THE UTILITY OF MATHEMATICS CURRICULUM-BASED MEASUREMENT TO PREDICT STUDENT RISK STATUS ON STANDARDIZED ACADEMIC ACHIEVEMENT MEASURES

by

Kyle Max Hancock

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Psychology

Approved:

Donna M. Gilbertson, PhD Major Professor	Gretchen Gimpel Peacock, PhD Committee Member
Melanie Domenech-Rodriguez, PhD Committee Member	Renee V. Galliher, PhD Committee Member
Nancy K. Glomb, PhD Committee Member	Byron R. Burnham, EdD Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2009

ABSTRACT

The Utility of Mathematics Curriculum-Based Measurement
to Predict Student Risk Status on Standardized
Academic Achievement Measures

by

Kyle Max Hancock, Doctor of Philosophy

Utah State University, 2009

Major Professor: Dr. Donna M. Gilbertson
Department: Psychology

The predictive utility of mathematics curriculum-based measurement (MCBM) to identify students who are at risk for failure on important educational measures is an emerging area of study in need of further investigation. The present study sought to identify which of four MCBM probes could be accurately used to determine students' risk status on selected subtests of three important educational measures commonly used to make educational placement decisions (WIAT-II, WJ-ACH-III, and KM 3) in Grades 2 (n = 49), 4 (n = 48), and 6 (n = 47). The study also sought to determine *which type of student performance measurement strategy (i.e., level, slope, or dual discrepancy) on each of the four types of MCBM probes proved to be the best method to determine student risk status.* The results of the study indicated that the ability of the MCBM probes to identify students' risk status was generally poor. However, evidence indicated that MCBM probes could be used more reliably and accurately to determine students in

the low-risk category than those in the high-risk category across all probe types and administration times. Finally, the level method generated the greatest support and the slope method generated the least support for identification of high- and low-risk student status on each probe or combination of probes.

(176 pages)

DEDICATION

I would like to dedicate this project to my Grandma and Grandpa Herrick. To my grandmother because, as a small child, she taught me that empathy and kindness is God's way of letting us know He cares and because her life exemplified this unremittingly. To my grandfather because, on a warm summer day, he taught me that anything worth doing is worth doing well. Their examples permeate my memory; it is my ambition that they will also permeate my life.

ACKNOWLEDGMENTS

I would like to thank my committee members, Dr. Gretchen Peacock, Dr. Melanie Domenech-Rodriguez, Dr. Renee Galliher, Dr. Nancy Glomb, and Dr. Donna Gilbertson, for their meaningful feedback and guidance, and for all of their efforts in my behalf. I would particularly like to extend especial gratitude to Dr. Donna Gilbertson for her time, effort, feedback, guidance, and support–which was invaluable in the orchestration of this project and, without which, successful completion would not have been possible. Dr. Gilbertson epitomizes true mentorship and I am deeply grateful for her support throughout my academic career–without which, success would have been unattainable.

I would also like to thank my family–both the one into which I was born and the one into which I married–for their ever-enduring faith, love, and support in my behalf. In particular, I would like to thank my mother for teaching me to love others, laugh at myself, and live a light-hearted life. I would like to thank my father for teaching me that all things are possible through faith and hard work. Above all, I am grateful for their unconditional belief in me, for their ceaseless support, and for allowing me the opportunities I needed to learn and grow throughout my life.

Above all, to my wife, Stephanie, I express my eternal love and gratitude for her infinite friendship, understanding, and extraordinary support as we have worked to fulfill our dreams. I wish to recognize her active role throughout this project and the completion of my education and training; without her, it would have been utterly impossible and, more importantly, meaningless. She is, as always, my angel of hope.

Kyle Max Hancock

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

An estimated 70% of American children experience difficulty successfully learning and applying mathematics principles during their elementary school careers (Manzo & Galley, 2003). Lack of basic skills continues to influence performance on more complex math applications in later grades (Manzo & Galley). Identification of children who would benefit from academic intervention or special education services in the area of mathematics when difficulties first emerge is needed to prevent the development of severe learning difficulties in mathematics.

A screening assessment system used to identify children's response to a math curriculum before problems become severe requires an assessment instrument that can be frequently administered and is sensitive to short periods of growth. One such assessment system that has been suggested to identify children who are at high risk for academic concerns is curriculum-based measurement (CBM). Curriculum-based measurement involves the use of brief, 1-5 minute probes consisting of academic content drawn specifically from the curriculum that the student is expected to know over a certain period of time (Shinn, 1989). This format allows the tracking of student progress many times during a school year to monitor the efficacy of the general curriculum or an intervention curriculum. The validity and reliability of this system has been extensively documented in reading; validity in the areas of math and writing, although less extensive, has also been demonstrated (Shinn; Shinn, 1998).

