A model for implementing Responsiveness to Intervention. *Teaching Exceptional Children*, *39*(5), 14–20.

Fuchs, L., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal, 21*(2), 449-460.

Fuchs, L. S., & Fuchs, D. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review*, *22*(1), 27–48.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989). Effects of instrumental use of curriculum-based measurement to enhance instructional programs. *Remedial and Special Education, 10*, 43-52.

Fuchs, L.J, Fuchs, D., Hamlett, C, Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22*, 27-48.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal, 28*, 617–641.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Hamlett, C. L. (2003). The potential for diagnostic analysis within curriculum-based measurement. *Assessment for Effective Intervention, 28*(3-4), 13-22.

Fuchs, L. S., Fuchs, D., & Zumeta, R. O. (2008). A curricular-sampling approach to progress monitoring: Mathematics concepts and applications. *Assessment for Effective Intervention*, *33*(4), 225–233. doi:10.1177/1534508407313484

Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1999). *Monitoring basic skills progress basic math manual.* Austin, TX: Pro-ed.

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Quarterly*, *18*(3), 157–171.

Fuchs, L. (2004). The, past, present, and future of curriculum-based measurement research. *School Psychology Review*, *33*(2), 188–192.

Glicking, E. E., Havertape, J. F. (1981). Curriculum-based assessment. In J. A. Tucker (Eds.), *Non-test-based assessment*. Minneapolis: National School Psychology Inservice Training Network, University of Minnesota.

Good, R. H., & Jefferson, G. (1998). Contemporary perspectives on curriculum-based measurement validity. In M. R. Shinn (Eds.), *Advanced applications of curriculum based measurement* (pp. 61–88). New York: Guilford Press.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, *5*(3), 257–288.

Gottfried, A. E., Marcoulides, G. A., Gottfried, A. W., Oliver, P. H., & Guerin, D. W. (2007). Multivariate latent change modeling of developmental decline in academic intrinsic math motivation and achievement: Childhood through adolescence.*International Journal of Behavioral Development, 31*(4), 317-327. doi:http://dx.doi.org/10.1177/0165025407077752

Graney, S. B., Missall, K. N., Martínez, R. S., & Bergstrom, M. (2009). A preliminary investigation of within-year growth patterns in reading and mathematics curriculum-based measures. *Journal of School Psychology*, *47*(2), 121–42. doi:10.1016/j.jsp.2008.12.001

Graney, S., & Shinn, M. (2005). Effects of reading curriculum-based measurement (R-CBM) teacher feedback in general education classrooms. *School Psychology Review, 34*, 184–201.

Graves, S. L., & Frohwerk, A. (2009). Multilevel modeling and school psychology: A review and practical example. *School Psychology Quarterly*, *24*(2), 84–94. doi:10.1037/a0016160

Gulek, C. (2003). Preparing for high-stakes testing. *Theory Into Practice*, *42*(1), 37–41.

Guskey, T. (2007). Multiple sources of evidence: An analysis of stakeholders' perceptions of various indicators of student learning. *Educational Measurement: Issues and Practice*.

Hagel, J., & Brown, J. S. (2002). Control vs. trust: Mastering a different management approach. Retrieved from http://www.johnhagel.com/paper_control.pdf

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*(3), 466–479. doi:10.1016/j.econedurev.2010.12.006

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis*, *19*(2), 141–164.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*(2), 297–327. doi:10.1002/pam.20091

Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics*, *4*(1), 131–157. doi:10.1146/annurev-economics-080511-111001

Harwell, M. (2010) .Review of the TerraNova Tests. In R. A. Spies J.F. Karlson, & K. F. Geisinger (Eds.), The eighteenth mental measurements yearbook (pp. 611-614). Lincoln, NE: Buros Institute of Mental Measurements.

Helwig, R., Anderson, L., & Tindal, G. (2002). Using a concept-grounded, curriculum-based measure in mathematics to predict statewide test scores for middle school students with LD. *The Journal of Special Education*, *36*(2), 102–112.

Hilton-Prillhart, A. (2011). Validation of the Monitoring Academic Progress: Reading (MAP: R): Development and investigation of a group-administered comprehension-based tool for RTI. *Unpublished Dissertation*.

Hinkle, R. W. (2011). Using growth rate of reading fluency to predict performance on statewide achievement tests. *Unpublished Dissertation*.

Hintze, J. M. (2001). The generalizability of CBM oral reading fluency measures across general and special education. *Journal of Psychoeducational Assessment*, *19*(2), 158–170. doi:10.1177/073428290101900205

Hintze, J. M., & Christ, T. J. (2004). An examinaton of variability as a function of passage variance in CBM progress monitoring. *School Psychology Review*, *33*(2), 204–217.

Hintze, J. M., & Shapiro, E. S. (1997). Curriculum-based measurement and literature-based reading: Is curriculum-based measurement meeting the needs of changing reading curricula? *Journal of School Psychology*, *35*(4), 351–375.

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review*, *34*(3), 372–386.

Hintze, J. M., Christ, T. J., & Keller, L. A. (2002). The generalizability of CBM survey-level mathematics assessments: Just how many. *School Psychology Review*, *31*(4), 514–528.

Hintze, J. M., Owen, S. V., Shapiro, E. S., & Daly, E. J. (2000). Research design and methodology section: Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly*, *15*(1), 52–68.

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

Holmstrom, B., & Milgrom, P., (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership and job design. *Journal of Law, Economics and Organization 7*, 24–52.

Hopkins, M. (2011). A validation of the Monitoring Academic Progress Mathematics: An experimental multidimensional group administered curriculum-based measure of mathematics fluency and problem solving. *Unpublished Dissertation*.

Hopkins, M., McCallum, S., Bell, S., & Mounger, A. (2010). *Monitoring Instructinal Responsiveness: Mathematics*. Knoxville, TN: Author.

Hosp, M. K., & Hosp, J. L. (2003). Curriculum-based measurement for reading, spelling, and math: How to do it and why. *Preventing School Failure*, *48*(1), 10–17.

Hussong, A. M., Curran, P. J., Moffitt, T. E., Caspi, A., & Carrig, M. M. (2004). Substance abuse hinders desistance in young adults' antisocial behavior. *Development and Psychopathology*, *16*(4), 1029–46.

Individuals with Disabilities Education Improvement Act of 2004, STAT. 2706 C.F.R. § 602 (2004).

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5-6), 761–796. doi:10.1016/j.jpubeco.2004.08.004

Jiban, C. L., & Deno, S. L. (2007). Using math and reading curriculum-based measurements to predict state mathematics test performance: Are simple one-minute measures technically adequate? *Assessment for Effective Intervention*, *32*(2), 78–89. doi:10.1177/15345084070320020501

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, *91*(433), 222–230.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80*, 437−447.

Kamata, A., Nese, J. F., Patarapichayatham, C., & Lai, C. F. (2012). Modeling nonlinear growth with three data points: Illustration with benchmarking data. *Assessment for Effective Intervention*, *38*(2), 105–116. doi:10.1177/1534508412457872

Keith T. Z., Fine J. G., Taub G. E., Reynolds M. R., & Kranzler J. H. (2006) Higher order, multi-sample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children-Fourth Edition: What does it measure? *School Psychology Quarterly. 35*(1):108–127.

Keller-Margulis, M. A., Mercer, S. H., & Shapiro, E. S. (2012). Differences in growth on math curriculum-based measures using triannual benchmarks. *Assessment for Effective Intervention*. doi:10.1177/1534508412452750

Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review, 37*(3), 374–390.

Kelley, B., Hosp, J. L., & Howell, K. W. (2008). Curriculum-based evaluation and math: An overview. *Assessment for Effective Intervention*, *33*(4), 250–256. doi:10.1177/1534508407313490

Klein, A. & Muthén, B. (2006). Modeling heterogeneity of latent growth depending on initial status. *Journal of Educational and Behavioral Statistics, 3*1, 357-375.

Lee, J., & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources: State NAEP 1990-2009 reading and math achievement gaps and trends. *Educational Evaluation and Policy Analysis*, *34*(2), 209–231. doi:10.3102/0162373711431604

Lei, P., & Wu, Q. (2007). An introduction to structural equation modeling: Issues and practical considerations. *Educational Measurement: Issues and Practice*, *26*(3), 33–43.

Littell, R.C., Miliken, G.A., Stroup, W.W., Wolfinger, R.D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, NC: SAS Institute.

Long., D., Tidwell, M., (n.d.) *TESTING 1-2-3: Your questions answered.* Tennessee Education Association. Retrieved from http://teateachers.org/sites/default/files/tvaastesting1-2-3.pdf

Marsh, H. (1993). Stability of individual differences in multiwave panel studies: Comparison of simplex models and one-factor models. *Journal of Educational Measurement*, *30*(2), 157–183.

