

To the Graduate Council:

I am submitting herewith a [dissertation or thesis] written by Michael Brandon Hopkins entitled “A Validation of the Monitoring Academic Progress Mathematics: An Experimental Multidimensional Group Administered Curriculum-Based Measure of Mathematics Fluency and Problem Solving.” 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

Vena M. Long

Michael K. Smith

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

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

A Validation of the Monitoring Academic Progress Mathematics: An Experimental
Multidimensional Group Administered Curriculum-Based Measure of Mathematics
Fluency and Problem Solving

A dissertation
Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Michael Brandon Hopkins
December 2011

Copyright 2011 © by Michael Brandon Hopkins
All rights reserved.

Dedication

To my wife, Amanda, son, Owen, and my parents, Mike and Dorothy, and my brother Aaron, for your unconditional love, support, and sacrifice.

“This moment contains all moments.”

– C.S. Lewis

Acknowledgements

This dissertation has been made possible by the contributions of numerous professional's generous contributions. I thank Dr. Steve McCallum for his guidance in helping me to better understand not only the technical aspects of test construction but also the practical outcomes and aspects of our profession that affect the lives of children and families. I thank Dr. Sherry Bell for the countless hours spent editing and reviewing the probes each week and for coordinating meetings, personnel, and other critical logistics. The experience of working with Dr. Bell and Dr. McCallum was exceptional and I appreciate their willingness to extend themselves and attempt innovative approaches to solving old problems.

I thank Dr. Vena Long and Dr. Mike Smith for serving on my committee. Dr. Long provided excellent resources and information concerning mathematics assessment. Dr. Smith was very helpful in providing information about curriculum-based measurement and the response to intervention process from a practical perspective.

I also acknowledge Steve Dunkin and Robyn O'Dell who were essential in carrying out the systems level leadership including training, grading probes, administering probes, and recording probe scores. Shannon Kamer was also central in the development of key technology. I also thank all of the principals, teachers, and assistance of the participating school system for your time and effort in completing this project.

Thank you also to my research group, including but not limited to Angela Hilton, Kelli Miller, Ashlee Cornett, and Kristin Maurer, who also spent hours upon hours over the year editing and listening to endless conversations about curriculum-based

measurement, whether you wanted to or not. Thank you all for your help, support, and patience.

Thank you Amanda and Owen, my wife and son, for your unwavering love and support throughout. I apologize, yet am thankful, that you both now know more about academic assessment than you ever cared to know. I also want to thank Dr.

Doofenshmirtz and Bertny who have both kept us smiling over the past year. I thank the Lord, my God. Through him, all things, including this are possible.

Abstract

The study investigated the psychometric properties of a newly developed math curriculum-based measure, the Monitoring Academic Progress: Mathematics (MAP:M), through examination of its internal consistency, alternate-form, slope, and test-retest reliability and validity. Participants included 1688 first through third-grade students from a school district in Northeast Tennessee. Application of Generalizability Theory produced reliability coefficients, score variances, and standard-error-of-measures (SEM) for both absolute and relative decisions based on a particular number of probes. MAP:M reliability coefficients for relative decisions ranged from .67 to .97 across eleven probes. The highest percentage of score variance at all three grades was attributed to the Person facet (academic growth), 58% to 66%, while variation in the Item facet (test difficulty) contributed the least score variance (11% to 16%). The SEMs for relative decisions ranged from 7.06 after one probe to 1.70 after eleven probes. SEMs were higher for absolute decisions. Results indicate that one MAP:M probe provides adequate reliability, although two probes provide good reliability estimates, exceeding .80. Alternate-form reliabilities' ranged from .65 to .83 at first grade, .51 to .79 at second grade, and .59 to .80 at third grade. Lower reliabilities are best explained by increased time between administrations. Test-retest reliabilities were .81, .79, and .76 for first, second, and third grades, respectively.

Correlation coefficients between the MAP:M and Monitoring Basic Skills Progress: Computation (MBSP) show adequate concurrent validity estimates ranging from .58 to .75 for first and third grades. Second-grade concurrent validity estimates were

somewhat lower and ranged from .37 to .52. A stepwise multiple regression indicates that the MAP:M provides stronger relative predictive power than does the MBSP when predicting end-of-year STAR Math scores, , $R^2 = .33$, $F(28) = 13.11$, $p < .01$. Results support adequate to strong reliability and validity of the MAP:M, in addition to successful implementation in a school environment.

Table of Contents

DEDICATION	III
ACKNOWLEDGEMENTS	IV
ABSTRACT	VI
TABLE OF CONTENTS	VIII
LIST OF TABLES	XI
LIST OF ACRONYMS	XIII
CHAPTER 1	1
Literature Review	1
Introduction	2
Curriculum-Based Assessment	3
Curriculum-Based Measurement	7
Responsiveness-to-Intervention and Curriculum-Based Measurement	8
Mathematics Curriculum-Based Measurement	13
Traditional Estimates of Reliability	13
Generalizability Theory	16
Mathematics Curriculum-Based Measurement Item Types	18
Math facts	19
Computation	20
Discrimination	23
Missing number.	24
Counting	24
Number Identification	25
Concepts and Applications.	26
Problem solving.	27
Statement of Purpose	27
Research Questions	29
CHAPTER II	31
Method	31

Qualitative pilot study.....	31
Participants and setting.	31
Instruments.....	32
Procedures.....	35
Results.....	35
School district collaboration.	35
First school district meeting.....	35
Second school district meeting.	36
Third school district meeting.	38
Fourth school district meeting.	41
Quantitative Pilot Study One.	42
Participants and setting.	42
Instruments.....	43
Procedures.....	44
Results.....	45
Modifications based on findings.....	49
Quantitative Pilot Study Two.....	49
Participants and setting.	49
Instruments.....	50
Procedures.....	50
Results.....	52
Summary.....	52
Current Study.....	53
Participants and setting.	53
Instruments.....	53
Monitoring Basic Skills Progress.....	53
STAR Math.....	54
Monitoring Academic Progress: Mathematics.....	54
Procedures.....	59
Training activities.	59
MAP:M probe administration.	59
Scoring directions.	61
Administration fidelity.....	62
Statistical procedures.	62
CHAPTER III.....	64
Results.....	64
G-Theory Estimates of Reliability.....	65
Estimated variance components.....	65
Relative decisions: G-Coefficients and standard error of measure.....	66

Absolute decisions: Index of dependability and standard error of measure.	68
Classical Test Theory Estimates of Reliability	70
Reliability of the Slope	72
Validity of the MAP:M.....	74
CHAPTER IV	76
Discussion	76
Uses of curriculum-based measurement of mathematics.....	76
G-Theory Estimates of Reliability	77
Estimated variance components.....	77
Relative decisions: G-Coefficients and standard error of measure.....	78
Absolute decisions: Index of Dependability and standard error of measure.	80
Classical Test Theory Estimates of Reliability	81
Reliability of the Slope	83
Validity of the MAP:M.....	84
Limitations and Future Research	85
REFERENCES	88
APPENDIX A.....	99
APPENDIX B	125
VITA.....	137

List of Tables

TABLE 1 MAP:M Quantitative Pilot Study One: Time-Unlimited Means, Standard Deviations, and Independent <i>t</i> tests.	100
TABLE 2 MAP:M Quantitative Pilot Study One: Time-Unlimited First-Grade Difficulty and Discrimination Indices by Item Type.....	101
TABLE 3 MAP:M Quantitative Pilot Study One: Time-Unlimited Second-Grade Difficulty and Discrimination Indices by Item Type.	102
TABLE 4 MAP:M Quantitative Pilot Study One: Time-Unlimited Third-Grade Difficulty and Discrimination Indices by Item Type.....	103
TABLE 5 MAP:M Quantitative Pilot Study One: Time-Limited Means, Standard Deviations, and Independent <i>t</i> tests.	104
TABLE 6 MAP:M Quantitative Pilot Study Two: Mean, Standard Deviations, Internal Consistency, and Standard Error of Measure	105
TABLE 7 MAP:M Quantitative Pilot Study Two: First-Grade Difficulty and Discrimination Indices.	106
TABLE 8 MAP:M Quantitative Pilot Study Two: Second-Grade Difficulty and Discrimination Indices.	107
TABLE 9 MAP:M Quantitative Pilot Study Two: Third-Grade Difficulty and Discrimination Indices.	108
Table 10 Demographic Information of the Current Study.....	109
Table 11 Descriptive Statistics for MAP:M and MBSP	110
Table 12 Grade 1 Generalizability Study Results: Estimates of Variance Components	111

Table 13 Grade 1 Decision Study Results	112
Table 14 Grade 2 Estimates of Variance Components	113
Table 15 Grade 2 Decision Study Results: Monitoring Academic Progress: Mathematics	114
Table 16 Grade 3 Estimates of Variance Components	115
Table 17 Grade 3 Decision Study Results: Monitoring Academic Progress: Mathematics	116
Table 18 Alternate Form Reliability: MAP:M.....	117
Table 19 Alternate Form Reliability: Monitoring Basic Skills Progress	118
Table 20 Alternate Form Reliability Correlation Coefficient Comparisons.....	119
Table 21 Test-Retest Reliability : MAP:M.....	120
Table 22 Slope and Standard Error of the Slope by Grade for the MAP:M and MBSP	121
Table 23 Concurrent Validity: Monitoring Academic Progress and Monitoring Basic Skills Progress.....	122
Table 24 Stepwise Multiple Regression Analysis Predicting STAR Math with MAP:M and MBSP Across Grades 2 and 3	123
Table 25 Correlations of MAP:M Probe Seven, MBSP, and STAR Math Across Second and Third Grades.....	124

List of acronyms

Cognitively Guided Instruction (CGI)
Comprehensive Test of Basic Skills (CTBS)
Computation (COMP)
Curriculum-based assessment (CBA)
Curriculum-based measurement (CBM)
Discovery Education Predictive Assessment: Math (DEPA:M)
Individuals with Disabilities Education Improvement Act (IDEIA)
Iowa Test of Basic Skills (ITBS)
Math Computation Test (MCT)
Mathematics-CBM (M-CBM)
Monitoring Academic Progress: Mathematics (MAP:M)
Monitoring Basic Skills Progress: Computation (MBSP)
National council of Teachers of Mathematics (NCTM)
No Child Left Behind Act (NCLB)
Number Pattern (NP)
Number sentence-quantity discrimination (NSQD)
Problem-solving model (PSM)
Quantity Discrimination using symbols (QDS)
Quantity discrimination using words (QDW)
Receiver Operator Characteristic (ROC)
Response to intervention (RTI)
Shape Pattern (SP)
Standard Deviation (SD)
Standard Error of Measure (SEM)
Standard-treatment-protocol model (STPM)
Stanford Achievement Test (SAT)
Wide-Range Achievement Test (WRAT)
Woodcock-Johnson III (WJ-III)
Woodcock-Johnson Revised Edition (WJ-R)

Chapter 1

Literature Review

Traditional mathematics curriculum-based measures (M-CBM) have relied on computation and other single-task probes to screen and track students' mathematics progress. Typically, these measures contain considerable measurement error even though they assess only one facet of mathematics (computation), i.e., they are limited by mono-operational bias. Consequently, in this study I describe the development and validation of a M-CBM, Monitoring Academic Progress: Mathematics (Hopkins, McCallum, Bell, & Mounger, 2010) which incorporates elements of traditional M-CBM, in addition to problem-solving elements to create a multi-dimensional ecologically valid assessment. Specifically this study addressed the following questions.

1. What are the percentages of score variance accounted for by the following facets: person (p), items (i), and the residual (p x i, e) of the (Monitoring Academic Progress: Mathematics (MAP:M), relative to a more traditional M-CBM, the Monitoring Basic Skills Progress: Computation (MBSP)?
2. What is the generalizability coefficient and standard error of measure (SEM) for relative decisions based on student's raw scores across one, two, three, five, seven, and ten probes for the MAP:M versus the MBSP?
3. What is the index of dependability and SEM for absolute decisions based on student's raw scores across one, two, three, five, seven, and ten probes for the MAP:M versus the MBSP?
4. What is the alternate-form reliability of the MAP:M versus the MBSP?

5. What is the test-retest reliability of the MAP:M?
6. What is the reliability of the slope of the MAP:M versus the MBSP?
7. What is the concurrent validity of the MAP:M with the MBSP as determined by a correlation coefficient expressing the relationship between the two?
8. What is the relative predictive power of the MAP:M and the MBSP when the criterion is STAR Math scores?

Introduction

The passage of No Child Left Behind (NCLB; 107th Congress, 2002) Act and the 2004 reauthorization of the Individuals with Disabilities Education Improvement Act (IDEIA; "Individuals with Disabilities Education Improvement Act of 2004,") has brought formative educational evaluation to the forefront in determining the presence of a learning disability and monitoring the progress of students with learning disabilities. NCLB placed states and school districts under increased pressure to provide at-risk students with scientifically based education and interventions from highly-qualified teachers, while imposing more stringent accountability standards. Prior to these legislative actions, accountability standards were assessed primarily using summative evaluations. Summative evaluations are state mandated and conducted at the end of the school year. Scores are intended to reflect reductions in the achievement gap between diverse student groups (107th Congress, 2002), although they provide inadequate information regarding individual student's rate of growth. In addition to yearly summative assessments, IDEIA encourages the use of formative assessment for purposes of screening, intervention, and special education classification. This additional

information provides educators and decision-makers routine academic progress monitoring.

The new legislation allows the use of innovative procedures for the identification of a learning disability. That is, not only can a learning disability be defined as an achievement-ability discrepancy, the traditional process (Tennessee Department of Education, 2004), but a learning disability can also be defined as a failure to respond to empirically supported interventions. This paradigm shift requires not only a change in the referral process but also a significant change in the assessment methodology used to identify students with specific learning disabilities, such as monitoring student academic performance using frequent formative evaluations (Torrance & Pryor, 1998). Educational programming decisions (R. Sattler, 1989, p. 120) are then made based on student's responsiveness to intervention (Deno, Fuchs, Marston, & Jongho, 2001; L. Fuchs & Fuchs, 1997) as demonstrated by evaluation of the data.

Curriculum-Based Assessment

Because formative assessment is a central feature of the new response to intervention (RTI) process and because curriculum-based assessment (CBA) systems are commonly used within RTI, I discuss briefly the history of CBA including the four most widely known procedures, and then the emergence of curriculum-based measurement (CBM) as the dominate CBA.

Although formative evaluations have a long history in education, these evaluation schemas were not systematically defined until the late 1980's. At that time, four prominent formative evaluations, characterized as CBA, began to appear in the literature.

These proposed systems varied from the use of teacher created assessments to the utilization of commercial academic achievement instruments. Although there are subtle, yet distinct differences across the models, each propose to use the actual classroom curriculum.

One of the early CBA models outlined by Blankenship (1985), required assessments to be derived from curriculum material to “obtain direct and frequent measures of a student’s performance on a series of sequentially arranged objectives” (p. 234). This approach assumes that student skill acquisition occurs in a sequential and predictable manner. Instructional time is devoted to promoting mastery of specific educational skills, while providing less time teaching skills already mastered. Pre-tests purport to measure student prerequisite performance on a particular academic skill. Performance on CBA determines the instructional goal, focusing efforts toward skills not yet mastered. Administering multiple CBA’s over the course of a few days provides evidence of a student’s typical performance on a given skill. Finally, a post-test is administered to determine if a student has achieved mastery of the targeted skill (Blankenship, 1985). This model of CBA is capable of tracking student performance on a successive series of isolated academic skills. The goal of the model is to maximize the amount of instructional time; however, students’ overall rate of growth cannot be determined. That is, instruction focuses on to-be-learned material, not on skills the students already possess. The potential problem with this model is the reliance on an assessment with no demonstrated reliability or validity for assessing the particular skill in

question. This can potentially result in inaccurate assessment of student's level of skill mastery.

A second CBA model relies on an analysis of sub-skills required to complete a given educational task (Howell & Morehead, 1987). This CBA model is similar to that described by Blankenship although the focus lies in assessment of splinter skills as opposed to the mastery of a skill as a whole. Through the identification of problematic sub-skill development, the instructor can provide customized course content targeting splinter skills. One striking difference between Howell and Morehead's (1987) model and Blankenship's (1985) model is Howell and Morehead's recommended use of a combination of teacher-designed, commercial and normative-based assessments, as long as they are representative of the curriculum content (Howell, 1986). The underlying assumption of this CBA model is that the development of specific sub-skills is central to learning a particular skill. Once a sub-skill deficit is identified and remediated through modifications in content, pacing, and difficulty (Howell, 1986), the skill should then be acquired.

A third CBA perspective contends that a student's lack of academic achievement is a curriculum artifact (e.g., curriculum pace) rather than an inherent within-the-student variable (Gickling, 1985). In other words, low student performance is not due to the student but rather the pace, language used, or other mediating factor in the curriculum itself. In order to monitor the impact of these mediating factors ongoing assessment is used to determine "the instructional needs of the students based on the student's ongoing performance in existing content" (Gickling, 1985, p. 206). This model of CBA is similar

to that of Blankenship, although Gickling uses the assessment information to prescribe changes of the instructional materials or strategies rather than focusing on skill mastery in the current materials. This model of CBA does not require tracking student progress but instead attempts to identify the instructional level and place the child within an appropriate educational environment.

The three aforementioned formative assessment models require that the teacher develop the CBA. A potential pitfall of using teacher-developed measures is the lack of instrument validity and/or reliability. Reliability is critical for accurate assessment; it is necessary because it provides an operationalization of systematic (non-error) variance. In addition, validity of CBA is important when assessing student mastery. If an instrument is not valid, decisions related to skill mastery are inherently inaccurate because the instrument is not measuring the target skill. Without psychometric data (i.e., reliability and validity), practitioners cannot determine whether a given CBA is measuring the intended construct.

Fuchs and Fuchs (1986) set curriculum-based CBM apart from other CBA's by focusing on descriptions of its standardization procedure and psychometrics. They define this type of CBA-based formative assessment as an inductive and psychometrically acceptable method of evaluating student performance within the natural classroom setting. Formative assessment is inductive because decisions begin ongoing assessment, using reliable and valid measures, that provides educators with a database of information in order to make empirically based educational decisions (Deno, 1985; L. Fuchs & Fuchs, 1986; Shinn & Marston, 1985). Following is a discussion of curriculum-based

measurement that highlights its history and distinguishing factors that separate it from the other three CBA models.

Curriculum-Based Measurement

Initial CBM research began in the late 1970's and early 1980's as Deno and colleagues examined various measurement procedures that could produce reliable and valid measurements of reading, (Deno, Mirkin, & Chiang, 1982; Deno, Mirkin, Chiang, & Lowry, 1980), written expression (Deno, Mirkin, & Marston, 1980), spelling (Deno, Mirkin, Lowry, & Kuehnle, 1980), and mathematics. These studies produced a series of single global tasks that had potential to provide a valid and reliable measurement of academic skills across repeated administrations, later identified as one of the hallmarks of CBM (Deno, 1985). Tasks used to measure reading included reading words in isolation, in context and from passages, in addition to various other strategies (e.g., cloze procedures to assess comprehension and vocabulary assessment). Other academic skills were also considered, including written expression and spelling. Written expression was operationalized by writing a passage under time constraints and counting the number of words written, number of words spelled correctly, number of correct letter sequences, and the use of mature words. Spelling was defined as the number of correct letter sequences or total number of correctly spelled words.

This early research provided the foundation for the CBM formative measurement model, an “approach that uses direct observation and recording of student’s performance in the local school curriculum as a basis for gathering information to make instructional decisions” (Deno, 1987). CBM consists of “a set of short duration (1 to 3 min) fluency

tests in the basic skill areas of reading, spelling, mathematics, computation, and written expression, used in a standardized manner to facilitate problem-solving assessment” (Shinn & Hubbard, 1992). Another characteristic of CBM is the strong focus on obtaining evidence to determine acceptable reliability and validity.

In general, CBM uses sensitive cost-effective, brief measures that can be administered repeatedly over time (Deno, 1985). The advantage of using CBM over other CBA methods is the standardized administration procedures, localized normative data, and research establishing the reliability and validity of the instrumentation.

Responsiveness-to-Intervention and Curriculum-Based Measurement

CBM is an integral part of the new RTI framework endorsed by recent legislation as a means of identifying students with learning disabilities. RTI is a system used to evaluate students on a regular basis to identify (D. Fuchs & Fuchs, 2006) those performing significantly below their peers on curriculum-based measures. Those students then move through two to four tiers (Ardoyn, Witt, Connell, & Koenig, 2005; Center; D. Fuchs, Mock, Morgan, & Young, 2003; Jankowski, 2003) of increasingly intense instruction, based on RTI (Hollenbeck, 2007).

Typically, CBM is used within a RTI system to identify struggling students through universal screenings (Ardoyn et al., 2004) and student progress monitoring (Lynn S. Fuchs & Fuchs, 1999; Lynn S. Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993; Hosp & Hosp, 2003; Stecker & Fuchs, 2000) based on local normative data. The typical RTI model considers all students part of tier 1 wherein they receive a prescribed number of minutes per day of adequate research-based instruction (Vaughn, 2003), such as

Cognitively Guided Instruction (CGI; Carpenter, Fennema, Franke, Levi, & Empson, 1999).

Because CGI is an example of a prototypical intervention to improve math fluency it will be described briefly. CGI is a research-based instructional practice built on cognitive-processing theory. Cognitive learning focuses primarily on developing efficient problem-solving strategies. CGI helps instructors move students through the milestones of mathematical cognitive development using strategies such as direct modeling, counting, and number facts as described by Carpenter et al. (1999). Direct modeling involves manipulation of physical objects to represent a math problem. Students then move to using a counting strategy, wherein the student no longer needs physical objects to represent the relationship between the numbers. Finally, the student moves onto learning number facts. Number facts represent the student's ability to recall from memory the answer to common math problems (e.g., $3 + 4$). These strategies are not exclusive to one another but may be used simultaneously to arrive at a solution. The ability to use these strategies efficiently and move to strategy automation represents progress in early mathematic understanding. The strategies match with the National Council of Teachers of Mathematics' (NCTM) suggestion that students reach a desired level of procedural fluency within the curriculum while increasing fluency via the use of efficient and accurate strategies. M-CBM is a means of incrementally measuring skill development and provides an estimate of each student's mathematics understanding and fluency.