Curriculum-based measurement focused on reading (RCBM) has been used extensively to identify children with reading problems who are likely to benefit from academic intervention or special education services; research has suggested that this can be done effectively using RCBM (Shinn, 1989, 1998). Student achievement outcomes improve when teachers use ICBM data to evaluate and modify their instruction (Fuchs, Deno, & Mirkin, 1984; Fuchs, Fuchs, & Hamlett, 1989a, 1989b, 1989c; Fuchs, Fuchs, Hamlett, & Allinder, 1991; Fuchs, Fuchs, Hamlett, & Ferguson, 1992; Fuchs, Fuchs, Hamlett, & Stecker, 1990; Stecker & Fuchs, 2000; Wesson, Deno, & Mirkin, 1988). Mathematics CBM (MCBM) has also been used to monitor progress on math calculations and math applications (Shinn, 1989), but few studies have been conducted that provide evidence for its ability to successfully and validly screen children who are at high risk for failing to succeed on important academic measures commonly used to make educational placement decisions (e.g., standardized norm-referenced tests [NRT]) or who achieve poorly due to a disability. Because many students are experiencing math difficulties, more research documenting the utility of MCBM in accomplishing these purposes is necessary.

Hence, this study extended the current research literature by examining the degree to which MCBM can be used to identify students who may require early or intensive intervention that might be provided through a school's supplemental intervention services. Given that the type or amount of math calculation skills surveyed on an MCBM probe may vary, probe content is likely to affect the accuracy with which the MCBM data can be used to identify high- and low-risk students. For example, a single skill MCBM probe contains a limited amount of skills assessed, whereas many NRTs often

assess multiple skills. Because of the disparity between the content assess by the two measures, using MCBM data to identify high- and low-risk students may result in a lower identification accuracy rate than what would be obtained using an NRT. Thus, one goal of this study was to explore the relationship between four types of MCBM probes and three NRTs frequently used to make decisions regarding students' educational placement (e.g., special education vs. regular education services) across several elementary school grades. There were four different types of survey level MCBM probes examined in this study. Specifically, probes consisted of a single skill that has previously been taught and students are expected to have mastered, a single skill that was to be acquired during the course of the study, multiple skills that were to be acquired over the course of the study, and an error analysis of multiple skill steps ranging from skills that have been mastered to skills that students were to acquire. All skills on each probe were math calculation skills only (e.g., addition, subtraction, multiplication) and did not represent math reasoning skills (e.g., rounding, estimating, story problems).

The final goal of this study was to evaluate the extent to which the MCBM probes accurately identified high- and low-risk students based on NRT scores varied as a function of the type of performance indicator used. The three performance indicators included in the study were: (a) static performance scores (i.e., level scores), (b) academic growth scores (i.e., slope scores), and dual discrepancy scores (i.e., students who were high risk on both level and slope scores).

The final goal of the study was to determine if combinations of MCBM probes added significantly to their ability to successfully predict student risk status based upon NRT performance.

CHAPTER II

REVIEW OF THE LITERATURE

Prevalence of Students with Math Problems and Outcomes

Research regarding academic assessment and intervention has gained substantial attention in the past decade, credited in part to a call for the demonstration of treatment utility and teacher accountability for adequate student progress as proposed in the ratification of the No Child Left Behind Act (NCLB; Daly & McCurdy, 2002). Largely as a result of this movement, there has been an abundance of research in the area of reading assessment, instruction, and intervention.

Research in the area of mathematics assessment and intervention has not received as much attention in the literature as reading (Badian, 1999; Daly & McCurdy, 2002). This is particularly problematic given the dismal outcomes of student performance on math proficiency tests. In a recent review of students' national test scores on yearly administered math proficiency tests, only 31% of all fourth-grade students scored at or above the proficiency standard (Manzo & Galley, 2003). In addition, several researchers have documented the importance of mathematics proficiency for subsequent successful employment as various occupations increasingly require employees to utilize mathematics (Saffer, 1999).

The lack of research in the area of mathematics assessment and intervention should not be attributed to a low prevalence of math disabilities. Several studies report the prevalence of math disability similar to rates reported for reading disability and

attention-deficit/hyperactivity disorder reporting a range of 3-8% of the school-age population (Jitendra, Sczesniak, & Deatline-Buchanan, 2005; Mazzocco & Myers, 2003; Shalev, 2004).

Many studies have indicated that early intervention has the potential to correct these deficits and to avoid any serious delays in academic progress (Hintze, Christ, & Keller, 2002). Without early identification of problems, low-achieving students continue to struggle over time until a special education evaluation is warranted or requested due to a substantial math performance gap relative to the students' peers. Remediation is costly in terms of time, finances, and personnel because severe math skill deficits are difficult to correct (Stecker & Fuchs, 2000). Frequent monitoring of progress for early identification of low achievers may lead to improved performance with lower remediation costs (Fletcher, Denton, & Francis, 2005).

Curriculum-Based Measurement as a Viable Solution

Methods of Assessing High-Risk Children Using CBM Data

Methods of assessing student performance within a CBM system consist of multiple approaches that primarily focus on three types: current status (or level), growth, or dual discrepancy (Fuchs, 2003). The current status (or level) approach provides an indication of a student's performance at the time the CBM is administered. This score can be compared to the student's standing relative to some norm (e.g., benchmarks, comparison to peers, or mastery level). However, when a child's current level is

assessed, information detailing why the child has performed at this level is not provided. The child's present level of performance could be due to multiple potential factors, including lack of motivation, poor instruction, or inadequate curriculum. According to IDEIA (Yell & Drasgow, 2007), these potential explanations must be excluded before the diagnosis of a learning disability can be provided.