McCallum, R.S., Bell, S.M., & Coles, J., (2012, August) *Screening twice exceptional students within an RTI model.* Poster accepted for presentation at the national convention for the American Psychological Association, Orlando, FL.

McCallum, R. S., Bell, S. M., Coles, J. T., Miller, K. C., Hopkins, M. B., & Hilton-Prillhart, A. (2013). A model for screening twice-exceptional students (gifted with learning disabilities) within a response to intervention paradigm. *Gifted Child Quarterly*, *57*(4), 209–222. doi:10.1177/0016986213500070

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T.A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67–101. doi:10.3102/10769986029001067

McGlinchey, M., & Hixson, M. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review*, *33*(2), 193–203.

McGrew, K. S. (2005). The Cattell-Horn-Carroll (CHC) theory of cognitive abilities: Past, present and future. In D. Flanagan, & Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues-Second Edition* (p.136-202). New York: Guilford Press

McGrew, K.. S., & Wodcock, R. W. (2001). Woodcock-Johnson III *Technical manual.* Itasca, IL: Riverside.

Meredith, W., Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55,* 107-122.

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, *16*(3), 283–301. doi:10.1016/S0272-7757(96)00081-7

Merton, R. K. (1968). The Matthew effect in science. *Science, 159*(3810), 56-63.

Miller, K. C. (2012). Predictive validation of the Monitoring Instructional Responsiveness: Reading (MIR:R): Investigation of a group-administered, comprehension-based tool for RTI implementation. *Unpublished Dissertation*.

Miller, K.C., Hays, E.A., Bell, S.M., McCallum, R.S., Hilton-Prillhart, A., Lyons, C., & Coles, J.T. (2013, February). *Predictive validation of CBM data slopes using varying time intervals*. Poster presented at the national convention for the National Association of School Psychologists, Seattle, WA.

Miller, N., DeLapp, R., & Driscoll, R. (2007). Group anxiety reduction in sixth grade students. *Education Resources Information Center, 11*, 1-8.

Missall, K. N., Mercer, S. H., Martinez, R. S., & Casebeer, D. (2012). Concurrent and longitudinal patterns and trends in performance on early numeracy curriculum-based measures in kindergarten through third grade. *Assessment for Effective Intervention*, *37*(2), 95–106. doi:10.1177/1534508411430322

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. Reston, VA: Author.

Neddenriep, C. E., Skinner, C. H., Hale, H., Oliver, R., & Winn, B. (2007). An investigation of the validity of reading comprehension rate: A direct, dynamic measure of reading comprehension. *Psychology in the Schools, 44*, 373-388.

No Child Left Behind Act, 20 U.S.C., (2001).

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, *45*(2), 365–397. doi:10.3102/0002831207308230

Phillips, N. B., Hamlett, C. L., Fuchs, L. S., & Fuchs, D. (1993). Combining classwide curriculum-based measurement and peer tutoring to help general educators provide adaptive education. *Learning Disabilities Research and Practice, 8*, 148–156.

Polignano, J. C., & Hojnoski, R. L. (2012). Preliminary evidence of the technical adequacy of additional curriculum-based measures for preschool mathematics. *Assessment for Effective Intervention*, *37*(2), 70–83. doi:10.1177/1534508411430323

Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment*, *23*(4), 326–338. doi:10.1177/073428290502300403

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational and Behavioral Statistics*, *13*(2), 85–116. doi:10.3102/10769986013002085

Raudenbush, S. W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, *92*(5-6), 1394–1415. doi:10.1016/j.jpubeco.2007.05.003

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

Renaissance Learning. (2002). *Star math*. Madison, WI: Author.

Richardson, R. D., Hawken, L. S., & Kircher, J. (2011). Bias using maze to predict high-stakes test performance among hispanic and spanish-speaking students. *Assessment for Effective Intervention*, *37*(3), 159–170. doi:10.1177/1534508411430320

Rouse, C. E. (1998). Private school vouchers and student achievement: An evaluation of the Milwaukee parental Choice program. *The Quarterly Journal of Economics*, *113*(2), 553–602.

Sayer, A. G., & Willett, J. B. (1998). A cross-domain model for growth in adolescent alcohol expectancies. *Multivariate Behavioral Research, 33*(4), 509-543.

Schoen, L., & Fusarelli, L. D. (2008). Innovation, NCLB, and the fear factor: The challenge of leading 21st-century schools in an era of accountability. *Educational Policy, 22*(1), 181-203.

Shapiro, E.S. (2011). *Academic skill problems: Direct assessment and intervention* (4th ed.). New York, NY: Guildford Press.

Shapiro, E. S., Keller, M. A., Lutz, G. J., Santoro, L. E., & Hintze J. M. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment*, *24*(1), 19–35. doi:10.1177/0734282905285237

Shinn, M. R. (1989). Identifying and defining academic problems: CBM screening and eligibility procedures. In M. R. Shinn (Eds.), *Curriculum based measurement: Assessing special children* (pp.90-129). New York: The Guilford Press.

Shin, J., Espin, C. A., Deno, S. L., & McConnell, S. (2004). Use of hierarchical linear modeling and curriculum-based measurement for assessing academic growth and instructional factors for students with learning difficulties. *Asia Pacific Education Review*, *5*(2), 136–148. doi:10.1007/BF03024951

Silberglitt, B., Burns, M. K., Madyun, N. H., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in the Schools*, *43*(5), 527–536. doi:10.1002/pits

Silberglitt, B., & Hintze, J. M. (2007). How much growth can we expect? A conditional analysis of R-CBM growth rates by level of performance. *Exceptional Children*, *74*(1), 71–84.

Silberglitt, B., Jimerson, S., Burns, M., & Appleton, J. (2006). Does the timing of grade retention make a difference? Examining the effects of early versus later retention. *School Psychology Review, 35*, 134–141.

Singer, J. D., & Willet, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* New York, NY: Oxford University Press.

Skinner, C. H., Neddenriep, C. E., Bradley-Klug, K. L., & Ziemann, J. M. (2002). Advances in curriculum-based measurement: Alternative rate measures for assessing reading skills in pre- and advanced readers. *Behavior Analyst Today, 3*, 270-281.

Skinner, C. H., Williams, J. L., Morrow, J. A., Hale, A. D., Neddenriep, C., & Hawkins, R. O. (2009). The validity of a reading comprehension rate: Reading speed, comprehension, and comprehension rates. *Psychology in the Schools, 46*, 1036-1047.

Springer, M. G. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review*, *27*(5), 556–563. doi:10.1016/j.econedurev.2007.06.004

Stage, S. (2001). Program evaluation using hierarchical linear modeling with curriculum-based measurement reading probes. *School Psychology Quarterly, 16*, 91–112.

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review*, *3*, 407–419.

Stoel, R. D., van den Wittenboer, G., & Hox, J. (2003). Analyzing longitudinal data using multilevel regression and latent growth analysis. *Metodologia de las Ciencias del Comportamiento*, *5*, 21–42.

Taub, G. E., Keith, T. Z., Floyd, R. G., & Mcgrew, K. S. (2008). Effects of general and broad cognitive abilities on mathematics achievement. *School Psychology Quarterly*, *23*(2), 187–198. doi:10.1037/1045-3830.23.2.187

Taylor, E., Hayes, E., Coles, J. T., McCallum, R. S., & Bell, S. M. (In Submission). Comparing prospective twice-exceptional students with high-performing peers on high-stakes of achievement.

Tennessee State Department of Education. (2013). *Tennessee Comprehensive Assessment Program*. Published test. State of Tennessee. Retrieved from http://tn.gov/education/assessment/achievement.shtml.

Thornblad, S. C., & Christ, T. J. (2014). Curriculum-based measurement of reading: Is 6 weeks of daily progress monitoring enough? *School Psychology Review*, *43*(1), 19–29.

Thum, Y. M. (2003). No Child Left Behind: Methodological challenges & recommendations for measuring adequate yearly progress. *Center for the Study of Evaluation Technical Report 590* (Vol. 1522, pp. 1–18.).

Thurber, R. S., Shinn, M. R., & Smolkowki, K. (2002). What is measured in mathematics test? Construct validity of curriculum-based mathematics measures. *School Psychology Review*, *31*(4), 498–513.

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, *1*, 31–65. doi:10.1146/annurev.clinpsy.1.102803.144239

Tucker, J. A. (1985). Curriculum-based assessment: an introduction. *Exceptional children*, *52*(3), 199–204.