Three times per year, typically once at the beginning, middle, and end of the school year, school personnel administer a universal screening using a CBM for reading,

math, and/or writing. Students who score at a designated percentile or lower (e.g., 10th percentile or below) based on the universal screening are placed in tier 2 where they receive additional instruction (e.g., 30 min per day) in reading, math, or writing.

Typically tier 2 of RTI consists of biweekly progress monitoring, small group instruction, and increased intervention intensity (Hollenbeck, 2007). Increased intensity refers to more frequent and lengthier instruction in smaller groups, with more experienced instructors (D. Fuchs & Fuchs, 2006). Each student's rate of progress is tracked and weekly or bi-weekly improvement recorded on a graph. The slope of the student's CBM scores indicates rate of progress. The slope is typically calculated using a least-squares regression (L. Fuchs et al., 2007). Least-squares regression produces a line-of-best-fit through the student's data points, producing a visual display of each student's performance.

After approximately 10 weeks of intervention at tier 2, each student's progress (i.e., slope) and performance (i.e., universal screening score) are used to determine possible movement either back to tier 1 or into tier 3. Students who demonstrate progress greater than those at given percentile (e.g., 25th percentile) and performance greater than those at a given percentile (e.g., 10th percentile) move back into tier 1. Students who demonstrate progress lower than those at a given percentile and performance lower than those at a given percentile move into tier 3 for more intense instruction. Tier 3 typically consists of 60 min per day of instruction on top of the 90 min per day of typical classroom instruction. Tier 3 also has a lower student-teacher ratio (e.g., 2:1, 1:1) and may utilize a "uniquely prescribed" problem-solving model (PSM), where tier 2 typically

uses a “tier-wide” standard-treatment-protocol model (STPM), both commonly used intervention methods in RTI.

The PSM requires the personalization of assessments and interventions to meet the needs of each student (Jankowski, 2003; Marston, Muyskens, Lau, & Canter, 2003). One promising PSM is the Heartland PSM that includes four levels (Jankowski, 2003). Level 1 consists of consultation between teachers and parents. This is an informal stage of intervention where the teacher informally monitors progress. Level 2 includes a teacher consultation team who provides information and support to teachers who are working with struggling students. During level 2, written instructional plans are developed and progress monitored. Consideration for tracking into level 3 occurs if adequate progress is not achieved. Level 3 includes extended support staff such as school psychologists, educational consultants, or speech/language pathologists. Level 3 provides the most powerful interventions and is the heart of the problem-solving model, including a formal intervention plan and ongoing progress monitoring.

The model defines the problem (Problem Identification) operationally, using specific behavioral descriptions. CBM then provides validation of the existence of the problem. Following, data are used to develop an intervention plan to correct the problem (Problem Analysis). Implementation of the plan follows while assuring treatment integrity (Plan Implementation). The final step in the PSM (Problem Evaluation) requires analysis of the success of the plan through ongoing CBM data collection. Data provide guidance for intervention modifications as needed.

These four problem-solving steps occur throughout the RTI process as students move from one tier to the next. The benefits of a problem solving RTI model is that each student is given an individualized intervention plan designed to meet each specific need as it arises, although the model is likely to identify more false positives than a STPM (D. Fuchs & Fuchs, 2006).

Level 4 of the Heartland model allows consideration for special education services. When considering an evaluation for special education services, data from the prior three levels of intervention and other assessment data are considered. In order to be placed in special education, using the Heartland model, a student must demonstrate failure to respond to multiple levels of intervention, a significantly lower level of performance than peers, and the student must require instruction that cannot be provided in the general education curriculum (Jankowski, 2003).

In contrast, the STPM provides all students with the same intervention at set intervals and durations within small groups at tier 2 and individually at tier 3 (D. Fuchs & Fuchs, 2006). Tier 3 remains very similar to tier 2, although the intensity of the interventions increase and may be more customized. Students move through the tiers using the same decision criteria as the PSM although the interventions are the same for all students. Decisions pertaining to eligibility for special education can occur in tier 3 for students who fail to make adequate progress. Depending on the system and the particular process employed to determine special education eligibility, cognitive and adaptive measures may be used to rule out intellectual disability and/or to identify cognitive processing strengths and weaknesses.

Mathematics Curriculum-Based Measurement

There is a rapidly developing literature focusing on the use of CBM for reading and writing; on the other hand, there has been little attention paid to the development of M-CBM. Although existing research describes a variety of probe types for measuring mathematics performance across grade levels, with the exception of Fuchs and Fuchs' (1999) *Monitoring Basics Skills Progress: Concepts and Applications* (a commercially available M-CBM), the breadth of M-CBM research has focused primarily on brief single tasks. *Monitoring Basic Skills Progress: Concepts and Applications* provides both screening and monitoring progress functions and includes a variety of probe types but is not time limited.

An in-depth analysis of the psychometric properties of M-CBM follows. Due to the number of differing item types used across and within studies, the analysis is broken into sections discussing the psychometric properties of each item type present in the M-CBM literature. The item types are arranged in order from most prevalent to least prevalent including Math Facts, Computation, Discrimination tasks, Missing Number, Counting, Number Identification, Concepts and Applications, and Problem Solving. First, there is a brief discussion of traditional methods for establishing psychometric properties (i.e., reliability and validity).

Traditional Estimates of Reliability

Establishing and compartmentalizing estimates of true score and error variance is at the heart of classical test theory. In classical test theory, a "true score" is a hypothetical score minus sources of error. An obtained score is the score an examinee gets on any

given administration of a test, and consists of systematic or true score variance and error variance. A combination of increased content mastery and measurement error results in variation in obtained scores over multiple CBM administrations. In classical test theory measurement error is assumed to be normally distributed; thus over the course of an infinite number of administrations an individual's true score would lie at the center of the normal distribution of obtained scores (Nunnally, 1978). Varieties of procedures are used to obtain estimates of reliability because it is impossible to administer a test an infinite number of times. Nunnally (1978) provided a widely accepted definition of reliability, i.e., the extent to which a construct can be measured over difference occasions with repeatable results. Particularly, a measure is said to be reliable if it is able to produce the same results time after time. The degree to which obtained scores vary, represents error, assuming there is no change in mastery. Error can occur in either a random or a systematic fashion. Classical test theory focuses primarily on random error because systematic error contributes to all examinees mean score and is least important when studying individual differences. A reliable instrument is one that can be generalized to future measurements of the same construct. Error is typically defined as one minus reliability ($E = 1 - \text{rel}$).

There are four estimates of reliability typically used in classical test theory: alternate-form, internal-consistency, inter-rater, and test-retest reliability. All these estimates are operationalized by a correlation coefficient or some variant, typically the Pearson product moment correlation coefficient that ranges between -1 and +1. Alternate-form reliability is a measure of the degree to which two parallel forms of a test are

correlated with one another. Alternate-form reliability is a useful measure of reliability if the construct being assessed produces considerable variability over short periods of time (Nunnally, 1978). Internal consistency is a reliability estimate based “on the average correlation among items within a test (Nunnally, 1978, p. 229). Coefficient alpha is the typical measure of internal consistency, which shows how much the items have in common and sets the upper limit of reliability. Inter-rater reliability is an estimate of reliability based on the correlation or percent agreement of multiple raters across multiple testing forms. Low inter-rater reliability can contribute to testing error. Finally, test-retest reliability refers to the degree to which a single test form correlates with itself within a short period-of-time. Low correlations would indicate a high degree of variability or error in the instrument where high correlations would indicate low levels of variability and a more reliable instrument.

Previous M-CBM research has relied primarily on classical test theory to establish reliability. Alternate-form reliability estimates have been used most often followed by inter-rater reliability. Internal consistency and test-retest reliabilities have also been reported, although not as frequently. It is possible, using traditional techniques, to create an estimate of error and to use it to produce a confidence band using the standard error of measure (SEM), for any distribution of scores once reliability is determined. Calculating the SEM consists of multiplying the standard deviation of the distribution by the square root of one minus the reliability ($SEM = SD\sqrt{1 - rel}$). The SEM is rarely reported in the current M-CBM literature, and the problem is compounded because multiple measures

are typically used in curriculum-based assessment to monitor progress, multiplying the error associated with decision-making.

Generalizability Theory

An alternate strategy for determining test error relies on Generalizability theory. “Generalizability [G] theory is a statistical theory about the dependability of behavioral measurements” (Shavelson & Webb, 1991, p. 1), that incorporates both reliability and validity (Silva, 1993). The Standards for Educational and Psychological Testing (American Educational Research, 2002) recommend the use of G-Theory for determining error variances that arise from multiple sources, when feasible. G-Theory allows “an investigator to identify and quantify the sources of inconsistencies in observed scores that arise, or could arise, over replications of a measurement procedure” (Brennan, 2001, p. 2). In other words, in contrast to classical test theory that is “decomposed into a ‘true’ score T and a single undifferentiated error term E , G-Theory differentiate between multiple sources of error where classical test theory cannot (Brennan, 2001). Additionally, G-Theory can guide instrument development in such a way as to maximize its reliability (Cardinet, Johnson, & Pini, 2010).

G-Theory can be defined in the context of two types of studies, a generalizability study (G-Study) and a decision study (D-Study). The purpose of a G-Study is to “anticipate the multiple uses of a measurement (Shavelson & Webb, 1991, p. 12) and “obtain estimates of variance components associated with a universe of admissible observations” (Brennan, 2001, p. 8). A universe of admissible observations is considered “the observations encompassed by the possible combinations of conditions that the G-

Study represents (L. J. Cronbach, Gleser, Nanda, & Rajaratnam, 1972, p. 20). In other words, the universe of admissible observations consists of all combinations of possible observations as defined by the researcher. In order to estimate the variance components the G-Study requires two or more scores for each person obtained across one or more conditions (L. J. Cronbach, et al., 1972). A Universe Score, similar to a “true” score in classical test theory, is obtained for each person. A Universe Score is the expected mean score across all acceptable observations (Brennan, 2001; L. J. Cronbach, et al., 1972). The variance associated with the Universe Scores across all persons is called the Universe Score Variance (Brennan, 2001). A “factorial ANOVA [is then used] to partition an individual’s score into an effect for the universe-score (for the object of measurement), an effect for each facet or source of error, and an effect for each of their combinations” (Shavelson & Webb, 1991, p. 16).

The purpose of a D-Study is to use information from the G-study (i.e., variance components) to design the best instrument (Shavelson & Webb, 1991), make decisions and/or draw conclusions (Brennan, 2001; L. J. Cronbach, et al., 1972). The results of a D-Study are used to make relative or absolute decisions in addition to providing G-Coefficients, similar to Cronbach’s alpha in classical test theory.

Relative decisions are those decisions made about individuals when compared to the performance of others (Shavelson & Webb, 1991). Relative decisions result in a G-Coefficient defined as the ratio of the universe score variance to the expected observed score variance. The ratio produces scores ranging from 0 to 1 and indicates how likely the observation is to locate an individual relative to others in the population, for example,

how well the measurement has ranked the objects of measure (Cardinet, et al., 2010; L. J. Cronbach, et al., 1972). This is similar to classical test theory reliability coefficients (Mushquash & O'Connor, 2006).

Absolute decisions are used to classify individuals in some way based on performance standards or pre-determined criterion (L. J. Cronbach, et al., 1972). The G-Coefficient, called the Index of Dependability under an absolute decision, indicates the extent to which accurate generalizations can be made from an individual's observed score relative to his or her universe score (Shavelson & Webb, 1991). The Index of Dependability is a "ratio of the universe score variance to itself plus relative error variance" (Brennan, 2001, p. 13). This coefficient indicates how accurately the measure can locate an individual on a scale without respect to others (Cardinet, et al., 2010).

The information from the D-Study provides information to estimate the SEM and modify the measurement procedure in order to increase the reliability of the measure. Reliability may be increased by either modifying the number of facet levels sampled, eliminating atypical levels of certain facets, or changing the number and nature of the facets (Cardinet, et al., 2010). Any one or combination of these modifications, depending on the D-Study results, can create more reliable measurements in future studies.

Mathematics Curriculum-Based Measurement Item Types

Even at the early levels, math skills are complex and multi-faceted; thus, the assessment of math skills can also be complex and multi-faceted. There have been various M-CBM item types developed with some being more promising than others.

Therefore, it is important to take a closer look at the psychometric properties of specific item types. Below I describe some of those most commonly used.

Math facts. Deno and Mirkin's (1977) originally designed math probes were based on the assumption that mastering math facts, with addends to 19 and subtrahends to 9, by the end of third grade was essential for success in mathematics. Subsequently, multiple studies have examined the technical adequacy of math-fact scores and other M-CBM probes. Math facts consist of two types of tasks: single-skill and multi-skill probes. Single-skill probes refer to a set of items assessing one particular mathematical skill such as adding two numbers without carrying. A multi-skill probe includes item sets that assess more than one mathematical skill at once, such as two digit by two-digit addition with regrouping and three digit by three-digit addition without regrouping. A discussion of the reliability and validity demonstrated by these items follow.

Studies have reported the alternate-form reliability of single-skill math-fact probes ranging from .48 to .72 (Tindal, Germann, & Deno, 1983). Addition and subtraction math facts have resulted in higher reliabilities (.70 to .72) than multiplication and division facts, .48 to .61 (Tindal, et al., 1983) . Test-retest reliabilities have ranged from .78 to .89 (Epstein, Polloway, & Patton, 1989; Tindal, et al., 1983) across single-skill probes. Addition and subtraction math facts have resulted in similar reliabilities (.80 to .89) as multiplication and division math facts (.78 to .79). Inter-scorer agreements are typically high for math-facts items (.90 to 1.00; Epstein, et al., 1989; L. Fuchs, et al., 2007; VanDerHeyden, Witt, & Naquin, 2003).

A study by Connell (2006) examined the criterion related validity of single-skill math facts with the Woodcock-Johnson, Third Edition (WJ-III) broad math score (.45 to .70), and the Iowa Test of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2005) composite scores (.45 to .53). Multi-skill probes were also correlated with the WJ-III broad math score (.67 to .76) and the ITBS composite scores (.34 to .57).

Shinn & Martson (1985) conducted the first study examining the discriminative validity of math-fact probes. Findings revealed a significant mean difference between scores of students who were normally achieving, mildly handicapped, and those who were Title 1 eligible. VanDerHeyden et al., (2003) extended the findings of Shinn and Marston (1985) by not only examining the ability of math fact probes to differentiate predetermined groups but to identify how well math fact probes were able to accurately classify children when used as a screening measure. VanDerHeyden and colleagues found that specificities ranged from .69 to .99 and sensitivity ranged from .33 to 1.00, depending on the criterion measure used. Specificity refers to the proportion of false positives and sensitivity refers to the proportion of true positives with respect to the identification of at-risk students. These results indicate a high degree of variability with respect to the decision-making utility of math-fact probes.

Computation. Computation probes have also been the focus of much of the M-CBM research. Computation probes include addition, subtraction, multiplication, and division of multiple numbers of varying lengths. Estimates of reliability of computation probes have included estimates of internal consistency, alternate-form reliability, and generalizability.

Adding to the reliability research base, Fuchs et al., (1994) reported internal consistencies (.94 to .98) across second, third, and fourth graders. The mean digits-correct-per-minute was reported to range from 13.99 to 33.64 with the standard error of estimates from 4.04 to 5.81. Additional studies have reported the standard error of estimate of the slope ranging from .15 to .31 (Shapiro, Edwards, & Zigmond, 2005).

The generalizability of single and multiple-skill computation probes was studied by Hintze, Christ, and Keller (2002) and Christ, Johnson-Gros, and Hintze (2005). Hintze, Christ, and Keller (2002) found that a single-skill computation form could be generalized across administrations of all single-skill computation probe forms. That is, scores obtained on one single-skill computation form may be reliably obtained on subsequent single-skill computation forms. Results were slightly different when administering a multiple-skill probe. A single-skill probe required the administrations in order to generalize results across all other probe forms. This difference was due to the variance in scores associated with the individual decreasing while variance for other factors, such as grade level, increased when given multiple-skill probes. A subsequent study by Christ, Johnson-Gros, and Hintze (2005) sought to determine the preferable administration time needed when administering grade-level computation probes in order to obtain a reliable and generalizable scores. The authors determined that when making norm-referenced decisions based on the data, a 1 min administration is sufficient when making low-stakes decisions; however, 4 min was needed to make high-stakes decisions. When using criterion-referenced scores, 4 min was needed for low-stakes and 13 min for high-stakes decisions. The authors also examined the measurement error associated with

computation probes. They determined that 64% to 68% of the variability in scores was due to measurement error. Additionally, variance in the scores associated with the individual ranged from 17% to 26%. Apparently, a substantial amount of variance is associated with measurement error relative to variance within the individual. Thurber, Shinn, and Smolkowski (2002) reported alternate form reliabilities ranging from .90 to .92. AIMSWeb, a commercially available M-CBM, has been reported to have alternate-form reliabilities ranging from .90 to .98 across first, second, and third grades (National Center on Response to Intervention, 2010). The Discovery Education Predictive Assessment: Math (DEPA:M), a newly developed commercially available M-CBM, reports Cronbach's alpha reliabilities for grades 3 through 10 with third-grade ranging from .76 to .90 (National Center on Response to Intervention, 2010). The DEPA:M also reports SEMs that range from 2.11 to 2.69 at third grade.

Shinn & Marston (1985) reported discriminant ability of computation probes, finding a statistically significant difference between the performance of normally achieving, mildly handicapped and Title 1 eligible students. Apparently, computation probes can discriminate between students with differing levels of math ability. Computation probes have demonstrated low to moderate criterion validity. Connell (2006) reported moderate correlations between both the WJ-III mathematics broad score and the ITBS composite scores, ranging from .57 to .76 among first through third graders. The Problem Solving and Procedures of the Stanford Achievement Test: Ninth Edition have also demonstrated moderately strong correlations with computation probes, ranging from .50 to .66 (Jitendra, Sczesniak, & Deatline-Buchman, 2005). Computation probes

have demonstrated the least correlation with a group-administered achievement test, TerraNova (.45 to .51; Connell, 2006), and an individually administered achievement test, Wide Range Achievement Test Third edition (.34; L. Fuchs, et al., 2007).

Discrimination. Quantity discrimination (e.g., identification of the larger of two numbers, placed side-by-side) and other discrimination tasks have demonstrated good reliability and moderate to good validity (Chard et al., 2005; Clarke & Shinn, 2004; Vanderheyden et al., 2004). These tasks yield good alternate form (.92 to .93) and test-retest (.85 and .86) reliability. Clarke and Shinn (2004) reported correlations (.79 and .80) with the with Woodcock Johnson, Revised Edition (WJ-R; Woodcock & Johnson, 1989) Applied Problems subtest and the Number Knowledge Test, respectively. Chard et al., (2005) examined the same quantity discrimination task for 486 kindergarten and 483 first-grade students. They reported somewhat lower correlations between the Quantity Discrimination probes and the Number Knowledge Test for first-grade (.45 to .53) and kindergarten (.50 to .55) students. Although not quantity discrimination, VanDerHeyden et al., (2004) constructed a discrimination task wherein children had to choose one object from an array of four objects that did not belong. This task was reported to have good alternate-form reliability (.88) and moderate correlation with the Brigance Screens (.55; Brigance, 1985) and the Test of Early Mathematics Ability (.50; Ginsburg & Baroody, 1990).

VanDerHeyden (2004) administered a series of probes requiring the examiner to say the name of a shape and the examinee to point to that shape. This discrimination task

yielded moderate alternate form reliability of (.40) and poor criterion related validity (.06 to .38).

Missing number. Missing number task typically consists of a string of numbers presented with one or more numbers missing from the string. Students determine which numbers are missing, and write them in the blank. Clarke and Shinn (2004) and Chard et al., (2005) studied the use of number sequences containing numbers from 1 to 20. Chard et al., (2005) reported strong alternate-form (.78 to .93) and test-retest (.79 to .81) reliabilities. Correlations with the Number Knowledge Test ranged from .61 to .74 and with the WJ-R Edition (Woodcock & Johnson, 1989) Applied Problems Subtest from .68 to .69.

Counting. Oral counting tasks, assessed via M-CBM, have demonstrated acceptable reliability although there is conflicting information about the validity of the measures. A variety of counting tasks have been studied, including orally counting numbers and counting objects (Clarke & Shinn, 2004; Vanderheyden, et al., 2004). These tasks have produced strong alternate-form reliabilities (.71 to .93). Clark and Shinn (2004) reported test-retest reliabilities ranging from .78 to .80 and inter-scorer agreement (.99).

VanDerHeyden et al., (2004) reported moderate to low correlations between counting tasks and the Brigance Screens (.44 to .56; Brigance, 1985) and the Test of Early Mathematics Ability (.19 to .49; Ginsburg & Baroody, 1990). Correlations with the Number Knowledge Test (Okamoto & Case, 1996) have ranged from .38 to .55 at kindergarten and .07 to .70 at first grade (Chard, et al., 2005). Modest correlations have

also been found between oral counting and M-CBM computation probes (.50) and the WJ-R Applied Problems subtest (.60; Clarke & Shinn, 2004).