The second method involves assessing students at different points over time during instruction or the course of an intervention in order to estimate student achievement growth (i.e., slope). The slope estimate provides an indication of how much the student has learned over the course of the intervention: that is, whether or not a student is learning as expected over time. This slope can also be compared to some standard to determine adequacy of growth (e.g., slopes of comparison peers). This score has the advantage of showing which students have the ability to obtain adequate growth that is similar to peers although the level of performance falls below peer current performance (level).

The final approach, dual discrepancy analysis, considers both the measurement of performance level and amount of growth. This final approach has been suggested as a viable method to provide meaningful and useful measurement of student performance within a CBM system because it provides a more thorough index of student performance (Fuchs, 2003). In addition, the combined information indicates how much a student has learned within a given time, if the student is demonstrating the ability to learn material when good instruction is given, what intensity of intervention is working (if adjustments to the initial intervention are implemented), and the discrepancy between the child's

current level and the expected level to gauge the amount of time and resources needed to obtain mastery or same-age peer level.

Advantages of a CBM System

Because of recent NCLB legislative requirements and a concern with the percentage of low student academic achievement, educators are increasingly seeking an effective and efficient tool that will improve the ability of teachers to monitor progress and to identify children who are at high risk for poor academic performance. CBM has been suggested as one viable method to be used as a screening device of academic underachievement by repeatedly sampling student proficiency on the school's curriculum at the student's instructional level on a frequent basis. Teachers can then use CBM to monitor student progress on reading and math probes over time. The primary advantages of CBM are that it can be used to quickly, accurately, and relatively easily identify students who are not performing adequately within their current curriculum (Deno, 2003). Moreover, CBM assessment validity and reliability have been well documented (Deno, 1985; Fletcher et al., 2005; Shinn, 1989, 1998).

Curriculum-based measurement probes are created by sampling all information a student should learn within a specified period of time. For math, the teacher selects from the core curriculum a sample of problems that adequately represents the type of skills the student should learn within a predetermined period of time (Deno, 1985). Procedurally, students are presented with printed problems on a probe and are asked to complete as many problems are they can within a certain time limit (Shinn, 1989). The total number of problems on the probes should exceed that which could be completed by a student

who has mastered the skill within the predetermined time limit. Probes are then scored as the number of digits correct per minute and progress is monitored as change along this index. The data can be used to determine if most students are adequately learning the curriculum material, if some students need additional intervention support, or if intervention support is working.

Research Regarding Effectiveness of CBM

Speece and Case (2001) investigated the accuracy with which a dual discrepancy model will identify children with reading disabilities relative to an intelligence quotient (IQ)-achievement discrepancy model and a low-achievement model. The authors hypothesized that, given an instructional environment in which most children are adequately learning and implemented general education interventions for children who are faltering academically, children who continue to exhibit a discrepancy from their peers on both mean level of performance and rate of progress on academic skills would be selected as candidates for special education services (Fuchs & Fuchs, 1998). Participants were selected from 694 first- and second-grade students attending three schools from a suburban district. During late September and early October, all first-grade students were administered two letter sounds fluency (LSF) CBM probes and all second-grade students were administered two oral reading fluency (ORF) CBM probes. Two groups of students were selected for participation in the study from these scores: a high risk group and a comparison group. Students were considered high risk for reading failure if their mean performance on the CBM probes placed them within the lowest 25% within their classroom. The comparison group was created by selecting 2 students

scoring at the median and 1 student at the 30th, 75th, and 90th percentiles from each classroom; students were selected from these cutpoints to ensure a wide range of skill. All children who had active IEPs and designated as English language learners without sufficient language proficiency were eliminated from the study. The final sample included 144 high-risk students and 129 comparison peers.

All second-grade participants were administered 20 ORF probes between November and May; all first-grade participants were administered 15 ORF probes between January and May (Speece & Case, 2001). In addition, the information, similarities, block design, and digit span subtests from the Wechsler Intelligence Scale for Children–Revised (WISC-R; Wechsler, 1974) were administered to all participants and a full scale IQ (FSIQ) was estimated based upon the formula provided by Sattler (1988). The reading achievement of all participants was assessed through the Letter-word identification and word attack subtests of the Woodcock-Johnson Tests of Achievement–Revised (WJ-R; Woodcock & Johnson, 1989).

Using the data obtained from these measures, students in the high-risk group were then placed into one of three subgroups: CBM dual discrepancy (CBM-DD), regression-based IQ-achievement discrepancy (IQ-DS), and low achievement (LA). Children (n = 84) were placed into the CBM-DD subgroup if his or her slope across the year and level of performance at the end of the year were both more than 1 standard deviation (SD) below the slope and level of their classmates on at least 10 CBM probes administered across the school year. Membership in the IQ-DS group was determined through the regression of the WJ-R Basic Reading Cluster scores on FSIQ scores; students whose actual achievement differed by 1.5 or more standard errors of prediction were placed in

this group (n = 17). Low reading achievement was defined as a standard score less than 90 on the WJ-R Basic Reading Skill Cluster (n = 28).