Tucker, J. (1987). Curriculum-based assessment is not a fad. *The Collaborative Educator, 1, 4, 10.*

VanDerHeyden, A. M., & Burns, M. K. (2005). Using curriculum-based assessment and curriculum-based measurement to guide elementary mathematics instruction: Effect on individual and group accountability scores. *Assessment for Effective Intervention*, *30*(3), 15–31. doi:10.1177/073724770503000302

VanDerHeyden, A. M., & Burns, M. K. (2009). Performance indicators in math: implications for brief experimental analysis of academic performance. *Journal of Behavioral Education*, *18*(1), 71–91. doi:10.1007/s10864-009-9081-x

VanDerHeyden, A. M., Witt, J. C., & Barnett, D. W. (2005). The emergence and possible futures of Response To Intervention. *Journal of Psychoeducational Assessment*, *23*(4), 339–361. doi:10.1177/073428290502300404

Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and the potential problems. *Learning Disabilities Research & Practice, 18*(3), 137-146.

Volpe, R. J., Mcconaughy, S. H., & Hintze, J. M. (2009). Generalizability of classroom behavior problem and on-task scores from the direct observation form. *School Psychology Review*, *38*(3), 382–401.

Wesson, C. L., King, R. P., & Deno, S. L. (1984). Direct and frequent measurement of student performance: If it’s good for us, why don't we do it? *Learning Disability Quarterly*, *7*(1), 45–48.

Wiley, H. I., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for english learners. *Remedial and Special Education*, *26*(4), 207–214.

Williams, J. L., Skinner, C. H., Floyd, R. G., Hale, A. D., Neddenriep, C., & Kirk, E. (2011). Words correct per minute: The variance in standardized reading scores accounted for by reading speed. *Psychology in the Schools, 48*, 87-101.

Wood, D. E. (2006). Modeling the relationship between oral reading fluency and performance on a statewide reading test. *Educational Assessment*, *11*(2), 85–104. doi:10.1207/s15326977ea1102_1

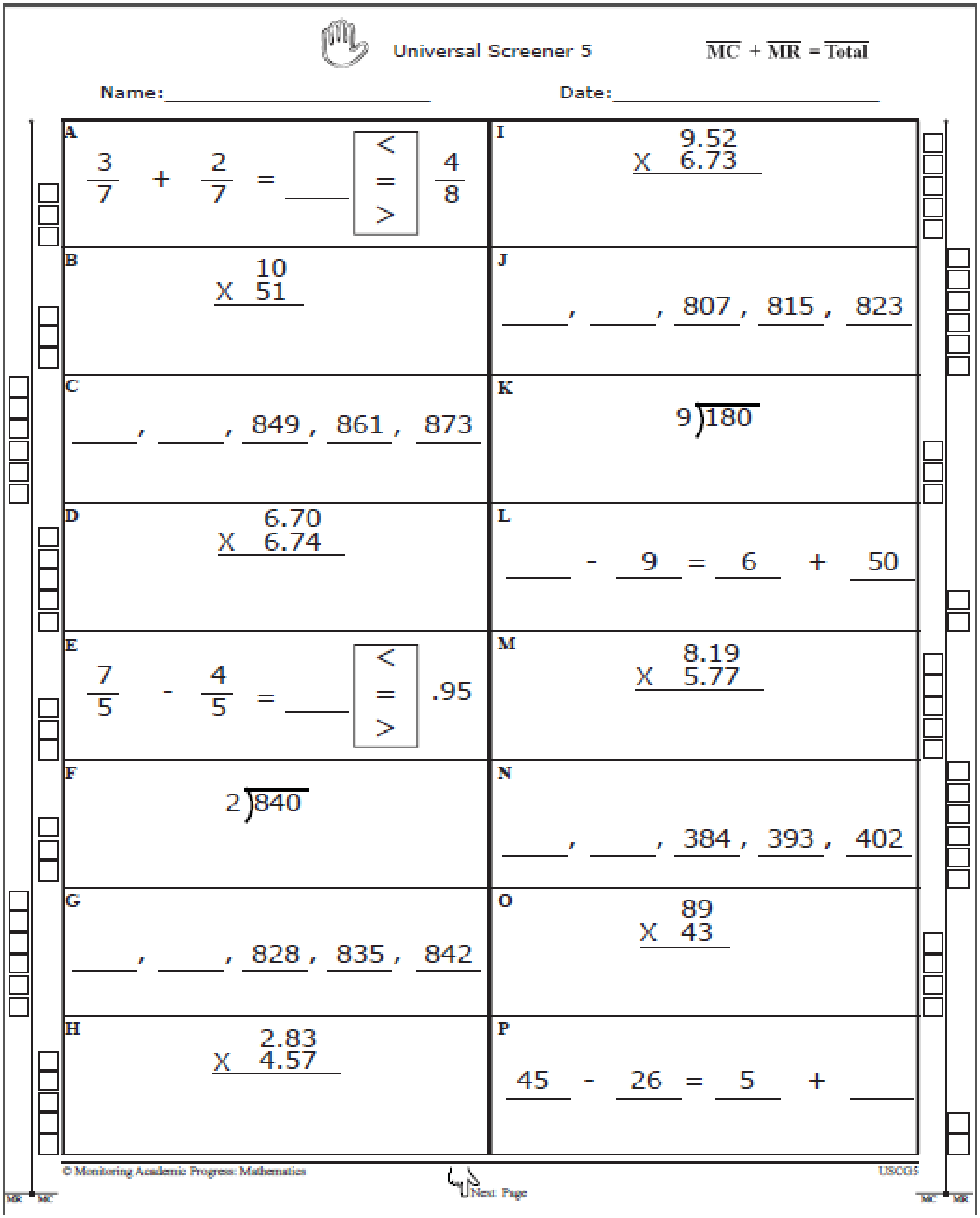
Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education*, *31*(6), 412–422. doi:10.1177/0741932508327463

Yeo, S., Fearrington, J. Y., & Christ, T. J. (2011). Relation between CBM-R and CBM-mR slopes: An application of latent growth modeling. *Assessment for Effective Intervention*, *37*(3), 147–158. doi:10.1177/1534508411420129

Yeo, Seungsoo, Kim, D.-I., Branum-Martin, L., Wayman, M. M., & Espin, C. A. (2012). Assessing the reliability of curriculum-based measurement: An application of latent growth modeling. *Journal of School Psychology*, *50*(2), 275–92. doi:10.1016/j.jsp.2011.09.002

APPENDIX

MIR: M EXAMPLE PROBE



Universal Screener 5 $\overline{MC} + \overline{MR} = \overline{Total}$

Name:______________________ Date:______________________

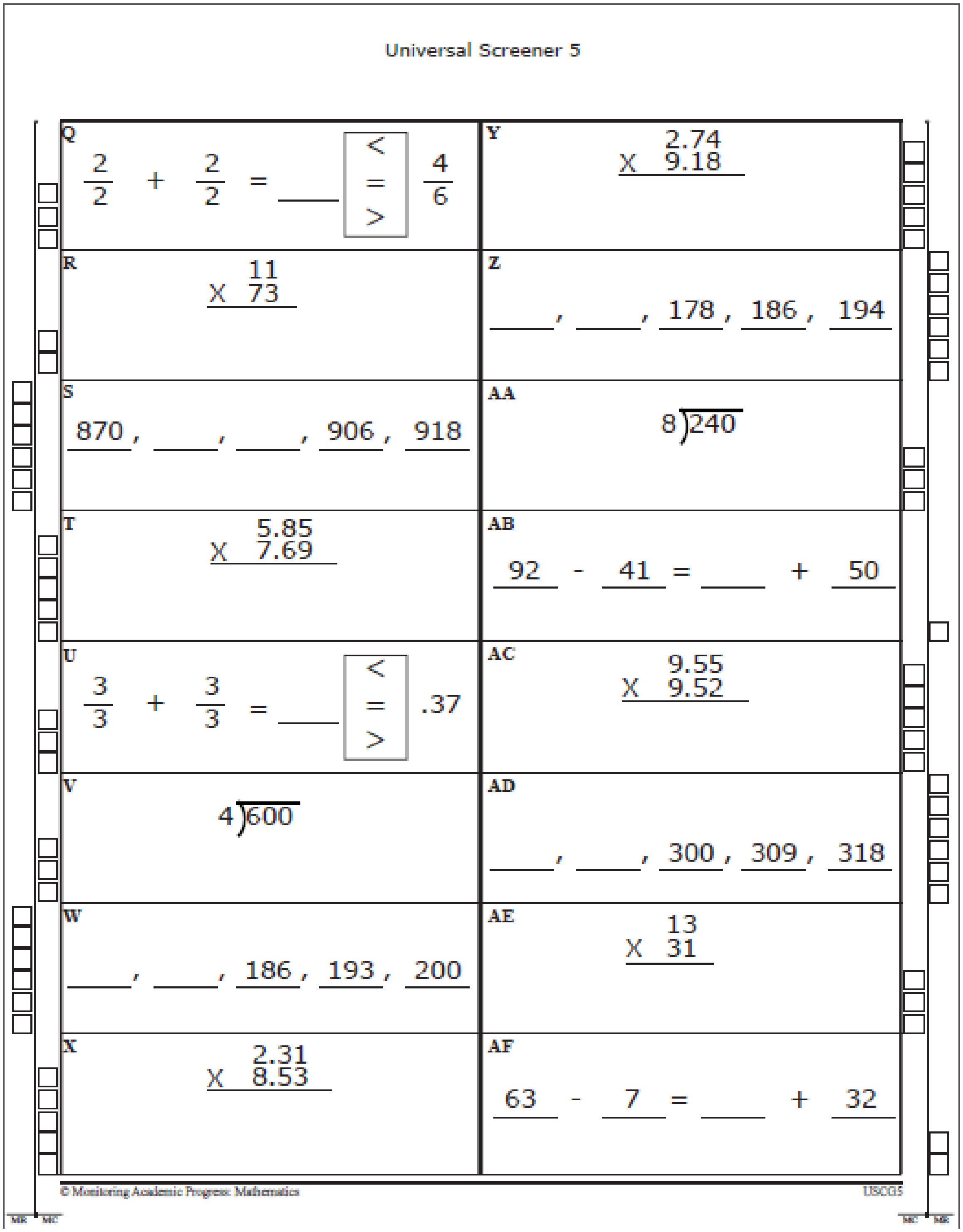
A $\frac{3}{7} + \frac{2}{7} =$ ____ $<$ $=$ $>$ $\frac{4}{8}$	**I** 9.52×6.73
B 10×51	**J** ____, ____, 807, 815, 823
C ____, ____, 849, 861, 873	**K** $9\overline{)180}$
D 6.70×6.74	**L** ____ - 9 = 6 + 50
E $\frac{7}{5} - \frac{4}{5} =$ ____ $<$ $=$ $>$ .95	**M** 8.19×5.77
F $2\overline{)840}$	**N** ____, ____, 384, 393, 402
G ____, ____, 828, 835, 842	**O** 89×43
H 2.83×4.57	**P** 45 - 26 = 5 + ____