Number Identification. Number identification tasks typically require the child to name a given number. The child is required to choose a specific number, to name a presented number, circle a number, or write a number. These tasks have been researched among pre-k through first grade students, typically as a screening measure. Number identification tasks have demonstrated strong alternate-form validity (.81 to .93; Chard, et al., 2005; Clarke & Shinn, 2004; Vanderheyden, et al., 2004; VanDerHeyden, Witt, Naquin, & Noell, 2001), while test-retest reliabilities have ranged from .76 to .85 (Clarke & Shinn, 2004). Number identification tasks have also shown moderate to strong correlations with criterion measures ranging from .44 to .70 (Chard, et al., 2005; Clarke & Shinn, 2004; Vanderheyden, et al., 2004; VanDerHeyden, et al., 2001). VanDerHeyden and colleagues (2004; 2001) found somewhat higher correlations between teacher ranking of student performance and criterion measures (.30 to .91).

Two studies examined identification of shapes in combination with identification of numbers as a method of screening children for math difficulties. VanDerHeyden et al., (2001) combined counting shapes with number identification. Students were required to count the number of circles on the left side of the page and match the quantity with the numerical value on the right side of the page by circling the correct number. The total score was determined by recording the number of correct choices in 1 min. Alternate-form reliabilities for kindergarten and first-grade students have been strong (.84) while criterion related validity with the Comprehensive Inventory of Basic Skills, Revised

(Brigance, 1999) has been slightly lower (.61). An alternate task required students to look at a number and draw the number of circles equal to the value of the number. This task resulted in alternate-form reliability of .70 and criterion-related validity of .44

Fuchs et al., (2007) developed a number identification task that required students to fill in two numbers in an incomplete number sequence. These number identification tasks in early first grade demonstrated a significant relationship to end of second grade Wide-Range Achievement Test Arithmetic scores (.34).

Concepts and Applications. Concepts and Applications M-CBM items were developed by Fuchs and colleagues, who recommend a multi-dimensional format for progress monitoring, rather than a mono-operational format (L. Fuchs, et al., 1994). Concepts and Applications in contrast to traditional time-limited M-CBM, uses a time unlimited format. This is important because the Concepts and Applications probes consist of items requiring students to calculate measurements, time, and count money in addition to solving applied computation and word problems. Concepts and Applications probes have demonstrated strong internal consistency (.94 to .98) and good criterion validity with the Comprehensive Test of Basic Skills (.64 to .81; CTBS; L. Fuchs, et al., 1994; MacMillan & McGraw-Hill, 1989). Studies have also reported student improvement slopes from .40 to .69 (L. Fuchs, et al., 2007; L. Fuchs, et al., 1994).

Fuchs et al., (2007) advocated use of a range of mathematics skills when conducting screening and progress monitoring using M-CBM. Based on a Logistic regression analysis using number identification, fact retrieval, M-CBM computation, and M-CBM Concepts and Applications tasks, specificity (79.6, 75.3) was greater than

sensitivity (69.6, 70.8) in predicting both math disability calculation and word problem status with hit rates of 78.2% and 74.7%, respectively.

The area under the receiver operator characteristic (ROC) curve was .85 for calculation and .81 for word problems. The ROC curve is a visual representation depicting the ranking of individuals along the x and y axes on two independent indicators of group inclusion (e.g., at risk or not at risk). These results suggest that a single screener that incorporates multiple item types such as mathematics Concepts and Applications may be useful for screening children at risk for math disability. This study represents one of the only studies to examine the degree to which M-CBM can accurately identify at risk elementary students.

Problem solving. A final item type that has been explored for use as a M-CBM is problem-solving tasks. These are typically include word-problems and are used in higher grades. One study has examined their use among third-grade students. Jitendra and Szczesniak (2005) found internal consistencies of word problem solving range from .60 to .75, while criterion related validity ranged from .38 to .71. There is limited research pertaining to the use of problem-solving M-CBM among younger students.

Statement of Purpose

Within the RTI paradigm, it is typical to use a single M-CBM math-fluency measure (e.g., computation or math-facts) to screen and monitor the progress of students. These mono-operational math fluency measures have demonstrated adequate reliability and validity, although sizeable variability exists with regard to decision-making accuracy (L. Fuchs, et al., 2007; VanDerHeyden, et al., 2003) because of considerable

measurement error (Christ, et al., 2005). Of particular note, traditional computation-based M-CBM does not take into account problem-solving skills, especially at the earlier grades. Some research has examined problem-solving M-CBM at higher grades (Jitendra, et al., 2005), although co-morbid deficiencies in reading and math make it difficult to distinguish whether poor reading or math skills is the cause of poor performance on word-based problem-solving tasks. The M-CBM probes in this study were designed to be an indicator of both math fluency and mathematic problem solving. These novel probes are designed to be more ecologically valid than typical M-CBM measures because they consist of a multi-dimensional array of mathematical tasks. They combine elements of traditional M-CBM such as math-facts, computation and discrimination tasks, while adding less traditional tasks such as number sequences and shape patterns. The use of shape patterns has not been incorporated as an element of M-CBM, although the use of shapes has been investigated by VanDerHedyn, et al. (2001) in Pre-Kindergarten and Kindergarten classrooms. The probes are consistent with end-of-grade outcomes while maintaining consistency across forms in order to reduce measurement error. Finally, because the current format includes items requiring a variety of response types including production and nonproduction responses error may be reduced.

The MAP:M was developed to overcome the limitations of existing M-CBM measures and requires validation as a M-CBM instrument. This study was designed with two primary objectives in mind: a) to establish the reliability of the MAP:M, and relationship to the MBSP, using generalizability theory in addition to alternate-form and test-retest reliability and b) establish concurrent and predictive validity of the MAP:M in

relationship to the MBSP. Techniques from generalizability theory provide an overall estimate of reliability in addition to estimates of variability from a variety of sources including the examinee, the items (test scores), and residual error. Subsequently, the SEM of MAP:M and MBSP probes will be determined. A G-Study analysis is unusual and is available only in a limited number of M-CBM studies (Christ, et al., 2005; Hintze, et al., 2002).

To address the second objective scores obtained from the MAP:M will be compared to scores obtained from the Monitoring Basic Skills Progress: Computation (L. S. Fuchs, et al., 1999). The MBSP is the most researched and well-documented M-CBM. A high degree of correlation between the two measures will provide support for use of the MAP:M to measure math fluency. The degree to which the MAP:M and MBSP scores predict performance on a standard mathematics measure used by the school district, i.e., Star Math (Renaissance Learning, 2002), will also be conducted. Specifically this study explored the following questions:

Research Questions

1. What are the percentages of score variance accounted for by the following facets: person (p), items (i), and the residual (p x i, e) of the MAP:M, relative to a more traditional M-CBM, the MBSP?
2. What is the generalizability coefficient and SEM for relative decisions based on student's raw scores across one, two, three, five, seven, and ten probes for the MAP:M versus the MBSP?

3. What is the index of dependability and SEM for absolute decisions based on student's raw scores across one, two, three, five, seven, and ten probes for the MAP:M versus the MBSP?
4. What is the alternate-form reliability of the MAP:M versus the MBSP?
5. What is the test-retest reliability of the MAP:M?
6. What is the reliability of the slope of the MAP:M versus the MBSP?
7. What is the concurrent validity of the MAP:M with the MBSP as determined by a correlation coefficient expressing the relationship between the two?
8. What is the relative predictive power of the MAP:M and the MBSP when the criterion is STAR Math scores?

Chapter II

Method

The experimental mathematics curriculum-based measure (M-CBM) used in this study, Monitoring Academic Progress: Mathematics (Hopkins, et al., 2010), was developed by this author and Colleagues for use in a local school district over one year. The development of the MAP:M included completion of a qualitative pilot study, collaboration with school district administrators (i.e., Literacy Leaders, Mathematics Consultant, Principals, and Teachers) and completion of two quantitative pilot studies, carried out by the research team (this author, another advanced doctoral student in School Psychology, a professor in School psychology, and a professor in special education). The school district provided feedback and guidance throughout the process so that the MAP:M reflected the district's needs and best practice. Following is a description of the instrument development process.

Qualitative pilot study. The purpose of the first pilot study was to gather qualitative feedback from participants regarding the initial item content and structure of the MAP:M, in addition to collecting information about the instrument's face and content validity. In addition, the study was useful in helping the authors determine the number of items needed to conform to a 3 min administration. Qualitative feedback was solicited from examinees to guide further MAP:M development.

Participants and setting. The pilot study took place in a large room at the University of Tennessee and consisted of ten participants, six females and three males, four examinees in the third-grade, two in the second-grade, and four in the first-grade.

The examinees' parents, who were students at the University, were also included in the study and provided feedback, based on observations of their children's performance.

Instruments. This first version of the MAP:M was designed to be group administered and to measure a combination of math fluency and math problem-solving skills for first through third grade students. The MAP:M consisted of three components and administration relied on a script for all three: Guided Practice (directions and items), Independent Practice (directions and items), and scored items. Guided Practice included items representing each item type. The examiner read the items aloud while demonstrating item completion using a script in the Guided Practice section. Next, the participants completed the Independent Practice section. This section included a sample item representing each item type. After the examinees completed the items on their own, the examiner provided the correct answers.

The National Council of Teachers of Mathematics Curriculum Standards (National Council of Teachers of Mathematics, 2000), Tennessee State Curriculum Standards (Tennessee Department of Education, 2004) and the Saxon Math (Larson, 2004) curriculum were consulted to determine item type and difficulty for each grade level. These sources contain content goals for the curricula used in many schools systems in Tennessee and around the country, and are the curriculum used in the participant's district. The scored items included in the MAP:M were designed to assess end of grade-level skills and included three basic item types: quantity discrimination using words (QDW), quantity discrimination using symbols (QDS), and number sentence-quantity discrimination (NSQD). The test items consisted of four sets: early first-grade, later first-

grade, second grade, and third grade. There were two sets of first-grade test items to accommodate the substantial growth within the first-grade.

Quantity Discrimination using Words. QDW items consisted of randomly assigned numerical values placed to the left and right of three vertically arranged phrases: “is more than”, “is the same as”, “is less than.” The examinee was required to indicate the relationship of the number on the left to the number on the right by circling one of the three phrases. Words were used instead of symbols to allow early first grade examinees to perform the quantity discrimination task prior to being taught the quantity discrimination symbols. Eight QDW items were included in early first-grade probes. In order to ensure a distribution of values, randomly selected values ranging from 1 to 50 were assigned to each item. As a reminder, a legend was placed at the top of each page containing the words “is greater than”, “is equal to,” and “is less than”, paired with their respective symbols and circles of the respective size.

Quantity Discrimination using Symbols. The QDS items consisted of randomly assigned numerical values to the left and right of three vertically arranged quantity discrimination symbols (i.e., $<$, $>$, $=$). The examinee was required to indicate the relationship of the number on the left to the number on the right by circling one of the three symbols. The QDS items were included in early first-grade probes. In order to ensure a distribution of values, randomly selected values, ranging from 1 to 50, were assigned to each item.

Number Sentence-Quantity Discrimination. The NSQD task consists of horizontally presented number sentences, immediately followed by a quantity

discrimination task. The symbols (i.e., $<$, $>$, $=$) separated the number sentence answer from a randomly assigned number. Addition items included randomly selected addends while subtraction items included randomly selected minuends and subtrahends selected from values less than the minuend. Examinees were required to first write the answer to the number sentence, and then circle the symbol representing the relationship between the written answer and the random number. The NSQD items were included in all ending-first through third-grade probes.

Early first-grade probes included 14 NSQD items with addend, minuend, and subtrahend values ranging from 1 to 10. Items were equally divided between addition and subtraction. Quantity discrimination random numbers ranged from 1 to 10. Eight QDW items and six QDS items with random numbers from 1 to 50 were also included.

Late first-grade probes included 32 NSQD items with addend, minuend, and subtrahend values ranging from 1 to 10. Items were equally divided between addition and subtraction. Quantity discrimination random numbers ranged from 1 to 10.

Second-grade probes included 32 NSQD items with addend, minuend, and subtrahend values ranging from 1 to 20. Items were equally divided between addition and subtraction. Quantity discrimination random numbers ranged from 1 to 20.

Third-grade probes consisted of 22 NSQD items with addend, minuend, divisors, dividends, factors, and subtrahend values ranging from 1 to 20. There were also seven multiplication and three division items with dividend, divisor, multiplicand, and multiplier values ranging from 1 to 10. All division items resulted in whole-number quotient. Quantity discrimination random numbers ranged from 1 to 30.

Procedures. The participants were divided into same-grade groups. One graduate student, trained in the administration of the probes, was assigned to each group as the examiner. First, the examiner explained the purpose of the test, and then proceeded to administer each grade-level probe to the group. Each administration included three parts: guided practice, independent practice, and finally completion of the scored test items. The examiner read a set of scripted directions prior to each part. The participants were given 3 min to complete the scored items at each grade-level. After the administration, participants were encouraged to provide feedback about the clarity of the directions and difficulty of the items.

Results. All examinees demonstrated an ability to make progress on testing items. Examinees were able to complete all below level probe items prior to the end of 3 min but typically were unable to complete all above grade-level probe items before time expired. Thus, the length of the test and time allotted was determined to be appropriate. Qualitative feedback from the examinees indicated that the legend was not used during the test administration. Examinees in first-grade made a large number of errors on both quantity discrimination tasks (i.e., QDS & QDW), although successful progress was noted from some participants.

School district collaboration. After completing the first pilot study, some items were modified. A number of additional MAP:M modifications were made based on feedback from school personnel.

First school district meeting. The MAP:M was presented to eight literacy leaders for the participating school district for review. School district personnel found all item

types (i.e., QDW, QDS, and NSQD) to be acceptable but requested that items be included that assess counting in sequence and the use of geometric figures. The information from the school district was used to modify the MAP:M.

The requested modifications were made to the MAP:M and a second meeting was scheduled with school district personnel, including literacy leaders and principals. The modified MAP:M was presented to 30 school administrators at their district office. The purpose of the meeting was to gather qualitative feedback and recommendations regarding the appropriateness of item content, item format, and clarify of directions.

Second school district meeting. The MAP:M presented to the school administrators during the second meeting had undergone the following modifications. The MAP:M retained the original items but now included new Number Pattern and Shape Pattern items. The legend was removed from the first-grade probes based on examinee feedback during the qualitative pilot study. As a replacement, administration directions were modified to require the examiner to write the words and symbols on the Blackboard prior to administering the MAP:M. Early first-grade and late-first grade probes were combined into one first-grade probe to allow for growth estimates across the school year.

Following is a description of the new Number Pattern and Shape Pattern items. The Number Pattern (NP) items consist of five numbers, presented horizontally and ordered from least to greatest. Items were randomly assigned and arranged according to grade specific criteria. First incrementing values for each grade level were placed in a Microsoft[®] Excel[®] worksheet (e.g., count by 1's, 5's, and 10's). Next, the first number in the number sequence was assigned randomly with grade specific parameters. Finally, the

subsequent numbers in the five digit number sequence were assigned by having Excel count incrementally using the incrementing value. One (first grade) or two (second and third grade) numbers were left out of each sequence by having Excel[®] randomly assign a value between one and four, indicating the first value to be substituted for a blank space for an examinee response. The examinees are instructed to write the missing number(s) in the sequence.

The NP items were included in the first through third grade probes. First-grade probes included four NP items including counting by 1's, 2's, 5's, and 10's. Numerical values ranged from 1 to 99. The second-grade probes included six NP items: counting by, 2's, 3's, 4's, 5's, 25's, and 100's with numerical values ranging from 1 to 999. Third-grade probes included six NP items: counting by 1's, 2's, 4's, 5's, 10's, and 25's with values ranging from 1 to 999. The values, inclusion and presentation order of any given counting sequence was randomly selected.

The Shape Pattern (SP) task was designed to measure mathematical problem solving while utilizing grade-level geometric knowledge. The SP items require identification of one missing shape from a shape pattern of varying length. Shapes used included squares, triangles, circles, pentagons, and hexagons. The shape patterns were presented horizontally with four possible shape choices to the right. The placement of the missing shape and the correct response was determined by random assignment. The answer was indicated by circling a shape from the four-shape choices to the right that completed the pattern. First-grade SP items included simple pattern sequences. Second-grade SP items included more complex sequences and rotations, while third grade

incorporated flips and turns of the shapes in more complex sequences. All three grades include four SP items.

Third school district meeting. After the completion of the second school district meeting, the feedback and recommendation were incorporated into the MAP:M. A third meeting was held wherein the MAP:M was again administered to same group of school administrators in small groups, following the same format and sequence as the procedures in pilot study one. This version of the MAP:M included Math facts and Computation items in addition to modifications of the NSQD task. QDS, NP, and SP items were modified slightly as reflected in their descriptions below. Following the administration of the probes and open forum was provided for feedback and recommendations. Because of the meeting, minor changes in formatting were suggested in order to eliminate clutter on the page. It was also suggested that the QDW items be dropped from the first-grade items due to redundancy with QDS and NSQD items.

Specific details of the modified MAP:M is as follows. First-grade probes included 16 math facts items divided equally between addition and subtraction. Half of the addition items and half of the subtraction items were made up of random numbers from 1 to 10 and the other half of each ranged from 11 to 20.

QDS items were included in the first-grade probes and consisted of eight items with randomly assigned numerical values to the left and right of three vertically arranged quantity discrimination symbols (i.e., $<$, $>$, $=$). Using this format the examinee was to indicate the relationship of the number on the left to the number on the right by circling one of the three symbols. The QDS items were included in early first-grade probes. In

order to ensure a distribution of values, half the QDS items included random values from 1 to 50 in four items and the other half from 51 to 99. Instead of the legend at the top of the page, the directions required the examiner to write the symbols, accompanied by the words corresponding to the symbols, on the dry erase or chalkboard in front of the room.

The NSQD items were designed to combine elements of math-facts and quantity discrimination tasks in order to consolidate the two tasks into a single, more challenging mathematical task. The NSQD task consists of horizontally presented number sentences that were immediately followed by a quantity discrimination task. The symbols (i.e., $<$, $>$, $=$) separated the number sentence answer from a randomly assigned number. Addition items included randomly selected addends while subtraction items included randomly selected minuends and subtrahends selected from values less than the minuend. In this format, the examinee is required first to write the answer to the number sentence and then circle the symbol that represented the relationship between the written answer and the random number. The NSQD items were included in all first through third grade probes.

Second-grade probes included 20 NSQD items with addend, minuend, and subtrahend values ranging from 1 to 999. Items were equally divided between addition and subtraction. Four addition and four subtraction items included digits ranging from 1 to 18. Two addition and two subtraction items included numbers from 10 to 99. Finally, two addition and two subtraction items included numbers from 100 to 999. Quantity discrimination random numbers ranged from 1 to 999.

Third-grade probes had 20 NSQD total items with addend, minuend, divisors, dividends, factors, and subtrahend values ranging from 1 to 999. Two addition, two

subtraction, two multiplication, and two division item ranged from 1 to 10. Two addition and two subtraction ranged from 10 to 99. Two addition and two subtraction ranged from 100 to 999. There were also seven multiplication and three division items with dividend, divisor, and Quantity Discrimination random numbers ranged from 1 to 999.

The Number Pattern (NP) items consisted of five numbers, presented horizontally and ordered from least to greatest. Items were randomly assigned according to grade specific criteria. First incrementing values for each grade level were placed in an excel worksheet (i.e., count by 1's, 5's, 10's). Next, the first number in the number sequence was assigned randomly with grade specific parameters. Finally, the subsequent numbers in the five digit number sequence were assigned by having Excel count incrementally using the incrementing value. Two numbers were left out of each sequence by having Excel randomly assign a value between one and four, indicating the first value to be substituted for a blank space for an examinee response. After all NP items for each probe were generated, they were randomly assigned for inclusion in each grade-level probe. The examinees were required to write the missing numbers in the sequence. One point was given for each correct response with two possible points per item.

The NP items were used included in first through third grade probes. First-grade probes included four NP items including counting by 1's, 2's, and 5's. Numerical values ranged from 1 to 99. The second-grade probes included eight NP items: counting by 2's, 3's, 4's, 5's, 10's, 25's, and 100's with numerical values ranging from 1 to 999. Third-grade probes included six NP items: counting by 1's, 2's, 4's, 5's, 10's, and 25's with

values ranging from 1 to 999. The values, inclusion and presentation order of any given counting sequence was randomly selected.

Research describing the use of shapes in M-CBM is limited. VanDerHeyden and colleges (2004; 2001) used shapes in sample of pre-k and kindergarten examinees to screen for school readiness. These tasks required the choosing and drawing of shapes. In addition, limited research has been conducted concerning problem solving in the context of M-CBM. The existing problem-solving M-CBM research is limited to the use of language and word-based items. The SP task in the current study is designed to measure the examinee's mathematical problem-solving ability while utilizing grade-level shape knowledge. The SP items required examinees to identify one missing shape from a nine-shape pattern. Sequences and shapes to be used in the SP items were determined prior to creating the probes. Each grade specific set of sequences and shape combinations were placed in an Excel spreadsheet and randomly selected for inclusion in each form. Shapes included squares, triangles, circles, pentagons, stars, and hexagons. The shape patterns were presented horizontally with four possible shape choices to the right. The placement of the missing shape and the correct response was determined by random assignment. The examinee was required to circle the shape that completed the pattern. First-grade SP items included simple pattern sequences. Second-grade SP items included more complex sequences and rotations, while third grade incorporated flips and turns of the shapes in more complex sequences.

Fourth school district meeting. A fourth meeting of school system administrators was held to consider the revised MAP:M. Following the standard administration of the

probes, an open forum was again provided for feedback and recommendations.

Administrators suggested some alternate administration direction wording such as “do some math” instead of “do some problems.” The administrators suggested that horizontally presented computation items that exceed one digit by one digit be presented in a vertical orientation. Additionally, it was recommended that graphics be added as reference points for examinees. With these changes, the district approved the items and format for administration in a pilot study to include two district schools.

Quantitative Pilot Study One. The primary purpose of this pilot study was to obtain data to inform item selection (e.g., to provide difficulty and discrimination indices). A secondary purpose was to determine differences in examinee performance within and across classrooms.