Evidence for the construct validity of the dual discrepancy approach was generated, indicating it is a potentially valid method to identify students with reading disability as evidenced by the following: (a) students in the CBM-DD group showed deficits on the majority of the dependent measures above and beyond those in the IQ-DS or LA groups, (b) teachers rated the students in the CBM-DD group as less academically competent, and, finally, (c) single-point measurements did not accurately identify the CBM-DD group or any of the poor readers. These results suggested that the dual discrepancy may be a more accurate and more adequate measure for identifying students with reading difficulties than the IQ-DS or LA models.

Results also revealed that there was no gender disproportion in the CBM-DD subgroup of high-risk children. In addition, the mean age of the CBM-DD subgroup was significantly younger than the other subgroups of high-risk students (ES = -1.16 for CBM-DD vs. IQ-DS and -.69 for CBM-DD vs. LA), thus indicating that a dual discrepancy model has the potential to identify students earlier than the other models and, thereby, has the potential to foster earlier intervention efforts. Finally, the racial distribution of the CBM-DD subgroup closely approximated the population from which it was taken.

Thus, Speece and Case (2001) concluded that a dual discrepancy model has the potential ability to identify students struggling with reading earlier and more effectively than other types of identification procedures. Given that many children are also

experiencing math difficulties, similar research on accuracy of the dual discrepancy model for identification of high-risk students in math is warranted.

Mathematics Curriculum-Based Measurement

Utility for Instructional Effects and Decision Making

Although much research has documented the psychometric qualities of RCBM, much less research has been conducted in the area of mathematics. What research that has been conducted on MCBM has primarily evaluated its use to improve individual or class-wide learning plans. A review of the utility of MCBM to improve student achievement between 1980 and 2005 was completed by Stecker, Fuchs, and Fuchs (2005). The authors identified five studies that investigated the relationship between implementing MCBM and subsequent improvement in math achievement relative to the math achievement of students in classes that did not use CBM. In all five studies, general education or special education teachers included in the CBM-trained group were asked to administer CBM probes weekly and review the data with consultants in order to make instructional changes at both the class and individual levels. Results indicated that students within the experimental CBM groups outperformed students in the control groups (i.e., no CBM training) on CBM probes at the end of the study. Four of the five studies reviewed provided evidence that MCBM can be used to improve academic performance among elementary students with and without learning disabilities when teachers utilize the MCBM information to make instructional changes.

Utility of Skills Analysis for Instruction

Information from MCBM to analyze performance on specific types of problems has also provided meaningful and useful information in identifying specific skill sets to teach or re-teach among individual students. A skills analysis approach reviews student answers on CBM probes to determine which skills presented on the probe require remediation efforts. This is possible when a CBM probe consists of problems requiring a variety of grade-level mathematics skill ability and displays the same types of problems in the same proportion on alternate test forms given multiple times during the school year.

Fuchs and colleagues (1990) investigated the effects of a skills-analysis approach within a CBM system on math performance. Teachers participated in this study with a total of 91 students in grades three through nine classified as either learning disabled or emotionally disturbed, each of whom had a current IEP with mathematics goals. Thirty special educators were randomly assigned to three groups within the study: (a) CBM monitoring with graphed math performance score on the CBM probe and skills analysis, (b) CBM monitoring with graphed total scores only, and (c) no systematic performance monitoring system (i.e., control). All students in the treatment conditions were administered a grade-level CBM probe twice weekly for 15 weeks. CBM probes consisted of 25 grade-level mathematics calculations problems based upon the state curriculum. The skill analysis consisted of a computer program that reported low performance on specific problems on the CBM probe to help teachers identify specific skills for which students require remediation.

Results indicated that students in the skills analysis condition significantly outperformed students in both the CBM-only condition and control condition on digits correct per minute ($p < .05$) and on a state-wide math assessment. Although teacher remediation activities were not monitored, the authors concluded that the inclusion of skills analysis may have assisted in instructional decision making. However, more research is needed to ascertain how the inclusion of skills analysis within a CBM system has the potential to benefit students relative to a CBM approach without a skills analysis component.

Reliability of MCBM

Data from several studies support the reliability of MCBM (e.g., Marston, 1989; Shinn, 1989, 1998). In general, the reliability of math probes has been reported with correlations in the $r = .90$ range across several studies (Marston). For example, a study by Tindal and Marston (1990) reported interrater reliability at $r = .97$, 1-week test-retest reliability at $r = .87$, and alternate form reliability at $r = .66$. Thurber, Shinn, and Smolkowski (2002) reported interrater reliability coefficients ranging from $r = .77$ to $r = .94$ with an average of $r = .87$. The same study produced support for alternate form reliability with a median correlation of $r = .91$ across three types of CBM probes (computation, application, and mixed probe types).