© Monitoring Academic Progress: Mathematics

Next Page

USCG5

MR MC MC MR



Universal Screener 5

Q $\frac{2}{2} + \frac{2}{2} =$ ____ $\boxed{\begin{matrix}<\\=\\>\end{matrix}}$ $\frac{4}{6}$

R 11×73

S 870 , ____ , ____ , 906 , 918

T 5.85×7.69

U $\frac{3}{3} + \frac{3}{3} =$ ____ $\boxed{\begin{matrix}<\\=\\>\end{matrix}}$ .37

V $4\overline{)600}$

W ____ , ____ , 186 , 193 , 200

X 2.31×8.53

Y 2.74×9.18

Z ____ , ____ , 178 , 186 , 194

AA $8\overline{)240}$

AB 92 - 41 = ____ + 50

AC 9.55×9.52

AD ____ , ____ , 300 , 309 , 318

AE 13×31

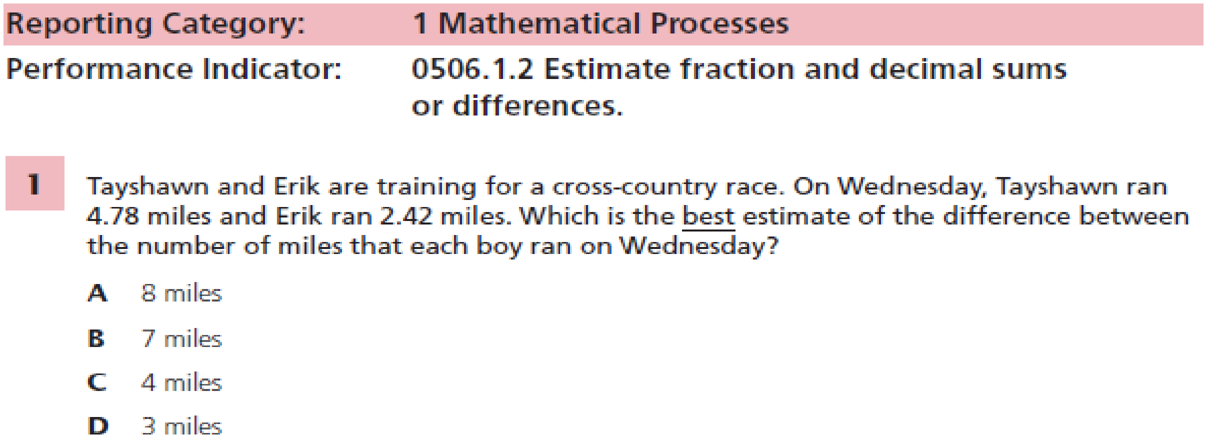
AF 63 - 7 = ____ + 32

© Monitoring Academic Progress: Mathematics

USCG5

MR MC MC MR

TCAP MATH PROCESSES ITEM SAMPLE



Reporting Category:	**1 Mathematical Processes**
Performance Indicator:	**0506.1.2 Estimate fraction and decimal sums or differences.**

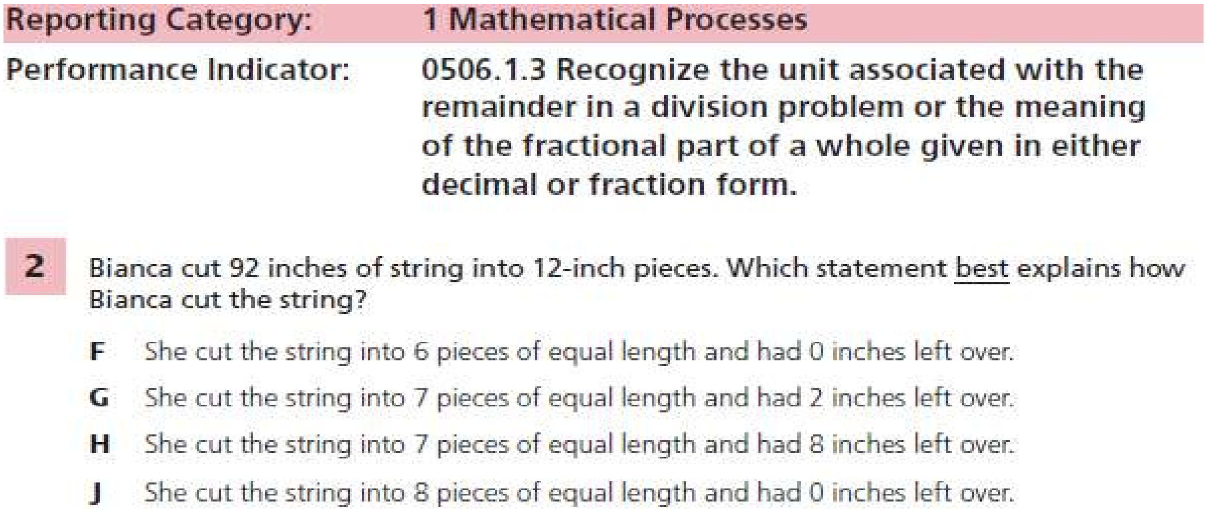
1 Tayshawn and Erik are training for a cross-country race. On Wednesday, Tayshawn ran 4.78 miles and Erik ran 2.42 miles. Which is the best estimate of the difference between the number of miles that each boy ran on Wednesday?

A 8 miles

B 7 miles

C 4 miles

D 3 miles



Reporting Category:	**1 Mathematical Processes**
Performance Indicator:	**0506.1.3 Recognize the unit associated with the remainder in a division problem or the meaning of the fractional part of a whole given in either decimal or fraction form.**

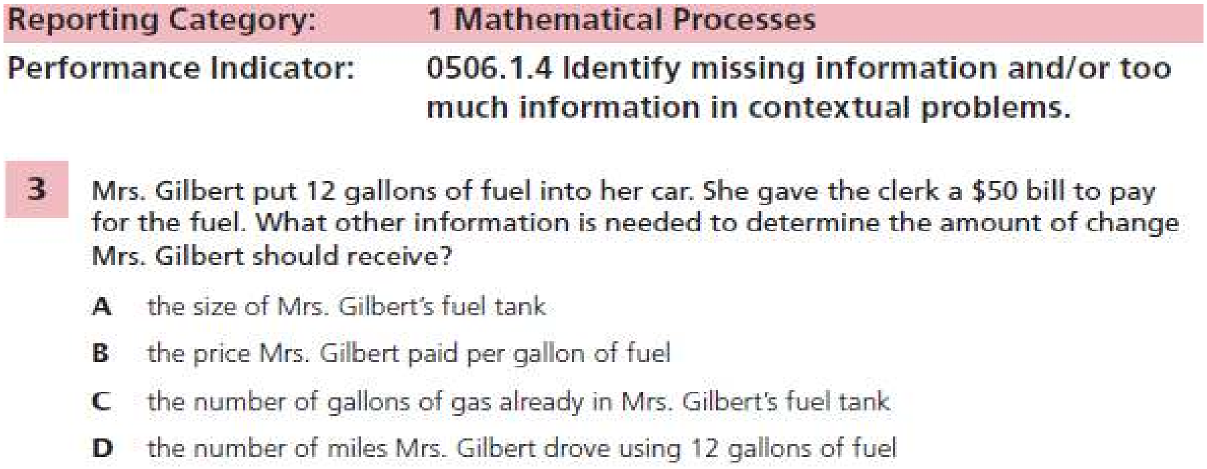
2 Bianca cut 92 inches of string into 12-inch pieces. Which statement best explains how Bianca cut the string?