Participants and setting. The revised MAP:M was administered to 181 examinees in two schools. There were 53 (29%) first-grade, 59 (33%) second-grade, and 69 (38%) third grade. The study was conducted in two phases. In phase one, the examinees of one school were provided unlimited time to complete the probes. Phase two limited administration time to 1.5 min.

The time-unlimited group was used to conduct an item-analysis to determine item difficulty across the probes. The time-unlimited group included 86 participants, 26 (30%) first-grade, 29 (34%) second-grade, and 31 (36%) third-grade examinees. The time-limited group was designed to provide an estimate of examinee performance under normal testing conditions. The time-limited group included 95 participants, 27 (28%) first-grade, 30 (32%) second-grade, and 31 (40%) third-grade examinees. The examinees

were administered the probes in their typical classroom, by the researcher or a trained graduate student, at a time designated by the classroom teacher.

Instruments. Based on the changes suggested by the latest administrative the following modifications were made to the MAP:M. The probes maintained the quantity discrimination, shape pattern, number sequence, and math fact items. As recommended by the administrators, Computation items were added, wherein addition and subtraction items exceeding one digit by one digit were presented vertically. The Computation (COMP) items required examinees to solve 2×2 or 3×3 addition and subtraction items. The COMP items did not require regrouping. Values for the 2×2 items ranged from 12 to 99, with any addition by one eliminated from second and third-grade items due to low difficulty and discrimination indices. Examinees received credit for each digit written correctly.

Eight 2×2 COMP items were included in first-grade probes. Eight COMP items were included in second and third-grade probes, divided evenly between 2×2 and 3×3 addition and subtraction items. The number sentence-quantity discrimination items were modified so that they included only math-fact items with randomly assigned quantity discrimination numbers ranging from one to nine. Pictures a fish and a cow were also added at the top of each column of items as a reference point for examinees during the presentation of directions and practice items.

First-grade probes included two COMP items, six MF, two NP, four NSQD, and two SP items. Second-grade probes included four COMP and six NSQD, four NP, and

two SP items. Third-grade probes contained six COMP, four NSQD, four NS, and two SP items.

The practice exercise was modified in the following way based on school personnel feedback. The guided and independent practices were collapsed into a single graduated learning task for each grade-level item type. The first item provided the examinee with the answer that they were asked to trace. The second item solicited examinee participation followed by the examiner immediately providing the correct answer. The third item was completed by each examinee independently in order to demonstrate understanding of the task demands.

Procedures. The following procedures were followed in both a time-limited and time-unlimited groups and both formats included two classrooms per grade level. This author and another graduate student served as examiners while being observed by designated school personnel. Prior to each administration, the probe examiner was introduced to the examinees by school staff. The practice items were distributed to the examinees. The examiner read directly from the scripted instructions designed enable the guided practice. Additional assistance was provided as needed to examinees that who experienced difficulty understanding the directions. The goal of the practice was to ensure examinees understood the nature of the task. Once all practice items were completed, the examiner collected the practice sheets and distributed the probes. The examiner then read from the scripted instructions and instructed the examinees to begin working. At the end of the time limit, or when everyone had finished all of the items, the examiner collected the probes.

Results. Descriptive statistics were calculated for the time-unlimited group based on total scores. First-grade probe scores ranged from 0 to 20, second-grade from 0 to 47, and third-grade from 0 to 49. Means and standard deviations were obtained for first-grade ($M = 8.69$; $SD = 4.04$), second-grade ($M = 28.26$; $SD = 6.90$), and third-grade ($M = 41.49$, $SD = 4.28$) probes. An independent t test was also conducted between each grade level probe to determine if independent administrations resulted in similar scores. There were no significant differences found between grade-level probe total scores at first-grade ($t[24] = -0.03$, $p > .05$), second-grade ($t[29] = -1.05$, $p > .05$), and third-grade ($t[29] = -0.14$, $p > .05$). Results are presented in Table 1.

An item analysis was conducted in order to determine item difficulty and item discrimination indices for each grade level. Results are presented in Tables 2, 3, and 4. The item difficulty index represents the proportion of examinees who received full credit for the item. Difficulty index scores range from 0.0 to 1.0, with items becoming more difficult as they approach 1.0 (Sattler, 2008). Item discrimination index refers to an item's ability to distinguish between examinees that are high or low achieving. The discrimination index is calculated by dividing the sample of scores into thirds based on each participant's total score. The proportion of examinees in the bottom third who answered an item correctly is subtracted from the proportion of examinees in the top third who also answered the item correctly. The difference between these proportions represents the discrimination index. In other words, the discrimination index represents the difference between the difficulty index for the top-third and the bottom-third. Discrimination indices range from -1.0 to +1.0 and values between -0.2 and +0.2 are

considered to poorly discriminate between examinees with high or low overall scores (Sattler, 2008). It is important to note that in the current study, participants have an opportunity to receive partial credit on any given item. The difficulty and discrimination indices are based on items for which examinees received full credit.

First-grade probes yielded difficulties ranging from $-.11$ to $.85$. Computation items were the most difficult (i.e., difficulty of $.04$). This outcome was expected because the items were written to reflect end-of-grade performance, and beginning first-grade examinees are learning math facts and have not yet begun to learn computation that is more complex. First-grade difficulty indices for math fact items ranged from $.22$ to $.78$, with subtraction facts ($.22$ to $.37$) being the most difficult and addition the least difficult ($.30$ to $.78$). Number sequences produced a range of difficulty. Counting by 1's items were easy (difficulty = $.85$) while counting by 2's was hard (difficulty = $.33$). Quantity discrimination using symbols proved to be moderately difficult, with indices ranging from $.33$ to $.59$. Finally, first-grade examinees found Shape Pattern items easy as evidenced by difficulty indices from $.74$ to $.85$.

First-grade probes, with the exception of the two computation items, demonstrated good item discrimination across all items ranging from $.33$ to $.78$. The two computation probes yielded discrimination indices of $.11$ for addition and $-.11$ for subtraction with only one participant receiving full credit on either item. Again, this outcome is most likely the result of the examinees not yet being exposed to more difficult computation items, though consistent with the goal of including end-of-year items.

Second-grade probe item difficulties ranged from .23 to .97. Computation items ranged from .42 to .94, while number sequence items ranged from .00 to .97. Participants found counting by 5's very easy (.97) where counting by 25's resulted in no full credit responses (.00), although 15 of 31 participants received partial credit. The number-sentence quantity-discrimination items resulted in discrimination indices from .23 to .71 and Shape patterns ranged from .74 to .81.

Second-grade probes demonstrated a range of discrimination indices ranging from .00 to .60. Computation items ranged from .10 to .60. Computation items that required addition or subtraction by 10's or 100's resulted in poor discrimination (.10 to .20) where other items resulted in good discrimination (.60). Number sequence items ranged from .00 to .60, with counting by 5's, resulting in .00 discrimination; all participants received full credit and counting by 25's resulting in .00 discrimination; no participants receiving full credit. Again, almost half the participants received some credit on items requiring counting by 25's. Number sentence-quantity discrimination items resulted in indices ranging from .20 to .60, with addition by 1's or series of 1's resulting in poor discrimination (.20 to .30). Shape patterns resulted in overall poor discrimination (.30). As a result, future probes included more difficult shape patterns.

Third-grade probe item difficulties ranged from .06 to .97. Computation items ranged from .63 to .94 and number sequence items ranged from .47 to .97. Number-sentence quantity-discrimination items ranged from .06 to .88. The most difficult item required division. This is expected because division is not taught early in the academic

year. The easiest item was an item with addition by one. Addition and subtraction by one was not included in future probes at the third grade level due to lack of difficulty.

Third-grade probes had discrimination indices from $-.10$ to $.68$. Computation probe discrimination ranged from $.11$ to $.58$ with addition by one items demonstrating the least discrimination. Overall, lower discrimination indices were a result of few errors in either the lower or the upper groups. Number sequences resulted in discrimination indices from $.00$ to $.68$ with counting by 10's and 1's providing no discrimination. Number-sentence quantity-discrimination items resulted in discrimination indices from $-.01$ to $.47$ where the negative discrimination resulted from one lower performing examinee answering an element of the division item correctly. The Shape Pattern items resulted in poor discrimination ($-.10$ to $.22$) for all but two participants answering them correctly. As a result, future probes included Shape Patterns of increased difficulty.

Descriptive statistics were also calculated for the time-limited group based on each participant's total score. Score ranges were the same as the time-unlimited group. Means and standard deviations were calculated for first-grade ($M = 2.79$; $SD = 2.08$), second-grade ($M = 9.27$; $SD = 4.46$), and third-grade ($M = 15.29$; $SD = 5.77$) probes. An independent t test was performed on second and third-grade probes to determine if independent grade-level administrations resulted in similar total scores. First-grade included only one classroom, thus an analysis was not conducted. The independent-samples t test showed no significant differences between mean class total scores at second-grade, $t(28) = 1.09$, $p > .05$, or third-grade, $t(36) = -0.85$, $p > .05$. Data were not available for first grade. Results are presented in Table 5. Item difficulty and item

discrimination indices were not calculated for the time-limited group because not all items were administered.

Modifications based on findings. This pilot study provided information to inform modifications to the probes in addition to providing evidence of score stability across independent administrations. As a result, the following modifications were made to the probe content, as indicated. First-grade probes were not modified, although it was noted that quantity discrimination using symbols was a difficult task, but one that might be mastered at the end of grade. Second-grade probes no longer included addition or subtraction by 10 or multiples of 100 due to poor discrimination. Number Sequence items containing numbers that are multiples of the step value were eliminated. Items that require addition or subtraction by 1's or series of 1's were excluded in future probes. The difficulties of the Shape Patterns were increased to improve their discriminant ability. Finally, third-grade probes included the same modifications as the second-grade probes with the elimination of counting by ones or tens. These modifications were included in the first Universal Screening. Because the items need to assess end-of-year performance, traditional difficulty and discrimination criteria were used as guidelines only.

Quantitative Pilot Study Two. The second quantitative pilot study was carried out in order to ensure that clarity of instructions and sufficient length. A second objective was to obtain an estimate of reliability of the test items under normal testing conditions.

Participants and setting. The current study was conducted in all first through third-grade classrooms in the participating school district. The probes were administered by the regular classroom teacher in their classroom, under the supervision of designated

school personnel. Approximately 1,600 examinees were included in the pilot study. A representative data sample were collected from one classroom per grade per participating school, resulting in 335 participants in first-grade ($n = 114$), second-grade ($n = 99$), and third-grade ($n = 122$).

Instruments. This pilot study included finalized practice items and test probes. The practice items at each grade-level included three items from each item type including one already completed item, one guided-practice item, and one independent-practice item.

The number parameters of each item type was the same as quantitative pilot study one for comparison purposes. The format of each probe was modified so that each item type was presented in the same sequence on the front and back pages. First-grade test probes included eight quantity discrimination, computation, shape pattern, and number sequence items. Second and third-grade test probes included eight of each of the following items: number sentence-quantity discrimination, computation, Number Pattern, shape pattern.

Procedures.

Preliminary Activities. Prior to beginning the study, the school literacy leaders and administrators were trained by the researchers in a four-hour training seminar. These trained school personnel were then responsible for providing supervision and training to the classroom teachers. During training, the research team administered each grade-level probe to the school personnel. Next, the administrators were required to administer the

probes to a peer. The school administrators then scheduled an in-service for classroom teachers, who were provided direct instruction.

Distribution of Instruments. The test probes, practice exercises, and administration instructions were downloaded from the internet by designated school staff, and then printed. The instruments were then distributed to classroom teachers for administration on the designated testing day.

Administration Procedures. The MAP:M was administered by the classroom teacher in their usual classroom at a time designated by the teacher. The classroom teacher first distributed the MAP:M practice exercise to each examinee. The practice instructions were read verbatim from the guided-practice script. The classroom teachers were directed to ensure that each examinee was able to understand the directions and were responding appropriately on the practice exercise form. After all practice exercise items were completed, the practice exercise forms were collected.

The test probes were then distributed to each examinee. The examinees were instructed to write their first name, last name, and date on the top of their paper. Once each examinee had finished writing the required information, the classroom teacher read the scripted administration instructions. The examinees were instructed to begin working and provided exactly 3 min to complete as many items as possible. The classroom teachers were instructed not to assist examinees during the administration. If assistance was requested, they were to instruct the examinee to skip the items they did not know and complete the ones they did know. At the end of 3 min, the classroom teachers told the examinees to stop and put their pencils down and then collected all materials.

Results. Descriptive statistics, Cronbach's alphas, and the SEM were calculated at each grade level as an estimate of reliability. Results are presented in Table 6. First-grade ($M = 8.58$; $SD = 4.76$) and third-grade ($M = 25.62$; $SD = 9.75$) examinees yielded good reliabilities of $\alpha = .78$ and $\alpha = .73$, respectively. The SEM was ± 2.23 and ± 5.07 for first and third-grade probes, respectively. The second-grade probes ($M = 13.81$; $SD = 7.31$) yielded lower reliability ($\alpha = .67$) and a SEM of ± 4.20 . As expected, means increased across grades.

An independent t test was conducted to compare the average item difficulty and item discrimination of quantitative pilot two and three in order to test form equivalency. No significant differences were found between either item difficulty, $t(62) = -.85$, $p > .05$, or item discrimination, $t(62) = -.73$, $p > .05$, indices at first-grade. Second grade item difficulty means were not significantly different, $t(62) = .07$, $p > .05$, nor were item discrimination means, $t(62) = -.64$, $p > .05$. The same was true for third-grade probes with no differences for item difficulty, $t(62) = -.15$, $p > .05$, or item discrimination, $t(62) = -1.10$, $p > .05$. Discrimination and difficulty indices tables are in Tables 7, 8, and 9.

Summary. The psychometric properties and score distributions were determined to be acceptable by the Research Team and the local school system personnel. These classical item analyses are not typically calculated for M-CBM, although they provide important information about the range of item difficulty and discrimination. Both the difficulty and discrimination indices support overall adequate difficulty and item discrimination.

Current Study

Participants and setting. The participating school district serves eight rural elementary schools, consisting of 1645 examinees.. The examinees include 895 (54%) males, and 749 (46%) females. There were 1546 (94%) White, 69 (4%) African American, 6 (<1%) Asian Pacific Islander, 18 (1%) Hispanic, and 2 (<1%) Native American/Alaskan participants. Participants were in first (n = 525), second (n = 555), and third grades (n = 565). Demographic information is presented in Table 10.

Permission to conduct the study was obtained from the school district Superintendent, building Principals, and the University of Tennessee Institutional Review Board. Probes were administered approximately bi-weekly and continued across 11 administrations. All MAP:M probes were administered in the examinee's classrooms by the classroom teacher. The MAP:M probes were then scored and logged in the school district's database by designated school personnel. The researcher was granted access to database reports that were cleaned of all personally identifiable examinee information.

Instruments. The experimental M-CBM progress monitoring measure, the MAP:M, is the focus of this study. The Monitoring Basic Skills Progress: Computation (MBSP) measures were used to establish concurrent validity. STAR math scores were used to establish the predictive validity of the MAP:M.

Monitoring Basic Skills Progress. The MBSP measures include standardized math computation (MC), and Concepts and Applications (CA) items with 30 alternate forms per grade level. The current study included only the MC forms. MC is a time-limited assessment, allowing examinees 2 to 3 min to complete as many grade level

computation items as possible. Total MC Score is the number of correct digits. Fuchs et al., (1999) report test-retest reliability for examinees with disabilities (.73 to .91) and those without (.73 to .84), across grades one through six.

Criterion validity of MC was established through correlation with the Math Computation Test (MCT; Lynn S. Fuchs, Fuchs, Hamlett, & Stecker, 1991) and the Stanford Achievement Test (SAT; Gardner, Rudman, Karlsen, & Merwin, 1982). Correlations with the MCT ranged from .66 to .91 across grades 2 through 5 while correlations with the Stanford Achievement Test (SAT) ranged from .49 to .93.

STAR Math. STAR Math (Renaissance Learning, 2002) is a computer-adaptive, individually or group administered, test of mathematics achievement for students in first through twelfth grades. STAR Math requires approximately 10 min to complete and provides norm-referenced, grade equivalent, and percentile rank scores that reflect overall performance on mathematics computation and Concepts and Applications. A composite Normal Curve Equivalence (NCE) score that has a mean of 50 and standard deviation of 21.06 represents overall math performance in the current study. STAR math has demonstrated adequate overall alternate-form (.91) and split-half (.94) reliability in addition to concurrent validity ranging from .40 to .91 (Renaissance Learning, 2009).

Monitoring Academic Progress: Mathematics. The MAP:M is a group-administered test of math fluency and item solving and the final form, as described here, resulted from the various pilot studies reported previously. In its final form the MAP:M requires 3 min, and consists of 11 alternate forms for first through third grade. All forms reflect learning objectives outlined by the National Council of Teachers of Mathematics

Curriculum Standards (National Council of Teachers of Mathematics, 2000) and the State of Tennessee's scope and sequence of mathematics (U.S. Department of Education, 2004) at each grade level. The MAP:M is made up of guided/independent practice and a variety of item types including QD, NP, SP, COMP, MF, and NSQD.

Guided/Independent Practice. The guided/independent practice is a graduated learning task designed to teach the expected response format and task demand of each item type. The first item of each item type provided the examinee with the answer. The examinee was then instructed to trace the answer. The second was designed to allow the examinee to demonstrate understanding of the task demands. The answers to each independently completed item were supplied after the examinees had a chance to complete them. These items have been described previously, and will be described only briefly below and examples presented in Appendix A.

Quantity Discrimination. Quantity discrimination and other discrimination tasks have demonstrated good reliability and moderate to good validity (Chard, et al., 2005; Clarke & Shinn, 2004; Vanderheyden, et al., 2004). The QD items consisted of three horizontally arranged numbers. The examinees were required to circle the largest number in a three-number sequence. Responses were counted as correct and one point earned if the examinee circled the correct number. The QD items were included in first-grade probes and consisted of eight items. In order to ensure an even distribution of values, four items were randomly assigned values ranging from 1 to 50 and the values of the other four items ranged from 50 to 99.

Number Patterns. Counting and finding the missing number have demonstrated good reliability and validity (Chard, et al., 2005; Clarke & Shinn, 2004; Vanderheyden, et al., 2004). These tasks were considered when developing the Number Pattern items that consisted of five numbers, presented horizontally and ordered from least to greatest. Items were randomly assigned according to grade specific criteria. In order to randomly assign values to each item an incrementing value for each grade level was placed in an excel worksheet (i.e., count by 1's, 5's 10's). Next, the first number in the number sequence was assigned randomly with grade specific parameters. Finally, the subsequent numbers in the five digit number sequence were assigned by having Excel count incrementally using the incrementing value. One or two numbers were replaced with a blank space, depending on grade level, of each sequence by having Excel randomly assign a value between one and four. The examinees were required to write the missing numbers in the sequence. One point was given for each correct digit written.

The NP items were used included in first through third grade probes. First-grade probes included eight NP items: counting backward by 1's and forward by 2's, 5's, and 10's. Numerical values ranged from 1 to 99. The second-grade probes included eight NP items: counting by 2's, 3's, 4's, 5's, 25's, and 100's. Counting by 1's and 10's were excluded from the final probes due to low difficulty and discrimination indices. Second-grade NP items had numerical values ranging from 1 to 999. Third-grade probes included eight NP items: counting by 3's, 5's, 6's, 7's, 8's, 12's, 25's, and 100's.

Shape Patterns. The SPs were presented horizontally from left to right with one shape missing from each pattern. Each item included four possible shape choices to the

right of the sequence. Examinees were required to circle the shape that completed the sequence.. The placement of the missing shape and the correct response were determined by random assignment.

Number Sentence-Quantity Discrimination. The NSQD items were developed by combining math facts and quantity discrimination items, both of which have demonstrated good reliability and validity (Chard, et al., 2005; Clarke & Shinn, 2004; Epstein, et al., 1989; Foegen, 2000; L. Fuchs, et al., 2007; Thurber, et al., 2002; Tindal, et al., 1983; Vanderheyden, et al., 2004; VanDerHeyden, et al., 2003). The NSQD task consisted of a horizontally presented number sentence immediately followed by a quantity discrimination task. The vertically arranged symbols (i.e., $<$, $>$, $=$) separated the number sentence answer from a randomly assigned number. Addition items included randomly selected addends while subtraction items included randomly selected minuends. Subtrahend values were selected randomly from values less than the minuend. First, the examinees were required to write the answer to the number sentence and then circle the symbol representing the relationship between the examinee's written answer and the random number. Examinees received credit for each digit correctly written in addition to one point for circling the correct symbol, regardless of the correctness of the answer to the number sentence. The NSQD items were included in first through third grade probes.

First-grade and second-grade probes included eight NSQD items with addend, minuend, and subtrahend values ranging from two to nine, equally divided between addition and subtraction. Quantity discrimination random numbers ranged from 1 to 99.

Second-grade probes included eight NSQD items with addend, minuend, and subtrahend values ranging from 1 to 10. Items were equally divided between addition and subtraction. Quantity discrimination random numbers ranged from 1 to 99.

Third-grade probes consisted of eight NSQD items. Four items had addend, minuend, and subtrahend values ranging from two to nine, divided equally between addition and subtraction. There were also two multiplication and two division items with dividend, divisor, multiplicand, and multiplier values ranging from 1 to 10. All division items resulted in a whole number quotient. Quantity discrimination random numbers ranged from 1 to 99.

Computation. COMP items have also been established as valid and reliable measures of math fluency (Christ, et al., 2005; Connell, 2006; L. Fuchs & Fuchs, 1997; L. Fuchs, et al., 1994; Jitendra, et al., 2005; Shapiro, et al., 2005; Shinn & Marston, 1985). The COMP items required examinees to solve 2×2 or 3×3 addition and subtraction items. The COMP items did not require regrouping. Values for the 2×2 items ranged from 12 to 99, with any addition by one eliminated from second and third-grade items due to low difficulty and discrimination indices. Examinees received credit for each digit written correctly. Eight 2×2 COMP items were included in first-grade probes. Eight COMP items were included in second and third-grade probes, divided evenly between 2×2 and 3×3 addition and subtraction items.