Construct Validity and Concurrent Validity of MCBM

Thurber and colleagues (2002) investigated the construct validity of MCBM with 207 fourth-grade students. In their study, confirmatory factor analysis procedures were

used to determine if MCBM measures computation, application of math principles, or both. Participants were administered an MCBM probe focused on math computation skills, two standardized math tests (math computation and math concepts and applications subtest of the California Achievement Test; CAT) and computation and applications subtest of the SDMT, and a reading maze probe. Approximately 75% of the probes consisted of basic multiplication and division facts with the remaining 25% representing basic addition and subtraction. Correlations ranged from $r = .36$ to $r = .63$ between the three MCBM probes and all of the achievement subtests. The results of model testing indicated that MCBM could be most accurately described as a measure of mathematics achievement (as characterized by the correlations generated on both the CAT and SDMT). In addition, data from this study provided construct evidence as to the types of math skills that MCBM measures. The results of confirmatory factor analysis indicated that MCBM can best be described as a two-factor model in which computation and applications are two separate, but highly related constructs with MCBM measuring math computation. Thurber and colleagues (2002) also reported a median correlation of $r = .82$ between computation MCBM and other measures of basic math facts (e.g., curriculum tests from math texts) and a median correlation of $r = .61$ with measures of math computation on the SDMT and CAT. Moreover, scores obtained from the reading mazes probes were also highly correlated with mathematics performance ($r = .76$).

In a review of approximately 80 studies of reading, writing, and mathematics CBM, Good and Jefferson (1998) discussed evidence that supports the concurrent validity of MCBM (i.e., correlation of two different math tests taken at the same time). Correlations reported in the reviewed studies between MCBM and various standardized

tests were reported at $r = .60$ or greater. In addition, construct validity was provided in a study by Shinn and Marston (1985), who found that MCBM results differentiated between students in general education, Title 1, and resource placement of students in Grades 5 and 6. Students with mild disabilities with lower MCBM scores were also distinguished from fourth grade general education.

MCBM as a Predictor of Performance on Norm-Referenced Testing

Norm-referenced tests are commonly administered to determine important placement decisions, such as grade advancement/retention and placement in general education. Traditionally, student performance is obtained on group-administered tests that are given to students one time each school year (or only on specific grades) to gauge adequate growth in curriculum compared to national or state norms. In addition, schools may use this data to determine which students are experiencing academic difficulties and require early and intensive intervention such as might be provided through the school's supplemental intervention or special education services (Klingner, Artiles, & Barletta, 2006). Commonly administered tests may include the CAT, the Iowa Test of Basic Skills (ITBS), and the Pennsylvania System of School Assessment (PSSA).

Individually administered NRTs are commonly used to provide diagnoses of learning disabilities. Commonly used tests include the Woodcock-Johnson Tests of Achievement–Third Edition (WJ-ACH-III), the Wechsler Individual Achievement Test–Two (WIAT-II), or the Key-Math Test. However, they are not without their weaknesses (Mazzocco, 2005; Mazzocco & Myers, 2003). First, these tests are often administered only after classroom mathematics performance has become sufficiently

severe to be detectable on standardized tests often consisting of a low number of math items. This is problematic because many studies have indicated that early intervention has the potential to correct learning deficits and to avoid any serious delays in academic success (Hintze et al., 2002). Another problem with this system is that different tests are used in different schools and districts to determine special education eligibility; this may seriously undermine the consistency with which disabilities are identified. Finally, these tests require substantial amounts of time to administer and score.

Investigations of the validity of MCBM as a more frequently administered screening instrument to identify students at risk for poor performance on NRTs are an emerging area of research on MCBM. Foegen and Deno (2001), for example, conducted a study within a middle-school population to determine if MCBM data could be used as indicators of mathematics proficiency. Their study expanded the research by examining the relationship between several types of MCBM probes with several criterion achievement measures. One hundred students in the seventh and eighth grades from an ethnically diverse middle school in an urban district participated in this study. Approximately 12% of the students were receiving special education services. Participants were administered one calculation-based, grade-level MCBM (the basic mathematics operation task; BMOT) probe on two occasions during a 1-week period in the spring. The BMOT was designed to index students' accuracy and fluency in mental computation of whole-number facts in addition, subtraction, multiplication, and division. Participants were given a 1-minute probe that contained 80 problems arranged randomly (20 single-digit computations for each mathematical operation—i.e., addition, subtraction, multiplication, and division). Participants were also administered an

application-based MCBM probe (basic mathematics estimation task, BMET) to index students' accuracy and fluency in the application of mathematics estimations skills.

The researchers also utilized test data collected by the school to serve as criterion variables to which the MCBM data could be compared. Specifically, MCBM scores were compared to participants' math grade point average (GPA), GPA for "core" classes, and standardized test scores from the CAT administered in the spring of the participants' sixth-grade year (i.e., 1 year prior to the study for the seventh-grade participants and 2 years prior to the study for the eighth-grade participants). Finally, teachers were asked to rate the students' performance in mathematics on a 5-point Likert scale on six dimensions of mathematics performance and abilities: (a) overall proficiency in mathematics, (b) value for mathematics, (c) confidence in his or her mathematics ability, (d) mathematical problem-solving ability, (e) mathematical communication ability, and (f) mathematical reasoning ability.