F She cut the string into 6 pieces of equal length and had 0 inches left over.

G She cut the string into 7 pieces of equal length and had 2 inches left over.

H She cut the string into 7 pieces of equal length and had 8 inches left over.

J She cut the string into 8 pieces of equal length and had 0 inches left over.



Reporting Category:	**1 Mathematical Processes**
Performance Indicator:	**0506.1.4 Identify missing information and/or too much information in contextual problems.**

3 Mrs. Gilbert put 12 gallons of fuel into her car. She gave the clerk a $50 bill to pay for the fuel. What other information is needed to determine the amount of change Mrs. Gilbert should receive?

A the size of Mrs. Gilbert's fuel tank

B the price Mrs. Gilbert paid per gallon of fuel

C the number of gallons of gas already in Mrs. Gilbert's fuel tank

D the number of miles Mrs. Gilbert drove using 12 gallons of fuel

TCAP NUMBERS AND OPERATIONS ITEM SAMPLES

Reporting Category:	**2 Number and Operations**
Performance Indicator:	**0506.2.1 Read and write numbers from millions to millionths in various contexts.**

4 Rosalba converted meters to yards using the information below.

> One meter = one and nine hundred thirty-six ten-thousandths yards

How is one and nine hundred thirty-six ten-thousandths written in standard form?

F 1.936

G 1.0936

H 1.00936

J 1.000936

Reporting Category:	**2 Number and Operations**
Performance Indicator:	**0506.2.3 Select a reasonable solution to a real-world division problem in which the remainder must be considered.**

5 Mrs. Garrett has 23 students in her class. She will give 2 pencils to each student. She will buy pencils in packages of 6. What is the minimum number of packages of pencils Mrs. Garrett needs to buy to have enough pencils?

A 3

B 4

C 7

D 8

Reporting Category:	**2 Number and Operations**
Performance Indicator:	**0506.2.4 Solve problems involving the division of two- and three-digit whole numbers by one- and two-digit whole numbers.**

6 Mr. Farris used 133 yards of cloth to make window curtains. He made 7 curtains and used the same amount of cloth for each curtain. Exactly how many yards of cloth did Mr. Farris use for each curtain?

F 10

G 16

H 18

J 19

Reporting Category:	**2 Number and Operations**
Performance Indicator:	**0506.2.4 Solve problems involving the division of two- and three-digit whole numbers by one- and two-digit whole numbers.**

7 Bettina has a book with 364 pages. There are 14 chapters in her book. Each chapter has the same number of pages. How many pages are in each chapter of Bettina's book?

A 25

B 26

C 27

D 28

Reporting Category:	**2 Number and Operations**
Performance Indicator:	**0506.2.5 Solve addition and subtraction problems involving both fractions and decimals.**

8 **Subtract:**

$$2.38 - 1\frac{9}{100} =$$

F 0.48

G 1.29

H $1\frac{31}{100}$

J $1\frac{39}{100}$

Reporting Category:	**2 Number and Operations**
Performance Indicator:	**0506.2.6 Add and subtract proper and improper fractions as well as mixed numbers.**

9 **Solve:**

$$3\frac{4}{5} + 2\frac{7}{10} =$$

A $5\frac{1}{2}$

B $5\frac{3}{4}$

C $6\frac{1}{10}$

D $6\frac{1}{2}$

Reporting Category:	**2 Number and Operations**
Performance Indicator:	**0506.2.7 Recognize equivalent representations for the same number.**

10 Which fraction shows another way to write 32.4?

F $32\frac{1}{10}$

G $32\frac{1}{4}$

H $32\frac{2}{5}$

J $32\frac{4}{5}$

Reporting Category:	**2 Number and Operations**
Performance Indicator:	**0506.2.8 Write terminating decimals in the form of fractions or mixed numbers.**

11 Rachel spent 3.75 hours doing homework last night. Which fraction shows another way to write 3.75?

A $3\frac{3}{4}$

B $3\frac{5}{7}$

C $3\frac{1}{4}$

D $3\frac{1}{75}$

Reporting Category:	**2 Number and Operations**
Performance Indicator:	**0506.2.8 Write terminating decimals in the form of fractions or mixed numbers.**

12 The height of a new library building is 12.03 meters. Which fraction is equivalent to 12.03?

F $12\frac{3}{100}$

G $12\frac{3}{10}$

H $12\frac{1}{3}$

J $12\frac{3}{5}$

Reporting Category:	**2 Number and Operations**
Performance Indicator:	**0506.2.9 Compare whole numbers, decimals and fractions using the symbols <, >, and =.**

13 The widths of three skateboards are shown in the table below.

Skateboard Widths

Skateboard	Width in Inches
X	8.125
Y	$8\frac{1}{8}$
Z	8.18

Which number sentence correctly compares the widths of two of the skateboards?

A $8\frac{1}{8} = 8.18$

B $8\frac{1}{8} > 8.18$

C $8.125 = 8\frac{1}{8}$

D $8.125 > 8\frac{1}{8}$

Reporting Category:	**2 Number and Operations**
Performance Indicator:	**0506.2.9 Compare whole numbers, decimals and fractions using the symbols <, >, and =.**

14 Which inequality is true?

F $\frac{5}{12} > \frac{1}{2}$

G $\frac{5}{8} > \frac{1}{2}$

H $\frac{1}{2} > \frac{8}{15}$

J $\frac{1}{2} > \frac{3}{5}$

TCAP ALGEBRA ITEM SAMPLES

Reporting Category: 3 Algebra

Performance Indicator: 0506.3.1 Evaluate algebraic expressions involving decimals and fractions using order of operations.

15 What is the value of the expression below, when $p = 61$?

$$p + 2 \times 1.5$$

A 62.7

B 64

C 91

D 94.5

Reporting Category: 3 Algebra

Performance Indicator: 0506.3.2 Evaluate multi-step numerical expressions involving fractions using order of operations.

16 Evaluate: $\frac{1}{3} + \frac{1}{2} \times \frac{1}{2}$

F $\frac{1}{4}$

G $\frac{2}{7}$

H $\frac{5}{12}$

J $\frac{7}{12}$

Reporting Category:	**3 Algebra**
Performance Indicator:	**0506.3.3 Find the unknown in single-step equations involving fractions and mixed numbers.**

17 What value of p makes this equation true?

$$p = 4\frac{1}{5}$$

A $\frac{20}{5}$

B $\frac{21}{5}$

C 19

D 21

Reporting Category:	**3 Algebra**
Performance Indicator:	**0506.3.4 Given a set of values, identify those that make an inequality a true statement.**

18 Look at the inequality below.

$$x -| 6 \leq 10$$

Which set contains only values of x that make this inequality true?

F {16, 18, 20}

G {15, 17, 19}

H {14, 16, 18}

J {12, 14, 16}

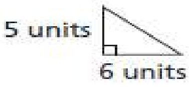
TCAP MEASUREMENT AND GEOMETRY ITEM SAMPLES

Reporting Category:	**4 Geometry and Measurement**
Performance Indicator:	**0506.4.1 Solve contextual problems that require calculating the area of triangles and parallelograms.**

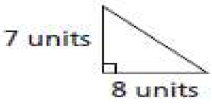
19 Justin drew a triangle that has an area of 15 square units. Which triangle shown below could be Justin's triangle?

Area $- \frac{1}{2}$ base × height

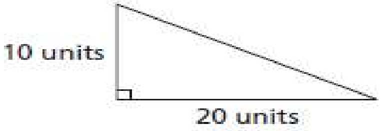
A



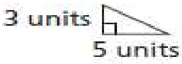
B



C

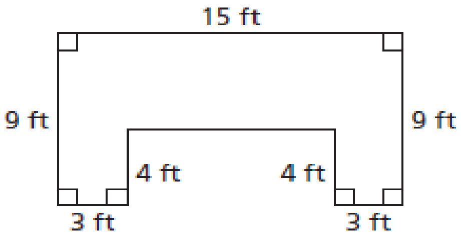


D



Reporting Category:	**4 Geometry and Measurement**
Performance Indicator:	**0506.4.2 Decompose irregular shapes to find perimeter and area.**

20 The diagram below shows the dimensions of the stage floor Mrs. Wilkins is building for a school play.



Area of Rectangle = length × width

What is the area, in square feet (sq ft), of the stage floor?