Math Facts. Math Fact items required examinees to solve 1×1 addition and subtraction items. First-grade probes included four horizontally arranged Math Fact items with randomly assigned values ranging from 1 to 10.

Procedures.

Training activities. Prior to beginning the study, literacy leaders and school administrators were trained by the researchers to use the final form of the MAP:M and the MBSP during a four-hour training seminar. The trained school personnel were then responsible for providing supervision and training to the classroom teachers. Training of school personnel involved the researchers administering each grade-level probe to the school personnel. Next, the administrators were required to administer the probes to a partner, under supervision of the research team. The school administrators then scheduled an in-service for each teacher wherein classroom teachers were provided direct instruction on MAP:M administration procedures.

The participating school district designated approximately 30 personnel to score the probes. Members of the research team provided a four-hour training seminar for all scorers of the MAP:M. The scorers were also provided with a detailed list of scoring guidelines, correct answers, and data recording procedures.

The participating school district also designated personnel to enter scores into the school district's custom CBM database. Research team members worked closely with the district's database designer to ensure reliable data-entry procedures were followed. The scorers were provided an opportunity to practice inputting and manipulating data prior to the initial implementation of the MAP:M.

MAP:M probe administration. The probes were group administered in each designated classroom on a bi-weekly basis with two exceptions. The delay between probe 3 and probe 4 was three weeks due to mid-year break and three weeks between probes 9

and 10 due to spring break. The MAP:M was administered the classroom teachers in their usual classrooms at a time designated by the teacher and according to scripted instructions as before, and on the same day of the week. The classroom teacher first distributed the MAP:M practice exercise to each examinee. The guided/independent practice instructions were then read verbatim from the practice-exercise script. Examinees were provided examples of expected responses followed by an opportunity to complete each item type independently prior to administration of the scored items. The administrators were instructed to monitor examinees to ensure their understanding of the task demands and expected responses. After all practice exercise items were completed, the practice exercise was collected. The Guided/Independent Practice exercise was included only on the first and second administrations. Following administrations of the MAP:M included only the test probes and accompanying instructions.

The test probes were distributed to each examinee. The examinees were instructed to write their first name, last name, and date on the top of their paper. Once each examinee had finished writing the required information, the classroom teacher read the scripted test-probe administration instructions. The examinees were instructed to begin working and provided exactly 3 min (timed with a stopwatch) to complete as many items as possible. The classroom teachers were instructed not to assist examinees during the administration of the test probes. If assistance requested assistance, they were to instruct the examinee to skip the items they do not know and do the ones they did know. At the end of 3 min, the classroom teacher instructed the examinees to stop, put their pencils

down and then collected all materials. The probes were then scored and entered into the database.

The MBSP was administered to one classroom at each grade level from each participating school on the same day MAP:M probes were administered. The MBSP and MAP:M was administered in a counterbalanced fashion. The administration of the MBSP began with the examiner distributing copies of the MBSP to each examinee. Each examinee was instructed to write his or her first name, last name, and date at the top of the page. The teacher then read the scripted instructions from the MBSP. Examinees in first and second grade were allowed 2 min to complete the MBSP, while third-grade was allotted 3 min, as per the directions. The teachers used a stopwatch to monitor the time. After the allotted time had elapsed, the examinees were instructed to stop; the tests were collected, scored, and entered into the database.

Scoring directions. Each MAP:M item included a series of boxes equal to the number of possible correct responses. Examinees received credit for each part of an item they answered correctly (e.g., digits written correctly, correct item circled). The scorer identified each correct response and made a mark in each box, representing the number of correct response per item. The QD, NS, and SP items required one response where the COMP, NSQD, and NS items allowed two or more possible responses. After each item was evaluated, the scorer tallied the total number of checked boxes. The score was placed in the upper right hand corner of the test. The MBSP was scored according to the publisher's directions. The number of digits correct was counted then recorded in the database.

Administration fidelity. School administrators were instructed to conduct fidelity checks in each of the classrooms by designated school personnel who were trained in conducting behavioral observations. They were also instructed to ensure that fidelity data revealed a minimum compliance (more than 80%); otherwise, the teacher should be retrained in proper administration procedures. Administrators were instructed to check administration fidelity using a checklist, provided by the researchers, which described 10 key areas of administration:

- 1) Probes were distributed
- 2) Examinees were instructed to write their names on the probes
- 3) Scripts were read verbatim
- 4) Examinees were instructed to begin
- 5) A stopwatch was used to measure time
- 6) Teachers actively monitored examinees
- 7) Examinees were redirected when necessary with the scripted prompts
- 8) Administration lasted exactly 3 min
- 9) The stop script was read
- 10) Examinees were instructed to put their pencils on their desk at the end of 3 min.

Statistical procedures. Descriptive statistics (means and standard deviations), and Generalizability and Decision study analyses were conducted to determine score variances, Generalizability coefficients and the standard errors of measure. Pearson product-moment correlations were calculated to determine test-retest reliabilities of the

MAP:M and alternate-form reliabilities of MAP:M and MBSP. In addition, z and t tests were used to make comparisons between select correlation coefficients. In order to determine the relative predictive power of the MAP:M and MBSP with respect to STAR Math scores a step-wise multiple regression was calculated.

Chapter III

Results

This study was designed to investigate the psychometric properties of the Monitoring Academic Progress Mathematics (Hopkins, et al., 2010) and to determine its acceptability as a mathematics curriculum-based measure (M-CBM). The research questions were designed to assess the reliability and validity of the instrument relative to the Monitoring Basic Skills Progress (MBSP; L. S. Fuchs, et al., 1999). The MAP:M and MBSP were administered to special education and general education students across first, second, and third grade at eight elementary schools. Table 11 displays the descriptive statistics for the MAP:M and MBSP, including means and standard deviations.

The results of the Generalizability (G) and Decision (D) studies were carried out using analyses based on G-Theory. The studies investigated the object of measurement or persons (p), items (i) or probe scores, and residual error (pi,e). All studies used a single facet fully crossed design (p x i). The calculations were performed using the G1.sps SPSS syntax designed by Mushquash and O'Connor (2006) for use with random facet balanced designs, such as the current study. The algorithms are modeled after those presented in Brennan (2001, p. 79 and 439). Generalizability studies require a complete dataset with no missing data, thus data imputation was conducted instead of data deletion to reduce the likelihood of biasing effects due to data deletion. Data imputation was performed using the Estimate-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977).

G-Theory Estimates of Reliability

Estimated variance components. G-Studies were conducted using scores from the MAP:M and MBSP. The results of the G-Studies provide information about the percentage of variance in probe scores attributable to either the Person facet (student growth), Item facet (test difficulty), or Residual error.

Estimated variance components were calculated independently for first, second, and third grade MAP:M probes. Tables 12, 14, and 16 provide a summary of the MAP:M estimated variance of Person facet, Item facet, and the Residual error. Results show the greatest amount of variance in first-grade scores (62%) can be attributed to the Person facet. The variability in scores derived from the Person facet should be interpreted as systematic score fluctuations due to individual differences and not error (Mushquash & O'Connor, 2006). Residual error produced the second largest score variance (22%), and the Item facet (test difficulty) contributed the least amount of variance (16%).

Second-grade MAP:M probes produced similar results. The estimated variance due to the Person facet contributed the most variance (58%) and the least variance was attributed to the Item facet (13%). Residual error accounted 29% of score variance.

The variance associated with Person facet, Item facet, and Residual followed a similar pattern for third-grade MAP:M probes. The Person facet accounted for the largest amount of variance (66%), followed by the Residual (23%), and the Item facet estimates (11%).

Likewise, MBSP probes were explored using G-Studies to determine the score variances associated with the Person facet, Item facet, and Residual. Results are

presented in Tables 12, 14, and 16. The most variance in first-grade scores was due to the Person facet, 68%, while the least amount of variance was due to the Item facet, 1%. The residual accounted for the second largest amount of score variance, 31%, across first-grade MBSP probes. Second-grade probes showed that the Person facet accounted for the largest amount of variance, 38%, while the Item facet and Residual accounted for 7% and 55%, respectively. Third-grade maintained the same trend as first and second-grade with most of the score variance attributed to the Person facet (69%), followed by the Residual (1%) and Item facet (30%).

The MAP:M produced more Item facet variance than the MBSP, although the MBSP yields more Residual error variance. The MAP:M and MBSP show similar percentages of Person facet variance at first and third grade, while the MAP:M yields greater variance attributed to the Person facet at second grade. The second-grade MBSP probes produced more Residual error than Person and Item facet variance combined. Overall the MAP:M variance estimates account more for individual performance and item difficulty variance and less error than the MBSP.

Relative decisions: G-Coefficients and standard error of measure. G-Coefficients (similar to reliability coefficients in classical test theory) were calculated to provide an estimate of reliability when considering a relative decision (comparing a student relative to their peers). Tables 13, 15, and 17 provide a summary of the MAP:M Generalizability (G) Coefficients and Standard Error of Measure (SEM) scores for one, two, three, five, seven, and ten MAP:M probes across first, second, and third grades. The G-Coefficient at first grade ranged from .73 after one probe to .97 after ten probes.

Second grade G-Coefficients ranged from .67 to .95 across the ten probes. Third-grade G-Coefficients followed suit with a G-Coefficient of .75 after one probe and G-Coefficients from probes two through 10 ranging from .85 to .97. First, second, and third-grade all required at least two probes to obtain G-Coefficients exceeding the desired .80 level for reliable educational decisions (J. Sattler, 2008).

The SEM was derived based on the Estimated Variance estimate provided by the Decision (D) Study for one, two, three, five, seven, and ten MAP:M probes. The SEM was calculated by obtaining the square root of the estimated variance components of relative decisions (α^2_{Rel}). The SEM is used in a similar fashion to the SEM in classical test theory in that it can provide a band of confidence surrounding an obtained score, wherein the theoretical “true” score should lie. The SEMs for Relative Decisions at first-grade ranged from ± 4.05 with one probe to ± 1.28 after ten MAP:M probes. Second-grade SEMs ranged from ± 7.05 to ± 2.23 across the ten probes. In the same manner, third-grade SEMs ranged from ± 6.78 to ± 2.23 .

MBSP G-Coefficients and SEM scores for one, two, three, five, seven, and ten probes across first, second, and third grades were calculated in the same manner as was the MAP:M. Results are presented in Tables 13, 15, and 17. First-grade G-Coefficients ranged from .68 to .96 across ten MBSP probes. One MBSP probe resulted in a reliability estimate of .68, while at least two MBSP probes were required to obtain reliabilities above .80. Second-grade G-Coefficients were somewhat lower than first-grade, ranging from .58 to .93, indicating that three or more second-grade MBSP probes are required to obtain reliability estimates greater than .80. One or two probes resulted in reliabilities of

.58 and .73, respectively. Third-grade G-Coefficients were similar to first-grade, ranging from .70 to .96 and two or more MBSP probes resulted in G-Coefficients greater than .80.

In addition, SEMs were obtained for MBSP probes. When making relative decisions the SEM for first-grade MBSP probes ranged from ± 4.12 at one probe administration to ± 1.30 after ten probes. Second-grade ranged from ± 5.88 to ± 1.89 , while third-grade ranged from ± 6.18 to ± 1.95 .

The G-Coefficients of the MAP:M and MBSP are similar at first and third grades as both measures require two probe administrations to obtain reliability estimates that exceed the .80 threshold. Second-grade MAP:M probes perform better than do MBSP probes as MAP:M probes show reliabilities of .80 after two probes, while the MBSP does not show reliabilities exceeding .80 until seven probes. Overall, the SEM of the MAP:M and MBSP were similar across grades.

Absolute decisions: Index of dependability and standard error of measure.

Index of Dependability Coefficients was calculated to provide an estimate of reliability when considering an absolute decision (comparing a student on a criterion or cut score). Tables 12, 14, and 16 provide a summary of the MAP:M's Index of Dependability Coefficients and SEM scores for one, two, three, five, seven, and ten MAP:M probes across first, second, and third grades. The Index of Dependability Coefficients at first-grade ranged from .62 after one probe to .94 after ten probes. Second-grade coefficients ranged from .58 to .93. Third-grade MAP:M coefficients ranged from .66 to .95. First and

second-grade coefficients indicate that three or more MAP:M probes are required to obtain good reliability while third grade requires two or more probes.

Estimates of the SEM were obtained from the Estimated Variance provided by the D-Study for one, two, three, five, seven, and ten MAP:M probes. The SEM was derived by calculating the square root of the estimated variance components of absolute decisions (α^2_{Abs}). The SEMs for absolute decisions at first-grade ranged from ± 5.28 after one probe to ± 1.67 after ten MAP:M probes. Second-grade SEMs ranged from ± 8.52 at probe one to ± 2.69 after ten probes. Similarly, third-grade SEMs ranged from ± 8.17 to ± 2.58 .

The MBPS Index of Dependability (Φ) Coefficients and SEM scores were also obtained for one, two, three, five, seven, and ten probes across first, second, and third grades. Index of Dependability estimates for first-grade MBSP probes ranged from .68 after one probe to .96 after 10 probes. Second-grade MBSP probes resulted in lower Index of Dependability estimates, ranging from .38 at one MBSP probe to .88 after 10 probes. Lastly, third-grade Index of Dependability estimates were similar to first grade and ranged from .69 at probe 1 to .96 after 10 probes. Two probes were required to obtain reliability estimates greater than .80 for first and third grades; however, second-grade probes required seven probes.

The estimated SEM of the MBSP was obtained from the Estimated Variance provided by the D-Study for one, two, three, five, seven, and ten MBSP probes. When making absolute decisions first-grade probes resulted in SEMs ranging from ± 4.16 at probe one to ± 1.32 at probe 10. Second-grade SEMs were slightly higher, ranging from ± 6.35 to ± 2.01 . Finally, third-grade SEM estimates ranged from ± 6.31 to ± 1.99 .

With respect to absolute decisions, the MBSP showed stronger reliability estimates at grade one, while the MAP:M was stronger at grade two. Both the MAP:M and MBSP showed similar reliability estimates at grade three. The MBSP demonstrated smaller SEM estimates across grades than did the MAP:M.

Classical Test Theory Estimates of Reliability

Typical alternate-form reliability estimates are obtained by determining the relationship of scores obtained on multiple forms administered close in time. The alternate-form reliabilities presented below are not typical due to extended time delays between probe administrations. Alternate-form reliabilities were calculated by grade and independently for both the MAP:M and the MBSP for first, second, and third grades. The Pearson Product Moment Correlation was used to calculate correlations between MAP:M alternate forms and between MBSP alternate forms. Results are presented in Table 18 and 19, respectively.

Traditionally, reliability coefficients greater than or equal to .70 have been considered acceptable for research purposes (L. Cronbach, 1951); however, reliability coefficients of greater than .80 are suggested for psychoeducational purposes (Bracken, 1987; J. Sattler, 2008), and particularly The National Center on Response to Intervention has provided psychometric guidelines concerning curriculum-based measurement (CBM) as follows. Estimates greater than or equal to .80 are considered to be convincing evidence of adequate systematic variance, or reliability, .60 to .80 is considered to be partially convincing evidence, and reliabilities less than .60 are considered to have unconvincing evidence.

The first-grade MAP:M probes demonstrated an overall alternate-form reliability of .73. Alternate-form correlations ranged from .65 to .83 across the eleven forms. Second-grade overall MAP:M alternate-form reliability (.66) was similar to first grade. Correlations across forms ranged from .51 to .79. Overall third-grade alternate-form reliability was .71, with reliability estimates ranging from .59 to .80 across the eleven MAP:M forms.

MBSP overall alternate-form reliability at first grade (.70) was similar to that found at third-grade (.71). Second-grade overall alternate-form reliability was somewhat lower (.42). Correlations across forms ranged from .67 to .74 at first-grade, .37 to .50 at second grade, and .65 to .74 at third grade.

Z tests were calculated to determine if MAP:M alternate-form correlation coefficients were significantly different from MBSP coefficients. The analysis was conducted by calculating the average r-value at each grade level then calculating the average n across at each grade level. The average correlations were then transformed to z values using the Fisher r-to-z transformation. The transformed z scores were then used to determine differences between the correlation coefficients as described by Glass and Hopkins (1984). Results are presented in Table 20.

The z test comparison showed that neither first nor second-grade MAP:M or MBSP probes showed significantly different overall alternate-form correlations. At second grade, z tests showed that the MAP:M had a significantly higher alternate form correlation than did the MBSP, $z = 3.29$, $p < .01$.

First and third-grade MAP:M and MBSP probes demonstrated similar yet minimally adequate alternate-form reliabilities. Second-grade MAP:M alternate-form reliabilities were also adequate; however, MBSP reliabilities were significantly lower than those of the MAP:M.

Test-retest reliabilities were calculated using a Pearson Product Moment Correlation between two administrations of the MAP:M probe 11, spaced one day apart. Results are presented in Table 21. The test-retest sample included 236 students from one participating school in the district. Reliability estimates were .81 at first grade, .79 at second grade and .76 for third grade.

Reliability of the Slope

The rate of improvement (slope) was calculated for the MAP:M and MBSP to determine the expected rate of improvement and estimated error of the slope. Ordinal least squares regression was used to determine the slopes, with probe scores as the dependent variable and time (number of probes administered) as the independent variable. Results were calculated for both the full MAP:M sample in addition to a matching sample of MAP:M and MBSP scores from participants who completed both probes and STAR Math. Due to the MBSP being administered once per month across three months instead of bi-weekly, the resulting slopes and standard errors of the slope have been converted to represent bi-weekly indices similar to conversions implemented by Christ (2006). Results are presented in Tables 22.

The MAP:M full sample resulted in a first-grade slope of 1.00, while second and third-grade probe scores resulted in slopes of 1.40 and 1.20, respectively. The standard

error of the slope for first, second, and third grades was .04, .05, and .06, respectively.

Analyses were conducted on a smaller matching MAP:M, consisting of participants who completed both the MAP:M probes and at least two MBSP probes. This sample showed increasing slopes at first (.63) and second grade (1.58) and decreasing slopes at third grade (-0.29). Standard error of the slope also increased significantly in the smaller sample as first grade was .29, second grade, .41, and .48 for third grade.

The slope and standard error of the slope were calculated for a smaller, representative sample, across three MBSP probes at first, second, and third-grade. Only students completing at least two MBSP probes were included in the analyses. Because the MBSP was administered once per month, the resulting slopes and standard errors were divided by two to represent bi-weekly growth, equivalent to the overall MAP:M sample. Results show that MBSP had a slope of -0.20, 1.07, and -0.27 for first, second, and third grades respectively. Standard errors of measure were .24, .25 and .35 for first, second, and third grades, respectively.

The Overall MAP:M sample provided positive and reliable slopes over the course of the academic year. However, the smaller sample that included only three administrations of the MAP:M resulted in significantly increased error of the slope in addition to a negative slope in third grade. The MBSP probes, administered to the same subjects as the smaller MAP:M sample, resulted in negative slopes for first and third grades. Both the MAP:M and MBSP had larger slope error estimates than did the full MAP:M sample. The MAP:M showed steeper rates of improvement than did the MBSP.

Validity of the MAP:M

Concurrent validity between the MAP:M and MBSP was determined by comparing coefficients from probes taken from probes administered on the same day. Results are presented in Table 23. First-grade MAP:M probes resulted in correlations ranging from .58 to .73, while second grade MAP:M probes provided slightly lower correlations of .37 to .52. The relationship between MAP:M and MBSP probes at third-grade yielded correlations from .60 to .74.

Coefficients of determination were also calculated to determine the amount of score variance shared between the MBSP and MAP:M. First-grade MAP:M probes' Coefficients of determination ranged from .34 to .53, meaning that 34% to 53% of the variance is shared by the MBSP and MAP:M. Similarly, third-grade coefficients range from .26 to .54. Second-grade coefficients are somewhat lower than first and second grade as they range from .14 to .27. These coefficients show that the MAP:M shares between 14% and 53% of score variance with the MBSP scores.

A stepwise multiple regression was used to determine whether the MAP:M or the MBSP probe best predict end-of-year STAR Math scores for 29 participants. Tests of assumptions of multiple regression were calculated prior to analysis. The Shapiro-Wilk test of normality indicated that MAP:M and STAR Math scores were normally distributed; however, MBSP scores were not. MBSP scores were transformed using the square root method and stepwise multiple regressions were calculated using transformed and non-transformed scores. Score transformations did not result in significant differences in the amount of variance explained. Table 24 reports results from the non-

transformed data and Table 25 contains the correlations coefficients. The correlations ranges from .38 to .69 with the MAP:M and MBSP demonstrating similar correlations at each grade level; hence, multi-collinearity should not be a problem.

Using stepwise multiple regression variables are included or excluded from the model based on whether they contribute significantly to the relative predictive power of the model. Results show that the MAP:M alone resulted in the strongest relative predictive power of end-of-year STAR Math scores, $R^2 = .33$, $F(28) = 13.11$, $p < .01$. MBSP scores were not included in the final predictive model as it did not significantly contribute to the relative predictive power of the model, $\beta = .26$, $t(28) = 1.26$, $p = .26$. The final predictive model shows that MAP:M scores predict 33% of the variance in end-of-year STAR Math scores.

Chapter IV

Discussion

The purpose of this study was to describe the development and validation for a newly developed math curriculum-based measure (M-CBM), the Monitoring Academic Progress: Mathematics (MAP:M; Hopkins, et al., 2010), by examining its psychometric properties and comparing its properties to commercially available measures. Results from multiple analyses reveal moderate to strong evidence of the psychometric quality of the MAP:M as discussed in the following sections.