Using regression analyses, the researchers examined the efficacy of using the BMOT and BMET to predict standardized test scores and teacher ratings, to determine the degree to which each of the measures contributed uniquely to the prediction of the criterion variables. Results of multiple forced-entry regression analyses indicated that the single best predictor of the computations subtest of the CAT was the BMOT; it accounted for 63% of the variance and was statistically significant at $p < .01$. The BMOT was also the single best predictor of math GPA and overall GPA. Finally, the BMOT was also the single best predictor of teacher ratings of students' proficiency, confidence, problem solving, and reasoning. The authors reported that the BMOT was generally the single best predictor of performance on the CAT and the teacher rating

scale and accounted for the majority of the variance even in conjunction with the predictor variables (i.e., the BMET).

These results suggested that calculation-based MCBM probes can be used to predict performance on several criterion measures commonly used (standardized tests, GPA, teacher rating) within the educational system to make important decisions (e.g., educational placement). However, this study also contained several weaknesses that must be addressed. First, the study used a small sample of students in only two grade levels. Second, the criterion variables were limited to only one standardized test: the CAT. Third, the data obtained from the CAT were 1-2 years old for the participants at the time of the study, so the two measures likely captured different abilities at different times. The disparity of time between administration of the CAT and the MCBM was a significant limitation of the study's internal validity as the predictor measures were likely measuring at least somewhat different constructs than the CAT scores had done. Finally, because of the static nature of the study, limited evidence of the measures' sensitivity to growth was provided. Without adequate change in slope, the predictability of a dual discrepancy analysis examining the relationship between student level, slope or both was not feasible or included in this study. Thus, although these results are promising indicators of the potential utility of MCBM data as an indicator of performance on other important outcome measures, future research must address the limitations of the study to enhance the external validity thereof.

Helwig, Anderson, and Tindal (2002) extended the research by including multiple norm-referenced tests as the criterion measures. In this study, 207 fourth graders from general education classrooms in four elementary schools within one district were

administered a mathematics calculation MCBM probe. The majority of students involved in this study received mathematics instruction within the general education curriculum (specifically, 74%), while the remaining students received their mathematics instruction in special education. Participants were administered three MCBM probes consisting of a variety of mathematics problems from the annual curriculum; the problems ranged in difficulty from basic addition, subtraction, multiplication, and division facts to more difficult multiplication and division problems that would require the use of algorithms and strategies (e.g., 362 x 25). Each 5-minute probe consisted of approximately 36% basic skill problems and 64% advanced skill problems. Participants in the study were also administered the computation subtest of the Green Level test of the SDMT and the mathematics computation subtest of the CAT.

The results of this study indicated that the MCBM tests correlated strongly with both the SDT and the CAT. Specifically, the three MCBM probes had a mean correlation of $r = .57$ with the SDT and $r = .61$ with the CAT. Hence, given the large correlation between the MCBM probes and the standardized tests, it is reasonable to conclude that students performing poorly on the MCBM probes would also likely perform poorly on the SDT and CAT. To further support this hypothesis, the researchers conducted a series of regression analyses to determine how much of the variance in CAT and SDMT scores could be explained by student performance on the MCBM probes. Results of simple regression analyses indicated that total MCBM scores accounted for approximately two thirds of the variance in CAT scores for the study sample.

Although Helwig and colleagues (2002) expanded the research by including multiple criterion tests, the limitations of their study must also be noted in an effort to

continue to further the research process. The first limitation is the lack of sensitivity of the MCBM probes to student growth over time. Although pre- and postintervention scores were obtained, no data were provided regarding the slope of the students' progress over time. A second limitation of this study was its failure to include a skills analysis component, thus failing to provide specific information regarding students' unique strengths and weaknesses within their mathematical repertoire. The third limitation is the limited number of NRTs investigated. In addition, both tests used in this study are typically used only to evaluate general education curriculum, so there was no extension to include those tests that are commonly used in the process of educational placement decisions (i.e., general education vs. special education placement). Finally, participants in this study represented only the fourth grade. Thus, future studies should seek to address a wider range of grades and demonstrate the sensitivity to growth of the particular MCBM probe being implemented within the study.

Jitendra and colleagues (2005) furthered the research process by including two types of skill probes on two separate criterion measures. Seventy-seven third-grade students representing four third-grade classrooms in a suburban district participated in an investigation to determine the concurrent and predictive utility of applications-based MCBM through correlation and regression analysis. According to the authors, approximately 15% of the school population in which the study was conducted represented ethnic minorities, 17% were economically disadvantaged, and 5% spoke English as a second language.

Participants were administered a 3-minute computation-based MCBM (computation fluency, CF) probe, consisting of 25 grade-level skill problems, once each

week for 4 weeks during the two primary phases of the study (i.e., winter and spring) for a total of eight probes. The average of the four scores obtained during winter and the average of the four scores obtained during spring was utilized in the statistical analyses of the study.

Participants were also administered an applications-based MCBM (word-problem solving fluency, WPS-F). Participants were administered the WPS-F once every 2 weeks for the duration of the 16-week study. Students were given 10 minutes to complete each eight-problem probe. Each probe was comprised of addition and subtraction word problems that were selected from commonly used third-grade mathematics textbooks; however, in an effort to ensure that students had not previously encountered the problems utilized in the study, all textbooks from which problems were taken were not included within the school's curriculum.