F 99 sq ft

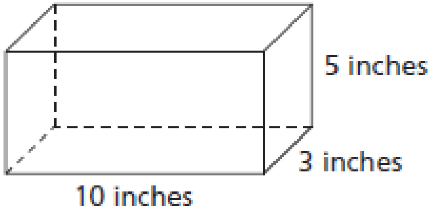
G 75 sq ft

H 54 sq ft

J 47 sq ft

Reporting Category:	**4 Geometry and Measurement**
Performance Indicator:	**0506.4.4 Solve problems involving surface area and volume of rectangular prisms and polyhedral solids.**

21 The dimensions of a rectangular prism are shown below.



Volume – length × width × height

What is the volume of this rectangular prism?

A 18 cubic inches

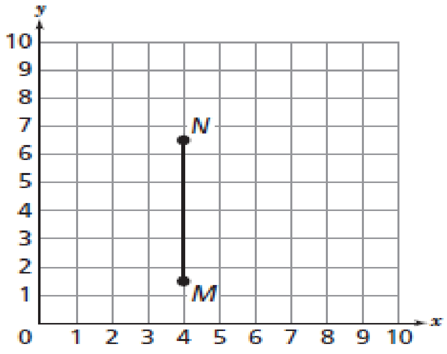
B 35 cubic inches

C 150 cubic inches

D 190 cubic inches

Reporting Category:	**4 Geometry and Measurement**
Performance Indicator:	**0506.4.5 Find the length of vertical or horizontal line segments in the first quadrant of the coordinate system, including problems that require the use of fractions and decimals.**

22 Look at the coordinate grid below.



Which is closest to the length of Line Segment *NM*?

F 5.0 units

G 5.5 units

H 6.0 units

J 6.5 units

TCAP DATA ANALYSIS, STATISTICS AND PROBABILITY ITEM SAMPLES

Reporting Category:	**5 Data Analysis, Statistics and Probability**
Performance Indicator:	**0506.5.1 Depict data using various representations, including decimal and/or fractional data.**

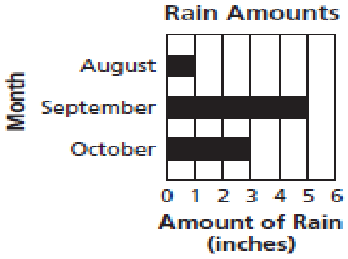
23 Darryl recorded the number of inches of rain in a city during 3 months in the table below.

Rain Amounts

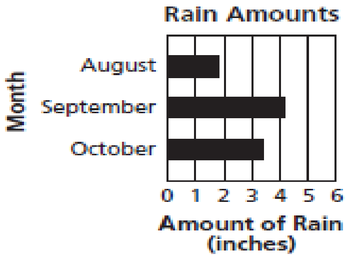
Month	Amount of Rain (inches)
August	1.8
September	5.2
October	3.4

Which graph best represents the data in the table?

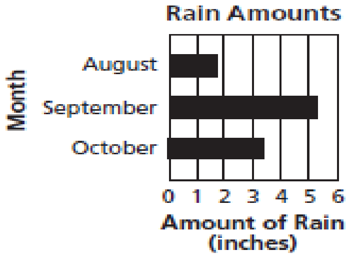
A



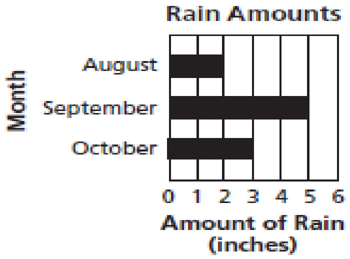
B



C



D



Reporting Category:	**5 Data Analysis, Statistics and Probability**
Performance Indicator:	**0506.5.3 Calculate measures of central tendency to analyze data.**

24 Terrence counted the number of cars that passed by his school every five minutes before he had to go inside. The data he recorded are shown below.

30, 28, 29, 32, 40, 37, 28

What is the mode of the data?

F 28

G 30

H 32

J 40

TABLES

Table 1

Descriptive Statistics of MIR:M Total by Probe

Probe	N	School Day[a]	Mean	SD	Minimum	Maximum
1	210	3.62	13.93	8.72	3	74
2	213	13.49	18.32	9.46	2	69
3	210	21.31	15.44	7.68	2	62
4	212	31.27	20.31	10.79	2	65
5	212	43.72	17.61	9.90	1	70
6	217	66.47	18.65	9.40	4	58
7	192	77.83	18.79	10.10	1	70
8	216	84.58	20.06	10.38	1	75
9	218	95.36	21.28	10.84	2	79
10	212	104.89	22.15	11.40	3	73
11	207	125.34	21.31	11.25	1	79
12	212	134.61	21.55	11.06	2	68

[a]- average school day of administration

Table 2

Descriptive Statistics of MIR:M Math Calculation by Probe

Probe	N	School Day[a]	Mean	SD	Minimum	Maximum
1	210	3.62	9.05	5.57	0	39
2	213	13.49	10.07	5.29	0	26
3	210	21.31	9.48	4.98	0	28
4	212	31.27	12.16	6.84	0	35
5	212	43.72	11.07	5.83	0	29
6	217	66.47	12.38	6.81	0	58
7	192	77.83	12.72	6.65	0	32
8	216	84.58	13.33	7.38	0	37
9	218	95.36	14.27	7.02	0	34
10	212	104.89	16.16	7.13	1	36
11	207	125.34	15.93	7.60	1	37
12	212	134.61	15.09	7.75	0	37

[a]- average school day of administration

Table 3

Descriptive Statistics of MIR:M Math Reasoning by Probe

Probe	N	School Day[a]	Mean	SD	Minimum	Maximum
1	210	3.62	4.88	5.66	0	39
2	213	13.49	8.25	6.86	0	46
3	210	21.31	5.96	6.15	0	35
4	212	31.27	8.15	8.29	0	48
5	212	43.72	6.54	8.06	0	48
6	217	66.47	6.27	7.28	0	35
7	192	77.83	6.06	7.63	0	46
8	216	84.58	6.73	7.61	0	51
9	218	95.36	7.01	8.18	0	52
10	212	104.89	5.99	8.36	0	45
11	207	125.34	5.38	7.61	0	54
12	212	134.61	6.46	7.98	0	41

[a]- average school day of administration

Table 4

Correlations of MIR:M Total by Probe

Probe	1	2	3	4	5	6	7	8	9	10	11	12
1												
2	.45											
3	.46	.63										
4	.56	.61	.63									
5	.47	.58	.69	.72								
6	.44	.49	.60	.54	.67							
7	.46	.57	.63	.64	.73	.64						
8	.47	.49	.56	.61	.67	.65	.76					
9	.52	.53	.53	.61	.67	.66	.76	.79				
10	.46	.52	.57	.66	.63	.62	.78	.78	.79			
11	.50	.52	.58	.59	.61	.59	.75	.75	.79	.76		
12	.51	.57	.55	.66	.65	.60	.67	.77	.77	.76	.82	

Note: All p-values significant at .0001 level

Table 5

Correlations of MIR:M Math Calculation by Probe

Probe	1	2	3	4	5	6	7	8	9	10	11	12
1												
2	.41											
3	.48	.68										
4	.46	.61	.62									
5	.45	.54	.63	.66								
6	.46	.45	.52	.50	.60							
7	.47	.53	.63	.62	.68	.61						
8	.52	.46	.58	.58	.67	.63	.72					
9	.46	.42	.51	.57	.66	.64	.67	.76				
10	.45	.52	.65	.64	.69	.55	.68	.72	.71			
11	.53	.52	.57	.59	.66	.60	.69	.75	.70	.70		
12	.55	.54	.58	.61	.68	.64	.65	.75	.72	.72	.83	

Note: All p-values significant at .0001 level

Table 6

Correlations of MIR:M Math Reasoning by Probe

Probe	1	2	3	4	5	6	7	8	9	10	11	12
1												
2	.46											
3	.52	.54										
4	.54	.53	.60									
5	.50	.54	.64	.66								
6	.40	.39	.58	.49	.68							
7	.40	.48	.56	.55	.65	.72						
8	.38	.44	.50	.47	.62	.70	.73					
9	.48	.44	.44	.49	.61	.71	.76	.72				
10	.38	.44	.44	.53	.53	.63	.75	.74	.77			
11	.37	.43	.40	.44	.53	.59	.70	.66	.81	.77		
12	.37	.44	.39	.52	.58	.55	.63	.67	.73	.71	.77	