Uses of curriculum-based measurement of mathematics. Screening and student progress monitoring within a response to intervention (RTI) model typically uses mono-operational measures of math fluency. Although these measures often have adequate reliability and validity, sizeable variability exists with regard to decision-making accuracy (L. Fuchs, et al., 2007; VanDerHeyden, et al., 2003) because of considerable measurement error (Christ, et al., 2005). Additionally, traditional computation-based M-CBM does not measure problem-solving skills at the earlier grades; although, some research has examined problem-solving M-CBM at latter grades (Jitendra, et al., 2005). Nonetheless, co-morbid academic deficiencies make it difficult to distinguish between students with reading, math, or combined academic deficits when utilizing language-laden problem-solving tasks.

The MAP:M was designed to be an indicator of both math fluency and math problem solving by using a multi-dimensional array of tasks designed to provide a variety of response types in an attempt to reduce error. The MAP:M items were derived from a

combination of prior math curriculum-based measurement research (Chard, et al., 2005; Clarke & Shinn, 2004; Epstein, et al., 1989; Foegen, 2000; L. Fuchs, et al., 2007; Shinn & Marston, 1985; Thurber, et al., 2002; Tindal, et al., 1983; Vanderheyden, et al., 2004; VanDerHeyden, et al., 2003) (Chard, et al., 2005; Clarke & Shinn, 2004; Epstein, et al., 1989; Foegen, 2000; L. Fuchs, et al., 2007; Thurber, et al., 2002; Tindal, et al., 1983; Vanderheyden, et al., 2004; VanDerHeyden, et al., 2003). Following is a discussion of reliability and validity estimates based on the current study and comparisons with previous research concerning the item types included in the MAP:M.

G-Theory Estimates of Reliability

Estimated variance components. Generalizability (G) Theory was used to determine the amount of variance in scores across administrations attributed to either the Person facet (academic growth), Item facet (test difficulty), and Residual error for both the MAP:M and MBSP. Variances were calculated independently across first, second, and third grades. The percentage of total score variance for the MAP:M for each of the three grades was consistent. The MBSP yielded similar results at first and third grades; however, more score variance was attributed to changes in academic growth for MAP:M scores than MBSP scores at second-grade.

Variability in test difficulty accounted for the least amount of score variances for both the MAP:M and MBSP; although the MAP:M demonstrated a larger amount of variability in test difficulty than did the MBSP. This discrepancy indicates that the consistent application of a mono-operational fluency-based assessment, such as the MBSP, results in less score fluctuation due to variations in probe difficulty.

Despite the lower score variance attributed to variations in test difficulty the MBSP demonstrates higher levels of score variation due to Residual error. At each grade, the MAP:M resulted in lower variations in score due to Residual error than did the MBSP. It would appear as though the MBSP provides a lower amount of variance due to item difficulty; however, the trade-off results in a greater amount of unexplained Residual error. These results provide adequate evidence of the reliability of the MAP:M and indicate that variations in MAP:M scores over time can be attributed most to real change in student progress.

The MAP:M G-Coefficients are encouraging in light of the findings by Christ et al. (2005) who found as much as 68% of M-CBM score variance to be accounted for by Residual error and only 17% by the Person facet. In a similar study, Hintze et al. (2002) examined single and multi-skill computation probes and found the Person facet accounted for 32% of the variance across both single and multi-skill probes; however, 74% of the variance on single-skill (and 32% on multi-skill probes) was attributable to the Person facet when parsed out of the full model. Apparently, the bulk of the fluctuations in MAP:M scores over time are due to changes in the Person facet (academic growth). According to Muschquash & O'Connor (2006), variability in scores attributed to the Person facet should be interpreted as systematic score fluctuations due to individual differences in skill level and not error.

Relative decisions: G-Coefficients and standard error of measure. The Generalizability (G) Coefficient is similar to the reliability coefficients of classical test theory and represent the reliability of making relative decisions, i.e., those which are

made about individuals when compared to the performance of others (Shavelson & Webb, 1991). Thus, the G-Coefficient indicates how consistently individual students maintain their ranking (Mushquash & O'Connor, 2006), based on raw scores, across probes. According to data from G-Coefficients and Standard Error of Measure (SEM) of MAP:M probe, administering one MAP:M probe provides reliability estimates below the suggested .80 level (Sattler, 2001); however, using National Center on Response to Intervention guidelines, one probe provides partially convincing evidence of reliability with estimates range from .68 to .74. Administering two probes provides convincing evidence of reliability (i.e., correlations exceeded .80). These findings indicate that at least two probes should be administered when making relative decisions. In a similar study using single-skill M-CBM probes, Hintz et al., (2002) found a G-Coefficient of .95 after one administration. Christ et al., (Christ, et al., 2005) also found G-Coefficients exceeding .80 after a 2 min administration of a multi-skilled M-CBM probe. Although the MAP:M obtained slightly lower estimates of reliability for relative decisions it provides more information about a variety of mathematical computation and reasoning skills than a single M-CBM, and hence is more useful to teachers. In fact, G-Coefficients could be calculated for individual MAP:M items and will be the focus of future research.

In order to aid in interpreting error produced by the MAP:M, the SEM was also calculated for each grade for one, two, three, five, seven, and ten MAP:M probe administrations. Because SEM is often used to calculate a confidence interval surrounding an obtained score, it is understood by most test users. Obviously, SEM is important when making educational decisions, especially for those students who fall on

the borderline of receiving assistance based on selection criteria for inclusion in the RTI tiers. At the recommended minimum, two MAP:M administrations the SEM was ± 2.87 for first grade, ± 4.84 at second grade, and ± 4.79 at third grade. The MAP:M demonstrates SEMs consistent with those of the MBSP. The MAP:M SEM estimates are slightly higher than SEMs of the Discovery Education Predictive Assessment (± 2.11 to ± 2.69 ; National Center on Response to Intervention, 2010), a newly developed commercial M-CBM. Obviously more precision and confidence can be gained by using three (or more) probes for decision-making.

Absolute decisions: Index of Dependability and standard error of measure.

The Index of Dependability is similar to the G-Coefficient, in that it provides an estimate of reliability similar to classic test theory; however, the Index of Dependability is used when making absolute decisions and indicates how accurately a student's obtained score can locate them on a fixed scale, without respect to other student's scores (Cardinet, et al., 2010). An absolute decision would be appropriate for comparing scores for determining student placement when a cut score is the criterion. Index of Dependability Coefficients of the MAP:M for one and two probes ranged from .58 to .79 across first, second, and third grades. At each grade level, at least three probes were required to obtain reliability estimates that exceeded the .80 threshold. In contrast, the MBSP D-Study provided evidence that two MBSP probes are acceptable at first and third grade, while second-grade requires approximately seven probes. Both the MAP:M and MBSP demonstrate strong reliability after two to three administrations for first and third grades. The MAP:M appears to be more efficient and produces stronger reliability at second

grade than did the MBSP. Results show somewhat lower reliability estimates for absolute decisions than found by Hintz et al., (2002), who estimated a single-skill probe Index of Dependability of .75 after a single 2 min probe administration. Christ et al., (2005) demonstrated that a 6 min multi-skill probe was required to obtain reliability estimates exceeding .80, which is equivalent to the time required to administer two MAP:M probes. In general, single-skill probes contain less error than multi-skill probes, but are less representative of actual math curricula and less useful to teachers.

When SEMs were calculated for absolute decisions for one administration values of 5.28, 8.36, and 8.17 were attained at first, second, and third grades, respectively. The SEMs were reduced by almost half when the “recommended” three probes were administered, obtaining SEM of 3.05 (first grade), 4.83 (second grade), and 4.71 (third grade). The SEM for MBSP probes were similar to those found for MAP:M.

Classical Test Theory Estimates of Reliability

Within classical test theory, alternate form data are often recommended to establish reliability for timed tests (Dick & Dagerty, 1971). Alternate-form reliability estimates of the MAP:M ranged from moderate to strong at all grades, while the MBSP ranged from moderate to strong at second and third grade; however, second-grade estimates were poor for the MBSP. The MAP:M reliabilities were moderate and similar to those found for single-skill probes (Tindal, et al., 1983), missing number identification (Chard, et al., 2005), and counting (Clarke & Shinn, 2004; Vanderheyden, et al., 2004) tasks. This is significant because it demonstrates that similar alternate-form reliability estimates can be obtained for both multi-skill and single-skill probes over a short time.

The MAP:M alternate-form reliability estimates are slightly lower than those found for computation-based probes (Thurber, et al., 2002); however, the current study shows MAP:M alternate-form correlations to be equivalent or stronger than the MBSP, a computation-based M-CBM probe. Alternate-form correlations for both the MAP:M and MBSP at second-grade for the current study were lower than those reported by the National Center on Response to Intervention for AIMSWeb Math, a commercially available M-CBM. Discrimination tasks alone (Chard, et al., 2005; Clarke & Shinn, 2004; Vanderheyden, et al., 2004) have also shown stronger alternate-form reliability than found for the MAP:M..

Lower alternate-form correlations, relative to some M-CBM measures reported in the literature, were found. The MAP:M alternate-form correlations were calculated across the span of the school year instead of collecting all the data within a very short period of time. This strategy may account for the lower correlations that fall between beginning of the year probes and end of the year probes. As noted in Table 13, as the year progresses, so do the standard deviation of MAP:M scores. This increasing variability is produced by differing levels of student growth over time and likely artificially reduced the correlations between beginning and end of year scores.

Test-retest reliability estimates of the MAP:M were calculated as a second measure of classic test-theory reliability. Results showed acceptable to good reliability ranging from .76 to .81. These estimates are similar to those found for single skill probes (Epstein, et al., 1989; Tindal, et al., 1983), discrimination tasks (Chard, et al., 2005; Clarke & Shinn, 2004; Vanderheyden, et al., 2004), counting tasks (Clarke & Shinn,

2004), and Number Identification tasks (Clarke & Shinn, 2004). Again, the similar reliability estimates between the MAP:M, multi-modal M-CBM, and other mono-operational M-CBMs that include elements from the MAP:M provide strong evidence for the use of multiple item types.

Reliability of the Slope

The slope, standard error of the slope, and reliability of the slope for the MAP:M full sample, MAP:M matching sample and MBSP were calculated. For comparative purposes, a MAP:M matching sample was created from existing data (from participants who had also completed the MBSP probes). The reported MAP:M full sample slopes represent the average change in scores on a bi-weekly basis across the entire school year. The full MAP:M sample demonstrated positive slopes and relatively small error estimates across all grades. The MAP:M matching sample and MBSP sample scores were taken once per month across three months. For comparative purposes, the matching sample slopes were divided by two to represent bi-weekly growth. Both the MAP:M full sample and MAP:M matching sample demonstrated positive slopes at first and second grades. Both also showed similar standard error of the slope. MBSP slope estimates for first and third grades were both negative, indicating an overall decrease in scores across the three MBSP probes. Second-grade probes resulted in an increasing slope similar to that of the MAP:M matching sample. The findings of the MAP:M and MBSP matching samples were consistent with previous research by Shapiro et al. (2005), i.e., computation probes have a standard error of the slope ranging from .15 to .31. In addition, slopes for both the MAP:M full and matching samples are similar to those from previous research by Fuchs

et al. (1994) and Fuchs et al. (2007) for Concept and Application probes that include problem solving.

Validity of the MAP:M

The MAP:M and MBSP were compared at three points in time to demonstrate concurrent validity of the MAP:M as a measure of math fluency. Strong correlations coefficients were obtained for first and third grade (.51 to .73), while second-grade correlations were slightly lower (.37 to .52). Computation probes, such as those included in the MBSP, have demonstrated criterion validity with achievement tests such as the WJ-III mathematics and ITBS test that range from .57 to .76. Apparently, the strong correlations found between the MAP:M and MBSP provide evidence that MAP:M is a good indicator of math fluency and may be a strong indicator of overall mathematics ability; although this is a hypothesis for future research.

In addition, results from a stepwise multiple regression showed that the best predictor of end-of-year STAR Math scores was the MAP:M. The MBSP scores did not add significantly to the predictive model. Although these findings are promising it should be noted that the results of this analysis were derived from a small sample ($n = 29$). Of note, correlations between both the MAP:M and MBSP with end-of-year STAR Math scores are similar to those found between AIMSWeb Math and a computerized individually administered math assessment (.52 to .54).

In summary, MAP:M produced adequate to strong reliability and validity when compared to other M-CBM measures including current commercially available M-CBM assessments. The Generalizability Studies provides evidence that the MAP:M is a viable

M-CBM, and is capable of measuring changes in individual academic growth over time utilizing multi-skill-item types. This is a unique contribution to the M-CBM literature and provides impetus for further research investigations of multi-skill item-types within a time-limited paradigm.

Limitations and Future Research

There were several limitations of this study, for example, timing of probe administration was not constant. At the onset of the current study a representative sample of students were scheduled to take the MBSP in addition to the MAP:M probes. Due to logistic considerations, the MBSP probes were not implemented until six MAP:M probes had already been administered. The MBSP probes were also administered one time per month in conjunction with the MAP:M. This may limit the fidelity of the predictive analyses because the STAR Math scores were obtained only a few months after the associated MAP:M and MBSP scores. Ideally, the MAP:M and MBSP scores would have been taken together and at the beginning of the school year. In addition, the comparisons of the Slope analyses were based on small sample sizes.

The timing of MAP:M probe administration may limit the interpretability of alternate-form reliability estimates. The correlation matrix of alternate-form reliabilities show that probes administered close in time with one another show higher reliability estimates than those administered further apart. The lowest reliability estimate are those that make comparisons between probes one through five and probes seven through eleven. In the future research should limit the time between administrations of alternate forms to obtain the most accurate estimate.

Limited sample size was also a limitation. STAR Math scores were obtained from only one participating school for analyses. In addition, STAR Math scores were not available for first-grade students. This resulted in a limited sample size with matching MAP:M, MBSP, and STAR Math scores. Thus, the predictive analyses contain only a small portion of the entire sample at each grade level, limiting generalizability and power.

Missing data was also another limitation. A missing value analysis conducted during data imputation for the Generalizability Study analysis showed that approximately 10% of data was missing for each probe, due to absenteeism, and students transferring in and/or out of the school district. The EM algorithm, discussed earlier, was used to impute values for the Generalizability Study due to data requirements for analysis. Although this method has been demonstrated to reliably and accurately estimate values (based on a maximum-likelihood model), there is no way to determine the extent to which missing scores affected analyses. All other analyses relied on pairwise exclusion of missing values. This procedure may result in biased reliability estimates to an unknown degree. In addition, two testing occasions relied on a three-week interval rather than a two-week interval, due to scheduling conflicts.

A final limitation of the study was researcher's inability to directly observe probe administration and scoring fidelity. The school officials were trained in implementation and scoring procedures on multiple occasions, but estimates of administration fidelity were not available.

In general, future researchers should explore the use of additional time-limited, multi-dimensional measures of math ability. Much of the current research focuses on time-limited assessment of elementary or splinter skill fluency; however, this may not provide the best information to guide instruction. Future MAP:M research should focus on obtaining additional alternate-form data, with probes administered close in time. Also the validity of the MAP:M should be further explored using end-of-year outcome measures such as State Mandated Achievement or Standardized Individual Achievement Tests as the criterion measures. Exploratory factor analysis studies may also focus on exploring the hypothesized structures, which include math fluency and math problem solving as two independent factors.

Future research should explore the accuracy of the MAP:M at identifying students with a math learning disorder versus those without a math learning disorder using data from students placed in tiers using MAP:R results. This important element is not addressed by the current study. Determining the sensitivity and specificity of the MAP:M will provide additional confirmation of the MAP:M as a viable M-CBM. Finally, acceptability data should be obtained to determine the perceived utility of MAP:M by teachers and administrators.

References

107th Congress. (2002). *No Child Left Behind Act of 2001*.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing. (2002). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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, 362-380. doi: 10.1177/073428290502300405

Ardoin, S. P., et al. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psychology Review*, 33, 218-233.

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

Bracken, B. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment*, 5, 313-326.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brigance, A. (1985). *Brigance Preschool Screen*. North Billerica, MA: Curriculum Associates.

- Brigance, A. (1999). *Comprehensive Inventory of Basic Skills, Revised*. North Billerica, MA: Curriculum Associates, Inc.
- Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York: Routledge.
- Carpenter, T., Fennema, E., Franke, M., Levi, L., & Empson, S. (1999). *Children's Mathematics: Cognitively Guided Instruction*. Portsmouth, NH: Heinemann.
- Center, T. I. STAR Legacy Modules Retrieved July 8, 2009, from http://iris.peabody.vanderbilt.edu/rti01_overview/rti01_05.html
- Chard, D. J., et al. (2005). Using measures of number sense to screen for difficulties in mathematics: preliminary findings. *Assessment for Effective Intervention*, 30, 3-14. doi: 10.1177/073724770503000202
- Christ, T. J. (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, 128-133.
- Christ, T. J., Johnson-Gros, K. N., & Hintze, J. M. (2005). An examination of alternate assessment durations when assessing multiple-skill computational fluency: The generalizability and dependability of curriculum-based outcomes within the context of educational decisions. *Psychology in the Schools*, 42, 615-622.
- 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.

- Connell, J. E., Jr. (2006). *Constructing a math applications, curriculum-based assessment: An analysis of the relationship between applications problems, computation problems and criterion-referenced assessments*. 66, ProQuest Information & Learning, US. Retrieved from <http://proxygsu-psin.galileo.usg.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2006-99002-004&site=ehost-live>
- Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 2, 151-160.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: NY: John Wiley & Sons, Inc.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1-38.
- Deno, S. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52, 219-232.
- Deno, S. (1987). Curriculum-based measurement. *Teaching Exceptional Children*, 20, 40-47.
- Deno, S., Fuchs, L., Marston, D., & Jongho, S. (2001). Using curriculum-based measurements to establish growth standards for students with learning disabilities. *School Psychology Review*, 30, 507.

- Deno, S., & Mirkin, P. (1977). *Data-based program modification: A manual*: Council for Exceptional Children (Publication Sales Unit), 1920 Association Drive, Reston, Virginia 22091.
- Deno, S., Mirkin, P., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, 49, 36-45.
- Deno, S., Mirkin, P., Chiang, B., & Lowry, L. (1980). Relationships among simple measures of reading and performance on standardized achievement tests.
- Deno, S., Mirkin, P., Lowry, L., & Kuehnle, K. (1980). Relationships among simple measures of spelling and performance on standardized achievement tests.
- Deno, S., Mirkin, P., & Marston, D. (1980). Relationships among simple measures of written expression and performance on standardized achievement tests.
- Dick, W., & Dagerty, N. (1971). *Topics in measurement: Reliability and validity*. New York, NY: McGraw-Hill Book Company.
- Epstein, M. H., Polloway, E. A., & Patton, J. R. (1989). Academic achievement probes: Reliability of measures for students with mild mental retardation. *Special Services in the School*, 5, 23-31.
- Foegen, A. (2000). Technical Adequacy of General Outcome Measures for Middle School Mathematics. *Diagnostique*, 25, 175-203.
- Fuchs, D., & Fuchs, L. S. (2006). Introduction to Response to Intervention: What, why, and how valid is it? *Reading Research Quarterly*, 41, 93-99.
- Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-Intervention: Definitions, evidence, and implications for the learning disabilities

- construct. *Learning Disabilities Research & Practice (Blackwell Publishing Limited)*, 18, 157-171.
- Fuchs, L., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, 53, 199-208.
- Fuchs, L., & Fuchs, D. (1997). Use of curriculum-based measurement in identifying students with disabilities. *Focus on Exceptional Children*, 30, 1-16.
- Fuchs, L., et al. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children*, 73, 311-330.
- Fuchs, L., et al. (1994). Technical features of a mathematics concepts and applications curriculum-based measurement system. *Diagnostique*, 19, 1994.
- Fuchs, L. S., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *School Psychology Review*, 28, 659.
- 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., Hamlett, C. L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review*, 22, 27.
- Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1999). *Monitoring basic skills progress basic math manual*. Austin, TX: Pro-ed.

- Gickling, E. T., V. (1985). A personal view of curriculum-based assessment. *Exceptional Children*, 52, 205-218.
- Ginsburg, H. P., & Baroody, A. J. (1990). *Test of Early Mathematics Ability* (2nd ed.). Austin, TX: Pro-ed.
- Glass, G., & Hopkins, K. (1984). *Statistical Methods in Education and Psychology* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Hintze, J. M., Christ, T. J., & Keller, L. A. (2002). The Generalizability of CBM Survey-Level Mathematics Assessments: Just How Many Samples Do We Need? *School Psychology Review*, 31, 514.
- 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.
- Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2005). *Iowa Test of Basic Skills*. Rolling Meadows, IL: Riverside Publishing.
- Hopkins, M., McCallum, S., Bell, S., & Mounger, A. (2010). *Monitoring Academic Progress: 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, 10-17.
- Howell, K. W. (1986). Direct assessment of academic performance. *School Psychology Review*, 15, 324-355.

- Howell, K. W., & Morehead, M. K. (1987). *Curriculum-based evaluation for special and remedial education: A handbook for deciding what to teach*. Columbus, OH: Charles E. Merrill.
- Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C., § 1400 et seq. Stat. (2004).
- Jankowski, E. A. (2003). Heartland area education agency's problem solving model: An outcomes-driven special education paradigm. *Rural Special Education Quarterly*, 22, 29-36.
- Jitendra, A. K., Sczesniak, E., & Deatline-Buchman, A. (2005). An exploratory validation of curriculum-based mathematical word problem-solving tasks as indicators of mathematics proficiency for third graders. *School Psychology Review*, 34, 358-371.
- Larson, N. (2004). *Saxon Math*. Norman, OK: Saxon Publishers, Inc.
- MacMillan, D. L., & McGraw-Hill. (1989). *Comprehensive Test of Basic Skills* (4th ed.). Monterey, CA: Macmillian/McGraw-Hill School Publishing Co.
- Marston, D., Muyskens, P., Lau, M., & Canter, A. (2003). Problem-solving model for decision making with high-incidence disabilities: The Minneapolis experience. *Learning Disabilities Research & Practice (Blackwell Publishing Limited)*, 18, 187-200.
- Mushquash, C., & O'Connor, B. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods*, 38, 542-547.