All participants were administered the mathematics procedures subtest of the Stanford Achievement Test-9 (SAT-9) in the winter and administered the mathematics computation subtest of the TerraNova achievement test in the Spring to assess their initial mathematics achievement. Concurrent validity coefficients were calculated in the winter and spring to address the relationship between students' scores on the CBM measures (WPS-F and CF) with the norm-referenced measures administered in the winter and spring. Results indicated that concurrent validity coefficients for the winter WPS-F were within the moderate range, with WPS-F more strongly correlated (r = .71) than the CF (r = .49) to SAT-9 problem solving scores obtained in winter. The SAT- 9 procedures subtest was also moderately correlated with the CBM scores, but with CF (r = .64) a stronger correlate relative to WPS-F (r = .58). In addition, concurrent validity

coefficients for the TerraNova concepts and applications subtest with spring WPS-F and spring CF measures were also in the moderate range, with performance on the WPS-F (r = .58) more strongly related than the CF (r = .45). Concurrent validity coefficients for the TerraNova computation against spring WPS-F and CF measures followed a similar pattern with the CF slightly more strongly correlated (r = .51) than WPS-F (r = .48) to TerraNova computation scores. This pattern of reversing strength of correlations among measures is to be expected given the matching underlying math construct purported to be measured by each respective instrument. Together, the evidence generated indicates that MCBM adequately measures math achievement as defined by the SAT-9 and TerraNova tests.

Predictive validity coefficients were also calculated for the WPS-F and CF probes on the criterion measures. Results indicated that the same pattern observed within the concurrent validity coefficients also existed within the predictive validity coefficients: application-based MCBM was the best predictor of application-based criterion measures and calculation-based MCBM was the best predictor of calculation-based criterion measures. Further predictive validity comes from the results of a forced-entry regression analysis which indicated that scores obtained in the winter on both MCBM measures successfully predicted performance on the scores obtained in the spring on the criterion measures at a statistically significant level ($p < .05$). Hence, the concurrent and predictive validity of MCBM grounded in either concepts or calculations appears to be within the moderate range when using the mathematics subtests of the SAT-9 and TerraNova assessments as the criterion variables. Thus, it is reasonable to conclude that

data obtained from MCBM may be used to predict subsequent performance on norm-referenced tests, including those utilized as important outcomes measures.

Again, the results of this study must be interpreted within its inherent constraints. The first limitation of this study is that it contained a relatively small sample size representing little ethnic diversity at one grade level. Third, few participants received special education services during the study ($n = 5$). Fourth, the MCBM instruments used within the study were designed only to measure only addition and subtraction, so no information regarding other mathematics operations can be derived. Finally, no evidence of sensitivity to growth over time was provided. Thus, future research should attend to and attempt to resolve these concerns.

Recently, Shapiro, Keller, Lutz, Santoro, and Hintze (2006) increased the available research by including two types of criterion measures and multiple MCBM probes over time within a much larger sample than had previously been utilized. In this study, the researchers investigated the accuracy of predicting performance on a statewide curriculum-referenced test administered in the spring (PSSA) from MCBM performance administered in the fall, winter, and spring. Participants were drawn from a stratified random sample across six elementary schools in two districts in Pennsylvania to more accurately represent the socioeconomic status (SES) of the districts from which they were drawn. A total of 906 students from Grades 3, 4 and 5 participated in the study ($n =$ 337 third graders, $n = 271$ fourth graders, and $n = 298$ fifth graders). Students were administered computation probes in this study consisting of 25 multiple-skills problems designed to assess mastery of computation skills typical for each grade level in the fall, winter, and spring. No information regarding the length of time limits imposed during

CBM administration was reported, however. The results indicated weak to strong correlations between MCBM and the SAT–9 scores (range of $r = .058$ to $r = .727$) and weak to strong correlations between MCBM scores and PSSA scores (range of $r = .058$ to $r = .727$) for students in the second through fifth grades. Specifically, the authors reported that of the 18 correlations calculated between MCBM administrations in the fall, winter, and spring for both second and fourth grade students, all were statistically significantly correlated with subsequent performance on the problem solving, procedures, and total math subscales of the SAT-9 at $p < .001$, with only one exception (second grade problem solving administered in the fall, $p > .001$). In addition, correlations between MCBM scores administered in the fall, winter, and spring for third- and fifth-grade students with PSSA scores ranged from $r = .072$ to $r = .644$; all correlations were statistically significant at $p < .001$. These results provide evidence for the utility of MCBM as a moderate predictor of educationally relevant assessment outcomes.

In addition to establishing the predictive validity of MCBM on the PSSA and SAT-9, Shapiro and colleagues (2006) discussed the utility and practicality of establishing benchmarks (or cut scores) based upon MCBM scores in the fall, winter, and spring to successfully predict students who would meet the criterion level of the PSSA administered in the subsequent spring. A series of receiver operating characteristics (ROC) curves were developed that modeled the diagnostic accuracy of the MCBM and PSSA scores over a range of benchmarks. This process identifies benchmarks that maximize the sensitivity and specificity trade-off (i.e., increasing one at the expense of the other) in an effort to maximize benefit while minimizing cost. Using this process, the authors reported that all measures utilized in their study showed positive predictive

power (i.e., correct prediction of students' failure on the PSSA based upon his/her MCBM score) ranging from 80-93% and negative predictive power (i.e., correct prediction of students' success on the PSSA based upon his/her MCBM score) ranging from 48-68%. Overall correct classification rate of PSSA success based upon MCBM score was approximately 85% with overall specificity and sensitivity ranging from .6 to .7.