Note: All p-values significant at .0001 level

Table 7

MIR:M Total Growth Models

Parameter	Model A	Model B	Model C	Model D	Model E	Model F
Fixed Effects						
Intercept	19.002 (0.547)	15.794 (0.584)	15.822 (0.498)	15.107 (0.627)	15.216 (0.544)	15.227 (0.546)
Time		0.479 (0.030)	0.476 (0.048)	0.812 (0.117)	0.769 (0.115)	0.763 (0.134)
Time^2				-0.025 (0.008)	-0.022 (0.008)	-0.021 (0.009)
Random Effects						
Level-2						
Intercept	62.617 (6.325)	63.363 (6.356)	44.362 (5.217)	63.378 (6.356)	44.697 (5.247)	46.011 (6.234)
Time			0.330 (.357)		0.392 (.357)	1.491 (0.387)
Covariance- Int./Time			0.412 (.049)		0.329 (.049)	-1.064 (1.191)
Time^2						0.005 (0.002)
Covariance- Int./Time^2						0.083 (0.078)
Covariance- Time/Time^2						-0.075 (0.025)
Level-1						
Within-students (σ^2)	45.490 (1.339)	41.011 (1.207)	34.759 (1.077)	40.874 (1.203)	34.639 (1.074)	33.354 (1.093)
-2*log-likelihood	17475.8	17241.8	17049.3	17240.8	17049.3	17034.5
Level-2 Pseudo R^2	-	(-1.2)	(29.2)	-42.9 (-1.2)	-0.8 (28.6)	-3.7 (26.5)
Level-1 Pseudo R^2	-	(9.8)	(23.6)	-17.6 (10.1)	0.3 (23.9)	4.0 (26.7)
Total Pseudo R^2	-	(3.5)	(26.8)	-31.8 (3.6)	-0.3 (26.6)	-0.3 (26.6)

Model A-Unconditional Means Model

Model B- Fixed Linear Growth with Random Intercept Only

Model C-Unconditional Linear Growth

Model D-Fixed Linear and Quadratic Growth with Random Intercept Only

Model E-Fixed Linear and Quadratic Growth with Random Intercept and Linear Growth

Model F-Fixed Linear and Quadratic Growth with Random Intercept and Linear and Quadratic Growth

Note: Pseudo R^2 represents percentage of variation accounted for from Unconditional Growth Model

Note: Pseudo R^2 in parentheses represents percentage of variation accounted for from Unconditional Means Model

Note: Negative Pseudo R^2 indicates an increase in variation from the comparison model

Table 8

MIR:M Math Calculation Growth Models

Parameter	Model A	Model B	Model C	Model D	Model E	Model F
Fixed Effects						
Intercept	12.557 (0.353)	9.099 (0.378)	9.136 (0.296)	8.789 (0.409)	8.881 (0.331)	8.883 (0.333)
Time		0.518 (0.020)	0.511 (0.029)	0.664(0.078)	.635 (0.077)	0.631 (0.091)
Time^2				-0.019 (0.006)	-.009 (0.005)	-0.009 (0.006)
Random Effects						
Level-2						
Intercept	25.658 (2.632)	26.269 (2.646)	14.555 (1.853)	26.282 (2.647)	14.657 (1.864)	15.348 (2.323)
Time			0.112 (0.018)		0.112 (0.018)	0.675 (0.174)
Covariance- Int./Time			0.525 (0.130)		0.518 (0.131)	-0.214 (0.488)
Time^2						0.005 (0.001)
Covariance- Int./Time^2						0.044 (0.033)
Covariance- Time/Time^2						0-.037 (0.012)
Level-1						
Within-students (σ^2)	23.465 (0.691)	18.236 (0.539)	16.055 (0.497)	18.213 (0.536)	16.033 (0.496)	15.342 (0.502)
-2*log-likelihood	15753	15176.6	14999	15184.4	15004.2	14986.7
Level-2 Pseudo R^2	-	(-2.4)	(43.3)	-80.6 (-2.4)	-0.7 (42.9)	-5.4 (40.2)
Level-1 Pseudo R^2	-	(22.3)	(31.6)	-13.4 (22.4)	0.1 (31.7)	4.4 (34.6)
Total Pseudo R^2	-	(9.4)	(37.7)	-45.4 (9.4)	-0.3 (37.5)	-0.3 (37.5)

Model A-Unconditional Means Model

Model B- Fixed Linear Growth with Random Intercept Only

Model C-Unconditional Linear Growth

Model D-Fixed Linear and Quadratic Growth with Random Intercept Only

Model E-Fixed Linear and Quadratic Growth with Random Intercept and Linear Growth

Model F-Fixed Linear and Quadratic Growth with Random Intercept and Linear and Quadratic Growth

Note: Pseudo R^2 represents percentage of variation accounted for from Unconditional Growth Model

Note: Pseudo R^2 in parentheses represents percentage of variation accounted for from Unconditional Means Model

Note: Negative Pseudo R^2 indicates an increase in variation from the comparison model

Table 9

MIR:M Math Reasoning Growth Models

Parameter	Model A	Model B	Model C	Model D	Model E	Model F
Fixed Effects						
Intercept	6.447 (0.388)	6.707 (0.420)	6.693 (0.384)	6.321 (0.458)	6.360 (0.419)	6.375 (0.386)
Time		-0.039 (0.024)	-0.036 (0.040)	0.149 (0.092)	0.125 (0.090)	0.121 (0.105)
Time^2				-0.014 (0.007)	-0.012 (0.006)	-0.012 (0.007)
Random Effects						
Level-2						
Intercept	31.331 (3.188)	31.313 (3.186)	26.350 (3.094)	31.305 (3.185)	26.395 (3.098)	20.903 (3.093)
Time			.0257 (0.035)		0.257 (0.035)	0.970 (0.238)
Covariance- Int./Time			-0.432 (0.242)		-0.434 (0.242)	0.268 (0.652)
Time^2						0.003 (0.001)
Covariance- Int./Time^2						-0.067 (0.042)
Covariance- Time/Time^2						-0.047 (0.015)
Level-1						
Within-students (σ^2)	25.649 (0.755)	25.632 (0.755)	20.781 (0.644)	25.595 0.754)	20.754 (0.643)	19.905 (0.653)
-2*log-likelihood	16000.7	16003.7	15766.3	16007.5	15770.7	15734.0
Level-2 Pseudo R^2	-	(0.1)	(15.9)	-18.8 (0.1)	-0.2 (15.8)	20.7 (33.3)
Level-1 Pseudo R^2	-	(0.1)	(19.0)	-23.2 (0.2)	0.1 (19.1)	4.2 (22.4)
Total Pseudo R^2	-	(0.1)	(17.3)	-20.7 (0.1)	0.0 (17.3)	13.4 (28.4)

Model A-Unconditional Means Model
Model B- Fixed Linear Growth with Random Intercept Only
Model C-Unconditional Linear Growth
Model D-Fixed Linear and Quadratic Growth with Random Intercept Only
Model E-Fixed Linear and Quadratic Growth with Random Intercept and Linear Growth
Model F-Fixed Linear and Quadratic Growth with Random Intercept and Linear and Quadratic Growth
Note: Pseudo R^2 represents percentage of variation accounted for from Unconditional Growth Model
Note: Pseudo R^2 in parentheses represents percentage of variation accounted for from Unconditional Means Model
Note: Negative Pseudo R^2 indicates an increase in variation from previous model

Table 10

Descriptive Statistics of TCAP Scales

Variable	Mean	SD	Minimum	Maximum	Skewness	Kurtosis
Raw Total	40.54	12.58	14	64	-0.057	-1.122
Scaled Score	746.13	37.79	610	900	-0.144	1.750
Mathematical Properties	60.51	18.65	20	100	0.190	-0.983
Numbers and Operations	64.97	22.32	5	100	-0.257	-1.078
Algebra	62.40	23.17	10	100	-0.108	-1.246
Geometry and Measurement	61.74	19.73	21	100	0.112	-1.052
Data Analysis, Statistics, and Probability	64.40	17.34	13	100	0.001	-0.747

Table 11

Correlations of TCAP Scales

Variable	A.	B.	C.	D.	E.	F.	G.
A. Raw Total							
B. Scaled Score	.964						
C. Mathematical Properties	.927	.894					
D. Numbers and Operations	.978	.938	.851				
E. Algebra	.946	.910	.836	.929			
F. Geometry and Measurement	.948	.922	.914	.893	.844		
G. Data Analysis, Statistics, and Probability	.955	.927	.990	.923	.871	.886	

Note: All p-values significant at .0001 level

Table 12

Correlations of TCAP Scales by MIR:M Components

	MIR:M Scales								
	Total			Math Calculation			Math Reasoning		
Variable	Int[a]	Slope[b]	Quad[c]	Int[a]	Slope[b]	Quad[c]	Int[a]	Slope[b]	Quad[c]
Raw Total	.452	.285	-.122	.456	.162	.154	.226	.256	-.198
Scaled Score	.416	.279	-.134	.435	.129	.261	.197	.273	-.212
Mathematical Processes	.398	.245	-.098	.396	.111	.053	.213	.246	-.198
Numbers and Operations	.448	.262	-.193	.452	.167	.058	.219	.225	-.174
Algebra	.432	.279	-.117	.444	.182	-.13	.284	.235	-.168
Geometry and Measurement	.443	.316	-.162	.433	.152	.116	.232	.333	-.244
Data Analysis, Statistics, Prob.	.492	.264	-.143	.413	.146	.277	.300	.237	-.183