- National Center on Response to Intervention (Producer). (2010, 06/03/2010). National Center on Response to Intervention: Screening tools chart. [Report] Retrieved from <http://www.rti4success.org/chart/screeningTools/screeningtoolschart.html#>
- National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. Reston, VA: Author.
- Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York, NY: McGraw-Hill Book Company.
- Okamoto, Y., & Case, R. (1996). Exploring the microstructure of children's central conceptual structures in the domain of number *The role of central conceptual structures in the development of children's thought: Monographs of the Society for Research in Child Development* (Vol. 1, pp. 27-58). Malden, MA: Blackwell Publishers.
- Renaissance Learning. (2002). *Star math*. Madison, WI: Author.
- Renaissance Learning. (2009). STAR Math: Understanding reliability and validity. Retrieved from <http://research.renlearn.com/research/pdfs/135.pdf>
- Sattler, J. (2008). *Assessment of children : cognitive foundations*. San Diego: J.M. Sattler.
- Sattler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119-144.
- Shapiro, E. S., Edwards, L., & Zigmond, N. (2005). Progress monitoring of mathematics among students with learning disabilities. *Assessment for Effective Intervention*, 30, 15-32. doi: 10.1177/073724770503000203

- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory : A primer*. Newbury Park, Calif.: Sage Publications.
- Shinn, M., & Hubbard, D. (1992). Curriculum-based measurement and problem-solving assessment: Basic procedures and outcomes. *Focus on Exceptional Children*, 24, 1-20.
- Shinn, M., & Marston, D. (1985). Differentiating mildly handicapped, low-achieving, and regular education students: A curriculum-based approach. *RASE: Remedial & Special Education*, 6, 31-38.
- Silva, F. (1993). *Psychometric foundations and behavioral assessment*. Newbury Park: CA: Sage.
- Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: the importance of individual progress monitoring. *Learning Disabilities Research & Practice*, 15, 128 - 134.
- Tennessee Department of Education. (2004). Tennessee Curriculum Standards Tennessee Department of Education (Ed.) Retrieved from <http://www.state.tn.us/education/ci/curriculum.shtml>
- Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review*, 31, 498.
- Tindal, G., Germann, G., & Deno, S. (1983). Descriptive research on the pine county norms. A compilation of findings. Minneapolis, Min: University of Minnesota.

- Torrance, H., & Pryor, J. (1998). *Investigating formative assessment*. Buckingham: Open University Press.
- Individuals with Disabilities Education Improvement Act of 2004, STAT. 2706 C.F.R. § 602 (2004).
- Vanderheyden, A. M., et al. (2004). Development and validation of curriculum-based measures of math performance for preschool children. *Journal of Early Intervention*, 27, 27-41. doi: 10.1177/105381510402700103
- VanDerHeyden, A. M., Witt, J. C., & Naquin, G. (2003). Developmental and validation of a process for screening referrals to special education. *School Psychology Review*, 32, 204.
- VanDerHeyden, A. M., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students.(Statistical Data Included). *School Psychology Review*, 30, 363(320).
- Vaughn, S. (2003). *How many tiers are needed for response to intervention to achieve acceptable prevention outcomes?* Paper presented at the National Research Center on Learning Disabilities Responsiveness-to-intervention symposium, Kansas City, MO.
- Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery-Revised*. Allen, TX: DLM Teaching Resources.

Appendix A

TABLE 1

MAP:M Quantitative Pilot Study One: Time-Unlimited Means, Standard Deviations, and Independent *t* tests.

	M	SD	<i>t</i>	Sig
First-grade	8.69	4.04		
Class 1	8.67	3.90		
Class 2	8.71	4.30	-0.03	.98
Second-grade	28.26	6.89		
Class 1	27.00	7.28		
Class 2	29.60	6.42	-1.05	.30
Third-grade	41.49	4.28		
Class 1	41.38	4.63		
Class 2	41.60	4.03	-0.14	.89

TABLE 2

MAP:M Quantitative Pilot Study One: Time-Unlimited First-Grade Difficulty and Discrimination Indices by Item Type.

Item Type	Item Number	Difficulty	Discrimination
Computation	B	.04	.11
Computation	L	.04	-.11
Math Fact	D	.33	.78
Math Fact	F	.33	.44
Math Fact	H	.78	.33
Math Fact	J	.37	.56
Math Fact	N	.22	.56
Math Fact	P	.30	.67
Number Sequence	C	.85	.44
Number Sequence	M	.33	.78
Quantity	A	.33	.44
Discrimination			
Quantity	E	.59	.33
Discrimination			
Quantity	K	.41	.67
Discrimination			
Quantity	O	.33	.67
Discrimination			
Shape Pattern	G	.85	.44
Shape Pattern	I	.74	.56

TABLE 3

MAP:M Quantitative Pilot Study One: Time-Unlimited Second-Grade Difficulty
and Discrimination Indices by Item Type.

Item Type	Item Number	Difficulty	Discrimination
Computation	A	.42	.20
Computation	D	.94	.10
Computation	M	.81	.10
Computation	P	.52	.60
Number Sequence	C	.97	.00
Number Sequence	H	.34	.60
Number Sequence	J	.45	.40
Number Sequence	O	.00	.00
Number-Sentence QD	B	.71	.30
Number-Sentence QD	E	.23	.20
Number-Sentence QD	G	.55	.40
Number-Sentence QD	I	.45	.60
Number-Sentence QD	K	.52	.40
Number-Sentence QD	N	.39	.40
Shape Pattern	F	.74	.30
Shape Pattern	L	.81	.30

TABLE 4

MAP:M Quantitative Pilot Study One: Time-Unlimited Third-Grade Difficulty and Discrimination Indices by Item Type.

Item Type	Item Number	Difficulty	Discrimination
Computation	A	.94	.11
Computation	D	.78	.11
Computation	G	.84	.22
Computation	J	.95	.33
Number Sequence	M	.63	.44
Number Sequence	P	.75	.58
Number Sequence	C	.81	.00
Number Sequence	H	.97	.02
Number-Sentence QD	I	.47	.33
Number-Sentence QD	N	.78	.68
Number-Sentence QD	B	.56	-.01
Number-Sentence QD	E	.59	.11
Number-Sentence QD	L	.88	.36
Number-Sentence QD	O	.06	.47
Shape Pattern	F	.91	-.10
Shape Pattern	K	.84	.22

TABLE 5
MAP:M Quantitative Pilot Study One: Time-Limited Means, Standard Deviations, and
Independent *t* tests.

	n	M	SD	<i>t</i>	<i>p</i>
First-grade					
Class 1	14	2.79	2.08		
Class 2					
Second-grade	30	9.27	4.46		
Class 1	14	10.21	5.07		
Class 2	16	8.44	3.83	1.09	.28
Third-grade	38	15.29	5.77		
Class 1	21	14.57	5.95		
Class 2	17	16.18	5.59	-0.85	.40

TABLE 6
MAP:M Quantitative Pilot Study Two: Mean, Standard Deviations, Internal Consistency,
and Standard Error of Measure

	n	M	SD	α	SEM
First-grade probe	114	8.58	4.76	.78	2.23
Second-grade probe	99	13.81	7.31	.67	4.20
Third-grade probe	122	25.62	9.75	.73	5.07

TABLE 7

MAP:M Quantitative Pilot Study Two: First-Grade Difficulty and Discrimination Indices.

Item Number	Item Type	Difficulty	Discrimination
A	QD	.63	.71
B	QD	.39	.27
C	COMP	.77	.30
D	NS	.83	.17
E	SP	.65	.65
F	SP	.60	.42
G	NS	.22	.28
H	COMP	.03	.05
I	QD	.63	.60
J	SP	.02	.05
K	NS	.17	.44
L	COMP	.41	.76
M	QD	.45	.42
N	COMP	.22	.43
O	NS	.39	.66
P	SP	.31	.64
Q	QD	.14	.31
R	COMP	.06	.18
S	NS	.10	.23
T	SP	.09	.15
U	SP	.04	.10
V	NS	.02	.03
W	COMP	.01	.03
X	QD	.01	.00
Y	SP	.03	.03
Z	NS	.00	.00
AA	COMP	.00	.00
AB	QD	.02	.03
AC	QD	.01	.00
AD	COMP	.01	.03
AE	NS	.01	-.03
AF	SP	.00	.00

TABLE 8
MAP:M Quantitative Pilot Study Two: Second-Grade Difficulty and Discrimination
Indices.

Item Number	Item Type	Difficulty	Discrimination
A	NSQD	.92	.12
B	COMP	.56	.39
C	NS	.63	.21
D	SP	.80	.03
E	NSQD	.73	.33
F	COMP	.52	.55
G	NS	.33	.36
H	SP	.53	.39
I	SP	.59	.48
J	NS	.24	.48
K	COMP	.19	.39
L	NSQD	.26	.21
M	SP	.18	.27
N	NS	.04	.06
O	COMP	.09	.27
P	NSQD	.21	.42
Q	NSQD	.08	.24
R	COMP	.06	.18
S	NS	.00	.00
T	SP	.06	.18
U	NSQD	.03	.09
V	COMP	.02	.06
W	NS	.01	.00
X	SP	.04	.09
Y	SP	.02	.06
Z	NS	.00	.00
AA	COMP	.01	.03
AB	NSQD	.01	.03
AC	SP	.01	.03
AD	NS	.00	.00
AE	COMP	.02	.03
AF	NSQD	.00	.00

TABLE 9
MAP:M Quantitative Pilot Study Two: Third-Grade Difficulty and Discrimination
Indices.

Item Number	Item Type	Difficulty	Discrimination
A	NSQD	.97	.10
B	COMP	.80	.32
C	NS	.78	.31
D	SP	.99	.02
E	NSQD	.98	.05
F	COMP	.89	.22
G	NS	.84	.29
H	SP	.75	.41
I	SP	.80	.54
J	NS	.47	.51
K	COMP	.49	.61
L	NSQD	.37	.73
M	SP	.39	.60
N	NS	.11	.30
O	COMP	.21	.38
P	NSQD	.08	.20
Q	NSQD	.10	.30
R	COMP	.08	.25
S	NS	.04	.13
T	SP	.04	.13
U	NSQD	.04	.13
V	COMP	.04	.13
W	NS	.00	.00
X	SP	.01	.03
Y	SP	.00	.00
Z	NS	.00	.00
AA	COMP	.00	.00
AB	NSQD	.00	.00
AC	SP	.00	.00
AD	NS	.00	.00
AE	COMP	.01	.03
AF	NSQD	.01	.03

Table 10
Demographic Information of the Current Study

	N	Sex		Ethnicity				
		Male	Female	W	B	A	H	P
Grade 1	538	291(54%)	234(44%)	485(90%)	29(5%)	3(1%)	7(1%)	0(<1%)
Grade 2	567	298(53%)	257(45%)	521(92%)	26(5%)	1(<1%)	4(1%)	1(<1%)
Grade 3	583	307(53%)	258(44%)	540(93%)	14(2%)	2(<1%)	7(1%)	1(<1%)
Total	1688	895(54%)	749(46%)	1546(94%)	69(4%)	6(<1%)	18(1%)	2(<1%)

Table 11
Descriptive Statistics for MAP:M and MBSP

	Grade								
	First			Second			Third		
	n	M	SD	N	M	SD	n	M	SD
MAP:M									
Probe 1	495	12.42	7.01	519	20.52	8.61	513	28.70	9.13
Probe 2	491	12.10	7.03	526	19.84	9.28	537	30.92	11.86
Probe 3	486	14.99	8.12	520	21.58	9.81	525	35.57	13.40
Probe 4	499	12.54	6.29	509	21.70	11.50	531	33.90	12.18
Probe 5	485	16.26	7.57	519	25.22	11.50	535	37.11	12.75
Probe 6	491	18.56	7.76	526	24.58	12.13	509	36.91	12.58
Probe 7	491	19.55	8.58	516	25.97	13.72	536	41.25	14.99
Probe 8	450	20.89	8.48	484	31.33	15.99	517	41.79	15.00
Probe 9	486	20.55	9.62	520	32.06	14.33	528	41.12	14.03
Probe 10	481	18.94	8.81	509	32.42	14.84	540	38.67	15.06
Probe 11	425	21.27	9.79	459	30.58	16.22	454	42.37	15.73
Average	480	17.10	8.10	510	25.98	12.54	520	37.12	13.34
MBSP									
Probe 1	115	18.88	6.27	129	16.03	7.08	137	26.23	10.76
Probe 2	144	16.39	10.24	115	18.57	6.50	177	22.42	15.85
Probe 3	119	17.66	7.84	135	20.41	9.66	131	24.79	10.87
Average	126	17.64	8.12	126	18.34	7.75	148	24.48	12.50

*Table 12**Grade 1 Generalizability Study Results: Estimates of Variance Components*

Sources of Variation	Sums of Squares	df	Mean Squares	Estimated Variance Components	Percentage of Total Variance
MAP:M					
Person (p)	251464.01	537	468.28	45.19	.62
Item (i)	55663.90	9	6184.88	11.47	.16
Residual (pi,e)	79243.21	4833	16.40	16.40	.22
MBSP					
Person (p)	12978.51	102	127.24	36.74	.68
Item (i)	98.18	2	49.09	0.31	.01
Residual (pi,e)	3469.83	204	17.01	17.01	.31

Table 13
Grade 1 Decision Study Results

	Number of Probe Administrations					
	1	2	3	5	7	10
MAP:M						
α^2_{Rel}	16.40	8.20	5.47	3.28	2.34	1.64
α^2_{Abs}	27.87	13.93	9.29	5.57	3.98	2.79
p^2	.73	.85	.89	.93	.95	.97
Φ	.62	.76	.83	.89	.92	.94
SEM p^2	4.05	2.86	2.34	1.81	1.53	1.28
SEM Φ	5.28	3.73	3.05	2.36	1.99	1.67
MBSP						
α^2_{Rel}	17.01	8.50	5.67	3.40	2.43	1.70
α^2_{Abs}	17.32	8.66	5.77	3.46	2.47	1.73
p^2	.68	.81	.87	.92	.94	.96
Φ	.68	.81	.86	.91	.94	.96
SEM p^2	4.12	2.92	2.38	1.84	1.56	1.30
SEM Φ	4.16	2.94	2.40	1.86	1.57	1.32

NOTE: α^2 = Estimated Variance Component; p^2 = Coefficient of Generalizability; Φ = Index of Dependability

Table 14
Grade 2 Estimates of Variance Components

Sources of Variation	Sums of Squares	df	Mean Squares	Estimated Variance Components	Percentage of Total Variance
MAP:M					
Person (p)	586902.02	566	1036.93	98.71	.58
Item (i)	116722.02	9	12969.11	22.76	.13
Residual (pi,e)	253868.35	5094	49.84	49.84	.29
MBSP					
Person (p)	11649.93	105	110.95	25.03	.38
Item (i)	1009.19	2	504.59	4.42	.07
Residual (pi,e)	7528.81	210	35.85	35.85	.55

Table 15

Grade 2 Decision Study Results: Monitoring Academic Progress: Mathematics

	Number of Probe Administrations					
	1	2	3	5	7	10
MAP:M						
α^2_{Rel}	49.84	24.92	16.61	9.97	7.12	4.98
α^2_{Abs}	72.62	36.31	24.21	14.52	10.38	7.26
p^2	.67	.80	.86	.91	.93	.95
Φ	.58	.73	.80	.87	.90	.93
SEM p^2	7.06	4.99	4.08	3.16	2.67	2.23
SEM Φ	8.52	6.03	4.92	3.81	3.22	2.69
MBSP						
α^2_{Rel}	35.85	17.93	11.95	7.17	5.12	3.59
α^2_{Abs}	40.27	20.14	13.43	8.06	5.75	4.03
p^2	.41	.58	.68	.78	.83	.88
Φ	.38	.55	.65	.76	.81	.86
SEM p^2	5.99	4.23	3.46	2.68	2.26	1.89
SEM Φ	6.35	4.49	3.66	2.84	2.40	2.01

NOTE: α^2 = Estimated Variance Component; p^2 = Coefficient of Generalizability; Φ = Index of Dependability

Table 16
Grade 3 Estimates of Variance Components

Sources of Variation	Sums of Squares	df	Mean Squares	Estimated Variance Components	Percentage of Total Variance
MAP:M					
Person (p)	785331.78	582	1349.37	130.36	.66
Item (i)	1111467.5	9	12385.29	21.17	.11
Residual (pi,e)	239931.45	5238	45.81	45.80	.23
MBSP					
Person (p)	36721.53	120	306.01	89.26	.69
Item (i)	447.32	2	223.66	1.53	.01
Residual (pi,e)	9175.34	240	38.23	38.23	.30

Table 17

Grade 3 Decision Study Results: Monitoring Academic Progress: Mathematics

	Number of Probe Administrations					
	1	2	3	5	7	10
MAP:M						
α^2_{Rel}	45.81	22.90	15.27	9.16	6.54	4.58
α^2_{Abs}	66.97	33.49	22.32	13.39	9.57	6.70
p^2	.74	.85	.90	.93	.95	.97
Φ	.66	.80	.85	.91	.93	.95
SEM p^2	6.78	4.79	3.91	3.03	2.56	2.23
SEM Φ	8.17	5.77	4.71	3.65	3.09	2.58
MBSP						
α^2_{Rel}	38.23	19.12	12.74	7.65	5.46	3.82
α^2_{Abs}	39.76	19.88	13.25	7.95	5.68	3.98
p^2	.70	.82	.88	.92	.94	.96
Φ	.69	.82	.87	.92	.94	.96
SEM p^2	6.18	4.37	3.57	2.77	2.34	1.95
SEM Φ	6.31	4.46	3.64	2.82	2.38	1.99

NOTE: α^2 = Estimated Variance Component; p^2 = Coefficient of Generalizability; Φ = Index of Dependability

Table 18
Alternate Form Reliability: MAP:M

	2	3	4	5	6	7	8	9	10	11
Grade 1										
Probe 1	.77	.75	.70	.74	.70	.67	.68	.65	.65	.69
Probe 2		.80	.71	.75	.70	.70	.69	.66	.68	.71
Probe 3			.74	.75	.74	.72	.72	.68	.68	.72
Probe 4				.78	.72	.70	.68	.67	.68	.69
Probe 5					.77	.73	.74	.71	.70	.74
Probe 6						.76	.74	.71	.71	.75
Probe 7							.79	.80	.76	.76
Probe 8								.80	.77	.79
Probe 9									.80	.79
Probe 10										.83
Grade 2										
Probe 1	.66	.67	.58	.61	.60	.62	.59	.60	.53	.54
Probe 2		.69	.63	.70	.64	.68	.59	.60	.58	.56
Probe 3			.66	.69	.64	.65	.56	.60	.54	.51
Probe 4				.75	.69	.67	.60	.62	.57	.55
Probe 5					.73	.77	.65	.72	.66	.62
Probe 6						.76	.68	.71	.70	.68
Probe 7							.75	.77	.74	.69
Probe 8								.75	.75	.70
Probe 9									.79	.74
Probe 10										.76
Grade 3										
Probe 1	.71	.70	.65	.65	.61	.65	.61	.60	.59	.60
Probe 2		.73	.74	.71	.67	.70	.69	.65	.69	.69
Probe 3			.73	.76	.70	.70	.65	.68	.71	.66
Probe 4				.77	.73	.74	.71	.68	.73	.71
Probe 5					.76	.73	.71	.71	.74	.71
Probe 6						.78	.73	.72	.75	.75
Probe 7							.79	.78	.79	.79
Probe 8								.78	.79	.79
Probe 9									.76	.80
Probe 10										.80

Table 19

Alternate Form Reliability: Monitoring Basic Skills Progress

	Probe 2	Probe 3
Grade 1		
Probe 1	.74	.67
Probe 2		.69
Grade 2		
Probe 1	.37	.39
Probe 2		.50
Grade 3		
Probe 1	.74	.65
Probe 2		.73

Table 20
Alternate Form Reliability Correlation Coefficient Comparisons

	n_{avg}	r_{avg}	z	p
Grade 1				
MAP:M	443.55	.73	0.57	.57
MBSP	112	.70		
Grade 2				
MAP:M	470.55	.66	3.29	< .01
MBSP	116.33	.42		
Grade 3				
MAP:M	478.93	.71	0	1
MBSP	132	.71		

Table 21

Test-Retest Reliability : MAP:M

	n	r	<i>p</i>
Grade 1	73	.81	< .01
Grade 2	66	.79	.02
Grade 3	97	.76	< .01

*Table 22**Slope and Standard Error of the Slope by Grade for the MAP:M and MBSP*

	N	b	SEb
Overall MAP:M			
Grade 1	538	1.00	.04
Grade 2	567	1.40	.05
Grade 3	583	1.20	.06
MAP:M Matching Sample			
Grade 1	126	0.63	.29
Grade 2	135	1.58	.41
Grade 3	143	-0.29	.48
MBSP			
Grade 1	126	-0.20	.24
Grade 2	135	1.07	.25
Grade 3	143	-0.27	.35

Table 23
Concurrent Validity: Monitoring Academic Progress and Monitoring Basic Skills Progress

	MAP:M Probe		
	7	9	11
MBSP: Grade 1			
Probe 1	.73		
Probe 2		.58	
Probe 3			.66
Grade 2			
Probe 1	.37		
Probe 2		.52	
Probe 3			.41
Grade 3			
Probe 1	.60		
Probe 2		.51	
Probe 3			.74

Table 24

Stepwise Multiple Regression Analysis Predicting STAR Math with MAP:M and MBSP Across Grades 2 and 3

Predictor	n	R ²	b	SEb	β	t	F	p
MAP:M P6	29	.33	.71	.20	.57	3.62	13.11	.00
Excluded MBSP	29				.26	1.16		.26

Table 25

Correlations of MAP:M Probe Seven, MBSP, and STAR Math Across Second and Third Grades

	n	MBSP	STAR
Grades 2 and 3	29		
MAP:M		.64	.52
MBSP			.53
Grade 2	13		
MAP:M		.38	.41
MBSP			.47
Grade 3	16		
MAP:M		.61	.65
MBSP			.69

Appendix B

First Grade Practice Instructions

Guided Practice

* **NOTE:** We will need student names (first and last), date, teacher's name, grade, and school. Assessments can be bundled by class with a cover sheet that gives this information so it doesn't have to be written on each child's form each time. But be sure students' first and last names are clearly written on each guided/independent practice (G/I) and practice probe form. This is critical because some classes will have two or more students with same first name. We'll troubleshoot the most efficient way to do this before we scale up for the whole district. When administering the G/I practice and practice probes, first hand out the G/I practice forms, work through them, collect them, and then hand out the probe practice forms. Throughout all assessments, we need to counterbalance reading and math assessments. For the pilot, 2 classes per grade are participating, so give reading first in one class, math first in the other per grade. Make a note (based on teacher name on cover sheet mentioned above) whether reading or math was done first: R1, M2 or M1, R2.

[Before handing out math probes, the teacher writes the $>$, $=$, and $<$ symbols on the board with the words greater, equal and less beside the appropriate word. This cue will remain on the board during all practice and probes.]

"Today you will {also} be doing some math. First we are going to practice."

[HAND OUT GUIDED/INDEPENDENT PRACTICE]

"Look at the top of the page. There are two hands **[POINT TO SECTION]**. We are all going to do these problems together."

- "Everyone put your finger on the cow **[POINT TO COW]**. Is 1 greater (more than) 2 **[POINT TO GREATER THAN]**; or is it equal (the same as) 2 **[POINT TO EQUAL]** or is it less (smaller) than 2 **[POINT TO LESS THAN]**. We know that 1 is not as much as 2, so we circle the symbol less than **[POINT TO SYMBOL ON BOARD AGAIN]**. Trace the circle around the less than symbol **[POINT TO SYMBOL]**. Any questions?"
- "Look at the next one **[POINT TO B]**. Is 2 greater (more than) 2 **[POINT TO GREATER THAN]**; or is it equal (the same as) 2 **[POINT TO EQUAL]** or is it less (smaller) than 2 **[POINT TO LESS THAN]**. We know that 2 is the same as 2, so we circle the symbol greater than **[POINT TO SYMBOL ON BOARD AGAIN]**. Everyone circle the equal to symbol. **[POINT TO SYMBOL ON THE PAGE]**.
- "Look at the next one **[POINT TO C]**. Everyone do this one on your own." **[WHEN EVERYONE HAS FINISHED]** "Everyone should have circled the greater than symbol **[CIRCLE GREATER THAN SYMBOL, POINT TO SYMBOL ON PAGE]** because we know that 2 is greater (more than) 1.
- "Put your finger on the duck. **[POINT TO DUCK]**. Fill in the missing numbers that complete the number sequence. The first one says 1, 2, 3, BLANK, 5. The numbers go up by 1. We count by 1's so 4 goes in the blank space **[POINT]**. Everyone trace the 4. Any questions?"
- "Look at the next one **[POINT TO E]**. This one is counting by 1's also. What number goes in the blank? 4, 5, 6 **[POINT]**, 7, 8. Write 6 in the blank.
- "Look at the next one **[POINT TO F]**. Everyone do this one on your own." **[WRITE A 6. WHEN EVERYONE HAS FINISHED]**. "Everyone should have written a 6 in the blank. 2, 3, 4, 5, 6 **[POINT TO 6]**.
- "Put your finger on the fish **[POINT TO FISH]**. Circle the shape that completes the sequence. The pattern begins with a triangle **[POINT]** followed by a square **[POINT]**. This pattern repeats itself. Triangle, square, triangle, square, triangle, square, triangle, square, BLANK. What shape goes in the blank? A triangle goes in the blank because it comes after square. Everyone trace the circle around the triangle **[POINT]**. Any questions?"
- "Look at the next one **[POINT TO H]**. This sequence starts with a square then a circle then a square then a circle. What shape should go in the blank? A square comes after a circle. Everyone circle the square from the shapes on the right **[POINT TO SHAPES]**."
- "Look at the next one **[POINT TO I]**. Everyone do this one on your own." **[CIRCLE THE CIRCLE. WHEN EVERYONE HAS FINISHED]** "Everyone should have circled the circle **[POINT TO THE CIRCLE]** because the circle comes after the square.
- "Put your finger on the dog **[POINT TO DOG]**. These are addition and subtraction problems. You will add or subtract the two numbers, then write your answer in the space **[POINT TO SPACE]**. $1 + 1$ is 2. Everyone trace the number 2. Any questions?"
- "Look at the next one **[POINT TO K]**. $2 + 1$ is 3. Everyone write 3 in the blank."
- "Look at the next one **[POINT TO L]**. Everyone do this one on your own." **[WRITE A 4 IN THE BLANK, WHEN EVERYONE HAS FINISHED]** "Everyone should have written a 4 in the blank **[POINT TO THE 4]**."

Administration Directions

Pilot Probe Practice **(After administering and collecting Guided/Independent Practice).**

- ☐ 1. Say these specific directions:

“ Today you will be doing some math like we just practiced. Some will be easy and some will be hard. Most people do not get all of the correct answers or finish reading. When I say begin you will start working. Everyone point to the cow. Begin with this problem , then do all of the problems under the cow [RUN FINGER DOWN COLUMN], then go to the problems under the fish [RUN FINGER DOWN SECOND COLUMN]. If you come to a problem that is too hard, you can skip it. Try your best but do not spend too much time on one problem. Any questions?

- Begin

- ☐ 2. **Start your stopwatch. This practice is for pilot information only, Make a note of how long it takes most students to finish the page of pilot practice probes. Make notes about any problems, apparent sources of confusion, etc.**

[COLLECT PRACTICE FORMS]



A

1 $>$ 2

$=$

$<$



G

$\triangle$ $\square$ $\triangle$ $\square$ $\triangle$ $\square$ $\triangle$ $\square$ _

Circle One

$\square$ $\hexagon$

$\bigcirc$ $\triangle$

B

2 $>$ 2

$=$

$<$

H

$\square$ $\bigcirc$ $\square$ $\bigcirc$ $\square$ $\bigcirc$ $\square$ $\bigcirc$ _

Circle One

$\square$ $\hexagon$

$\bigcirc$ $\pentagon$

C

2 $>$ 1

$=$

$<$

I

$\bigcirc$ $\square$ $\bigcirc$ $\square$ $\bigcirc$ $\square$ $\bigcirc$ $\square$ _

Circle One

$\square$ $\hexagon$

$\bigcirc$ $\pentagon$



D

1, 2, 3, 4, 5



J

$1 + 1 =$ 2

E

4, 5, __, 7, 8

K

$2 + 1 =$ __

F

2, 3, 4, 5, __

L

$3 + 1 =$ __

Second/Third Grade Practice Instructions

Guided Practice

* **NOTE:** We will need student names (first and last), date, teacher's name, grade, and school. Assessments can be bundled by class with a cover sheet that gives this information so it doesn't have to be written on each child's form each time. But be sure students' first and last names are clearly written on each guided/independent practice (G/I) and practice probe form. This is critical because some classes will have two or more students with same first name. We'll troubleshoot the most efficient way to do this before we scale up for the whole district. When administering the G/I practice and practice probes, first hand out the G/I practice forms, work through them, collect them, and then hand out the probe practice forms. Throughout all assessments, we need to counterbalance reading and math assessments. For the pilot, 2 classes per grade are participating, so give reading first in one class, math first in the other per grade. Make a note (based on teacher name on cover sheet mentioned above) whether reading or math was done first: R1, M2 or M1, R2.

[Before handing out math probes, the teacher writes the $>$, $=$, and $<$ symbols on the board with the words greater, equal and less beside the appropriate word. This cue will remain on the board during all practice and probes.]

"Today you will {also} be doing some math. First we are going to practice."

[HAND OUT GUIDED/INDEPENDENT PRACTICE]

"Look at the top of the page. There are two hands **[POINT TO SECTION]**. We are all going to do these problems together."

- "Put your finger on the cow. **[POINT TO COW]**. Fill in the missing numbers that complete the number sequence. The first one says 1, 2, BLANK, BLANK, 5. We count by 1's so 3 goes in the first blank and 4 in the second **[POINT]**. Everyone trace the 3 and the 4. Any questions?"
- "Look at the next one **[POINT TO E]**. This one is counting by 1's also. What number goes in the blank? 4, BLANK, BLANK **[POINT]**, 7, 8. Write 5 in the first blank and 6 in the second blank.
- "Look at the next one **[POINT TO F]**. Everyone do this one on your own." **[WRITE 5 & 6. WHEN EVERYONE HAS FINISHED]**. "Everyone should have written 5 in the first blank and 6 in the second blank. 2, 3, 4, 5, 6 **[POINT TO 5 & 6]**."
- "Put your finger on the duck **[POINT TO DUCK]**. This problem has 2 parts. First solve the addition or subtraction problem, then tell whether your answer is less than, equal to, or greater than the number to the right **[POINT]**. In S3, **[POINT]** first solve the problem, $2 + 1$. The answer to $2 + 1$ is 3. Trace the number 3 in the blank **[POINT]**. Next compare your answer (3) **[POINT]** to the number to the right of the symbols **[POINT]**. Is 3 less than, equal to, or greater than 4? 3 is less than 4. Trace the circle around the less than sign. Any questions?"
- "Look at the next problem **[POINT TO E]** $2 + 2$ is 4. Write 4 in the blank. Our answer 4 **[POINT TO 4]** is equal to the 4 on the right **[POINT TO 4]**. Circle the equal to sign **[POINT TO SIGN]**."
- "Look at the next problem **[POINT TO F]**. Everyone do this problem on your own." **[WRITE 2 IN BLANK AND CIRCLE GREATER THAN, WHEN EVERYONE HAS FINISHED]** "Everyone should have written a 2 in the blank **[POINT TO 2]** and circled the greater than sign **[POINT TO SIGN]**."
- "Put your finger on the fish **[POINT TO FISH]**. Circle the shape that completes the sequence. The pattern begins with a triangle **[POINT]** followed by a square **[POINT]**. This pattern repeats itself. Triangle, square, triangle, square, triangle, square, triangle, square, BLANK. What shape goes in the blank? A triangle goes in the blank because it comes after square. Everyone trace the circle around the triangle **[POINT]**. Any questions?"
- "Look at the next one **[POINT TO H]**. This sequence starts with a square then a circle then a square then a circle. What shape should go in the blank? A square comes after a circle. Everyone circle the square from the shapes on the right **[POINT TO SHAPES]**."
- "Look at the next one **[POINT TO I]**. Everyone do this one on your own." **[CIRCLE THE CIRCLE. WHEN EVERYONE HAS FINISHED]** "Everyone should have circled the circle **[POINT TO THE CIRCLE]** because the circle comes after the square."
- "Put your finger on the dog **[POINT TO DOG]**. These are addition and subtraction problems. You will add or subtract the numbers, then write your answer below the line **[POINT TO SPACE]**. $111 + 111$ equals 222. Everyone trace the two's in the answer. Any questions?"
- "Look at the next one **[POINT TO K]**. $21 + 12$ is 33. Everyone write 33 below the line."
- "Look at the next one **[POINT TO L]**. Everyone do this one on your own." **[WRITE A 4 IN THE BLANK, WHEN EVERYONE HAS FINISHED]** "Everyone should have written a 342 below the line **[POINT TO THE 4]**."



A

1, 2, 3, 4, 5



G

△ □ △ □ △ □ △ □

Circle One

☐ ☐
☐ ☒

B

4, , , 7, 8

H

□ ○ □ ○ □ ○ □ ○

Circle One

☐ ☐
☐ ☐

C

2, 3, 4, ,

I

○ □ ○ □ ○ □ ○ □

Circle One

☐ ☐
☐ ☐



D

$2 + 1 = \underline{\quad}$

☒ $<$
 $=$
 $>$

4



J

$$\begin{array}{r} 111 \\ + 111 \\ \hline 222 \end{array}$$

E

$2 + 2 = \underline{\quad}$

$<$
 $=$
 $>$

4

K

$$\begin{array}{r} 21 \\ + 12 \\ \hline \end{array}$$

F

$1 + 1 = \underline{\quad}$

$<$
 $=$
 $>$

1

L

$$\begin{array}{r} 121 \\ + 221 \\ \hline \end{array}$$

Bi-Weekly Administration & Scoring Directions
Administration Directions

- ☐ 1. Make sure practice worksheets have been collected.
- ☐ 2. Distribute copies of MAP: Mathematics. Be sure students have the correct side up.
- ☐ 3. Instruct the students to write their first and last name on their paper.
- ☐ 4. Write the date on the board (using slashes mm/dd/yy). Instruct students to write the date on their paper
- ☐ 5. Say these specific directions:

“Today you will be doing some math {like we just practiced}. Some problems make you write the answer to a math problem and then compare it to another number by circling the correct symbol **[POINT TO A]**. You will do some arithmetic and write your answer below the line **[POINT TO B]**. You will write the missing numbers on the line to complete the number sequence **[POINT TO C]**. Finally, you will circle the missing shape from the choices given **[POINT TO D]**. Some will be easy and some will be hard. Most people will not get all of the correct answers or finish. Everyone point to the cow. When I say begin, start here with A **[POINT TO A]**, do all of the problems under the cow **[RUN FINGER DOWN COLUMN]**, then go to the problems under the fish **[RUN FINGER DOWN SECOND COLUMN]**. If you finish the first page, turn your paper over to the back and continue working. Be sure to work the problems you know and skip those you do not know. If you come to a problem that is too hard, skip it and go to the next. Try your best but do not spend too much time on one problem. We will do this for three minutes. If you change your mind, do not erase. Just make an X on the wrong answer and keep working. Any questions?”

Begin

- ☐ 6. **Start your stop watch.**
- ☐ 7. Observe students to ensure they understand the directions.
 - a. Are they actively engaged?
 - b. Are they writing answers in the correct places?
 - c. Do they have the correct side of the paper up?
- ☐ 8. If a student is not actively engaged, say **“Remember to skip the problems you do not know and do the problems you do know”**
If a student asks for help, say **“Do your best, if you do not know one go to the next one”**
- ☐ 9. **Stop the stopwatch at 3 minutes: “Stop, turn your papers face down. Put your pencil on your desk.”**
- ☐ 10. Collect the papers.

Scoring Directions

- 1. Check a box for each part of the problem that is correct. The number of boxes beside each problem indicates the total number of elements in that item. ☒
- 2. If the student did not mark an answer or responded incorrectly then leave the box blank ☐
- 3. Count the number of check marks on each page and write in the space provided at the bottom of the page
- 4. Add the total number of checkmarks and write the score in the “SCORE” at the top of the page.
- 5. Record all student scores on the Record Form Sheet



Name: _____



Date: _____

A 12 9 14	I <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> Circle One ☐ ☐ ☐ ☐
B 43 + 56	J 2, , 12, 17, 22
C 58, 57, 56, , 54	K 37 - 23
D <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> Circle One ☐ ☐ ☐ ☐	L 70 66 56
E 26 36 40	M <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> Circle One ☐ ☐ ☐ ☐
F 3 - 2 =	N , 59, 61, 63, 65
G , 13, 23, 33, 43	O 9 + 7 =
H <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> <u> </u> Circle One ☐ ☐ ☐ ☐	P 93 85 83



Name: _____



Date: _____

A

$$5 + 7 = \underline{\quad}$$

$<$
 $=$
 $>$

 12

I

Circle One

$\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$

$\square$ $\hexagon$
 $\bigcirc$ $\triangle$

B

$$\begin{array}{r} 857 \\ - 332 \\ \hline \end{array}$$

J

$$\underline{269}, \underline{271}, \underline{\quad}, \underline{\quad}, \underline{277}$$

C

$$\underline{691}, \underline{695}, \underline{\quad}, \underline{\quad}, \underline{707}$$

K

$$\begin{array}{r} 59 \\ - 37 \\ \hline \end{array}$$

D

Circle One

$\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$

$\square$ $\hexagon$
 $\bigcirc$ $\triangle$

L

$$9 - 2 = \underline{\quad}$$

$<$
 $=$
 $>$

 4

E

$$9 - 6 = \underline{\quad}$$

$<$
 $=$
 $>$

 3

M

Circle One

$\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$

$\square$ $\hexagon$
 $\bigcirc$ $\triangle$

F

$$\begin{array}{r} 32 \\ + 24 \\ \hline \end{array}$$

N

$$\underline{644}, \underline{\quad}, \underline{\quad}, \underline{653}, \underline{656}$$

G

$$\underline{673}, \underline{\quad}, \underline{\quad}, \underline{703}, \underline{713}$$

O

$$\begin{array}{r} 453 \\ + 723 \\ \hline \end{array}$$

H

Circle One

$\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$ $\underline{\quad}$

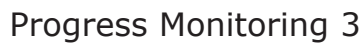
$\square$ $\hexagon$
 $\bigcirc$ $\triangle$

P

$$9 + 5 = \underline{\quad}$$

$<$
 $=$
 $>$

 15



$$\overline{\text{MC}} + \overline{\text{MR}} = \overline{\text{Total}}$$


$$7 + 8 = \underline{\quad} = 4$$
$$\forall \Rightarrow \forall$$
$$\begin{array}{r} 886 \\ - 253 \\ \hline \end{array}$$

331, 343, ____, ____, 379

137, 237, ____, ____, 537

$$\begin{array}{r} 79 \\ - 37 \\ \hline \end{array}$$
$$6 \times 3 = \underline{\quad} = 18$$

$\forall$
 $\parallel$
 $\forall$

$$5 - 3 = \underline{\quad} = 10$$
$$\begin{array}{c} \vee \\ || \\ \vee \end{array}$$
$$\begin{array}{r} 56 \\ + 83 \\ \hline \end{array}$$

775, , , 850, 875

453, , , 468, 473

$$\begin{array}{r} 752 \\ + 937 \\ \hline \end{array}$$
$$8 \div 2 = \underline{\quad} = \boxed{2}$$

∇
 $\parallel$
 ∇

MAP: M
Scoring Directions

A

1

2

3

- » Must circle correct answer
- » Each problem has one potentially correct answer
- » If correct check the box beside the problem

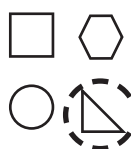
B

1, 2, 3, 4, 5

- » Must write correct number or numbers in blank
- » Transposed numbers are counted as correct
- » Double digit numbers have a potential score of 2, triple digit numbers have potential score of 3
- » Check one box correct for each number written in the correct place.
- » Example: If correct answer is 954 and student writes 924 they will receive credit for the 9 and 4.

C

Circle One



- » Must CIRCLE the correct shape
- » Shapes drawn in the blank are not counted as correct
- » If correct shape has been circled, check the box beside the problem

» SCORING TIP

- » AT THE END OF EACH COLUMN, TOTAL THE NUMBER OF MATH CALCULATION (MC) AND MATH REASONING (MR) PROBLEMS CORRECT AND RECORD THE SCORE AT THE BOTTOM

MAP: M
Scoring Directions

D

$$2 + 1 = \underline{\quad} \quad \begin{array}{|c|} \hline < \\ \hline = \\ \hline > \\ \hline \end{array} \quad 4$$

- » Must write a number in the blank, THEN circle the correct sign
- » Transposed numbers are counted as correct
- » Signs circled in the absence of a number written in the blank is not correct
- » If arithmetic is incorrect but the correct symbol is circled comparing the answer to the number then credit 1 checkbox.
- » Two digit answers have potentially 3 correct responses (2 digits, correct symbol circled)
- » Check the boxes next to the problem to indicate the total number of correct responses

E

$$\begin{array}{r} 111 \\ + 111 \\ \hline 222 \end{array}$$

- » Must write answer below the line
- » Transposed numbers are counted as correct
- » Two digit answers have 2 correct responses
- » Three digit problems have 3 potentially correct responses
- » Any number written in the correct position (i.e., ones, tens, hundreds) is counted as a correct response irrespective of the answer as a whole.
- » Credit is given for 0's or empty spaces left in a subtraction problem that results in a 0 in left most position (e.x., $15-14 = 1$; this answer has 2 correct responses, 01 or 1, both answers receive 2 points because the 0 in the second answer is inferred.
- » Check the boxes next to the problem to indicate the total number of correct responses

GENERAL DIRECTIONS

1. Checkbox one box to the right of the problem for each element correct. The number of boxes beside each problem indicates the total number of elements in that item.
2. If the student did not mark an answer or responded incorrectly then leave the box blank
3. Written answers are given credit for the correct numbers written in the correct position (i.e., ones, tens, hundreds). Example: If the correct answer were 92 and the student writes 192, they receive full 2 point credit for having the 9 and the 2 in the correct position.
4. Be SURE to check ALL parts of ALL problems for student responses, remember they can skip around, even to the back page.
5. Count the number of check marks for MC and MR on each page and record in the designated position.
6. Sum the MC and MR scores and record at the top of page 2.
7. Sum the MC and MR scores for the Total Score.
7. Record all student scores on the Record Form Sheet

Vita

Michael Brandon Hopkins was born in Marion, VA on June 11, 1981. He graduated from Chilhowie High School in 1999. Subsequently he graduated from Radford University with a Bachelor of Science in psychology and coursework in business management. Michael then attended Richmond Graduate University (aka. Psychological Studies Institute) where he earned a Master of Arts degree in Marriage and Family Therapy with specializations in both Child and Adolescent Therapy and Christian Sex Therapy. He moved to Knoxville, TN in 2007 where he began doctoral work in School Psychology at the University of Tennessee.

Michael is currently completing a doctoral internship at the Onarga/Iroquois County site at the Illinois School Psychology Internship Consortium. He will graduate with a doctoral degree in School Psychology from the University of Tennessee in December 2011.