[a]-Intercept

[b]-Linear Slope

[c]-Quadratic Slope

Table 13

Correlations of TCAP Scales by MIR:M Predicted Values at TCAP Administration

	MIR:M Predicted Models		
Variable	Total	Math Calculation	Math Reasoning
Raw Total	.526	.433	.345
Scaled Score	.492	.392	.337
Mathematical Processes	.463	.366	.316
Numbers and Operations	.540	.430	.317
Algebra	.510	.429	.327
Geometry and Measurement	.532	.470	.377
Data Analysis, Statistics, Probability	.487	.399	.320

Note: Predicted Values from Quadratic Growth Models

Table 14

Percentage of Variance by MIR:M Components, MIR:M Scale, and Hierarchical Levels

	MIR:M Scores		
Parameter	Total	Math Calculation	Math Reasoning
Level-2 (Between-Teachers)			
Intercept	39.7%	46.0%	5.4%
Linear Slope	-3.5%	-8.1%	-6.5%
Quadratic Slope	-7.3%	5.7%	-8.5%
Level-1 (Between-Students)			
Intercept	18.1%	12.6%	6.4%
Linear Slope	9.8%	9.0%	8.8%
Quadratic Slope	6.1%	4.8%	5.1%
Combined Levels			
Intercept	26.1%	24.9%	6.0%
Linear Slope	4.9%	2.6%	3.1%
Quadratic Slope	1.2%	5.1%	0.0%

Note: Student level variables are fixed and random variables are free to vary within class
Note: The three components (i.e., intercept and slopes) were modeled individually.

Table 15

Best Predictive Model for Total Composite, MR Global, and MC Global Scores

Parameter	Unconditional	Total	MC	MR
Fixed Effects				
Intercept	40.126 (2.302)	40.494 (1.854)	40.490 (1.740)	40.365 (2.231)
Level-2 (Teacher Level)				
MIR:M Intercept		1.212 (0.557)	2.191 (0.715)	2.378 (1.51)
Level-1 (Student Level)				
MIR:M Intercept		0.710 (0.195)	1.155 (0.365)	.731 (0.212)
MIR:M Linear Slope[a]		7.194 (0.195)		7.781 (1.804)
MIR:M Quadratic Slope[a]		95.92 (35.711)		115.74 (44.565)
Random Effects				
Level-2 (Teacher Level)				
Intercept (τ_{π})	58.081(26.835)	36.917 (18.181)	31.358 (15.973)	54.408 (26.237)
Level-1 (Student Level)				
Intercept (σ^2)	99.143 (9.647)	72.641 (7.364)	86.963 (8.690)	84.423 (0.196)
MIR:M Intercept		.269 (0.218)	.856 (0.700)	.089 (0.196)
-2*log-likelihood	1682.9	1611.2	1655.3	1635.1
Level 1-Pseudo R^2	-	26.7%	12.3%	14.8%
Level 2- Pseudo R^2	-	36.4%	46.0%	6.3%
Total- Pseudo R^2	-	30.3%	24.9%	11.7%

[a]-Indicates grand mean centered variable

Table 16

Comparison of Common Modeling Procedures of Growth

	Regression	**Latent Growth Modeling**	**Growth Curve Modeling**
Modeling Framework	Ordinary Least Squares	Structural Equation Modeling	Hierarchical Linear Modeling
Handling Missing Data	Poor	Fair	Good
Estimation Flexibility	Poor	Good	Good
Model Specification	Fair	Fair	Good
Individual Intercepts	Poor	Good	Good
Individual Slope	Poor	Good	Good
Varying Time Intervals	Poor	Good	Good
Varying Time Occasions By Subjects	Poor	Poor	Good
Between-Subject Model	Good	Good	Good
Within-Subject Model	Poor	Good	Good
Time-Varying Covariates	Fair	Good	Fair
Modeling Error Structures	Poor	Good	Fair
Modeling Hierarchical Levels	Poor	Fair	Good
Growth Model as part of larger model	Poor	Good	Fair

Note: This is a simplified comparison of modeling procedures and is based on the utility of the procedures for this study. For detailed comparisons refer to Stoel et al. 2003; Tomarken & Waller, 2005; Curran & Hussong, 2003.

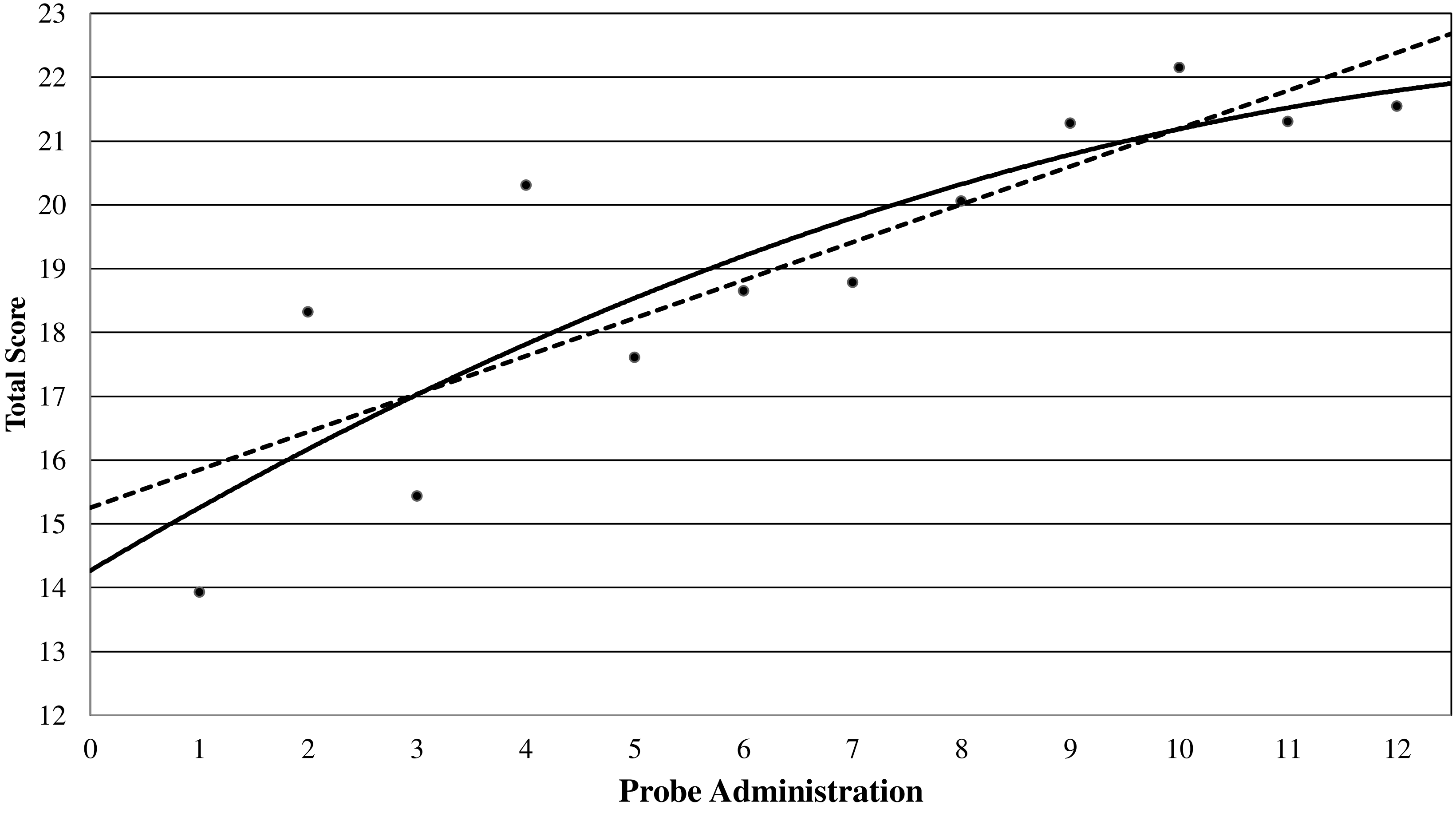


Figure 1 Mean Total Score by Probe Administration
Note: Dashed line represents linear trend
Note: Solid line represent quadratic trend

VITA

Jeremy Thomas Coles graduated from The Ohio State University in June of 2008 with a Bachelor of Arts in Psychology. He received his Master of Education in School Psychology at Bowling Green State University in August of 2010; he subsequently enrolled at the University of Tennessee in School Psychology. He is currently completing his pre-doctoral internship at Sweetwater City School District within the Tennessee Internship Consortium. He will graduate in August of 2014 with a PhD in School Psychology and a Master of Science in Statistics.