Although this study provided a promising expansion of previous research on the utility of MCBM data to predict performance on important outcome measures, the results must be interpreted within the confines of its limitations. First, attrition was a potential threat to the internal validity of the study because only those students who had full data sets were included in the analysis at the end of the study. Second, problems with data collection resultant from incorrect MCBM administration procedures were noted in at least one data set (i.e., all data collected at one time point) in the study. Although the contaminated data were excluded from the analyses reported in the study, concerns relative to the validity of the included data remained because treatment integrity was not monitored 100% of the study; hence, some probability of data corruption through unspecified study procedures remained. Thus, future research should seek to implement a stringent treatment fidelity procedure.

Conclusions and Recommendations

Results of the present review provide empirically based evidence that MCBM is a useful screening instrument that can be used to screen for the existence of academic problems among elementary and possibly secondary students (Foegen & Deno, 2001;

Helwig et al., 2002; Jitendra et al., 2005; Shapiro et al., 2006). Further, the results of the review also indicate that MCBM can be used effectively to help improve student achievement through skills analysis and instruction modification (Foegen & Deno; Helwig et al.; Jitendra et al.; Shapiro et al.; Stecker et al., 2005). Finally, results indicated that MCBM shows at least moderate correlations with a variety of NRTs (Foegen & Deno; Helwig et al.; Jitendra et al.; Shapiro et al.). Thus, the available evidence indicates that MCBM may be a promising approach to identify students as high risk for failing to meet relevant standards on important outcome measures.

However, although the studies reviewed provide evidence of the validity of MCBM, they are not without their limitations. Perhaps the greatest limitation is that few studies combined multiple NRTs and MCBM probes in the same study. Thus, no direct comparison of any one MCBM probe has been made with multiple NRTs, including those commonly used to identify students with learning disabilities. This is particularly important because one major problem associated with intervening with students who would likely benefit from additional education services (e.g., academic intervention or specialized instruction) is the limited evidence of reliability of identification procedures across systems (Augustyniak, Murphy, & Phillips, 2005; Mazzocco, 2005; Mazzocco & Myers, 2003). However, if students' performance on MCBM can be correlated across multiple tests commonly used to identify students requiring additional educational services, the accuracy with which students would be identified as high risk across school systems would be increased.

Given the results of the present review, additional research designed to augment and extend that which is currently available is also clearly necessary. Several

suggestions can be generated from the results of the current review of literature. First, the correlation of various types of MCBM probes with various types of NRTs within a larger, more diverse sample representing multiple grades is important. Second, future studies should attempt to investigate the sensitivity of MCBM probes utilized within the study to growth over time in an effort to ensure more accurate identification of students likely to fail to meet standards. Studies using reading CBM assessments suggest that this can be accurately accomplished by utilizing a dual discrepancy analysis within the MCBM system (Burns & Senesac, 2005; Speece & Case, 2001). Third, probes should be designed to provide information that can be utilized to define instructional treatment plans for those individuals identified as high risk for failure (Speece & Case). Fourth, probes should be created so that they are able to predict performance on NRTs and be used for error analysis.

Purpose of Study

Using MCBM for the purposes of screening and progress monitoring of math skill development over time is useful because they can be easily administered at any time to make educational decisions. Data obtained from MCBM can be used in schools to answer such questions as: (a) what fluency score determines that a child has developed math skills to a degree that skills will proficiently be demonstrated across various tests? (b) is the current instruction effective in increasing math skills for most children? and (c) which children are at risk for math difficulty because of inadequate math computation skills and thus need additional instruction or special education?

Although previous research has examined the concurrent and predictive validity of MCBM on group-administered tests, questions still exist regarding the degree to which MCBM scores predict math performance on individually administered tests used to identify students at high risk for math academic failure. A complex issue related to effective assessment is the question about which students are experiencing academic difficulty and require early and intensive intervention such as might be provided through the school's supplemental intervention or special education services (Klingner et al., 2006). If MCBM is to be used for identification and placement of children in services, the measure should accurately differentiate between children who have not yet acquired skills and those who have acquired skills. Further, research on the decision validity of a dual discrepancy approach to the identification of risk level with RCBM provides guidance for future research in other areas of CBM including mathematics. Future research on the technical adequacy and utility of MCBM as a predictor of mathematics achievement on educationally relevant testing may lead to its increased use within education systems to efficiently and effectively identify and support the right child at the right time.

This study replicated and extended the current research literature by investigating the extent to which MCBM could be used to identify children likely to fail to meet proficiency standards on individually administered outcome measures without support. In this study, students were administered four types of MCBM probes over a 9-week period of the regular academic school year; the results have been correlated with scores obtained on several NRTs administered at the end of the study (math fluency subtests of the WJ-ACH-III, Key Math 3, and WIAT-II) to determine the relationships between

different MCBM probe scores at different points in time and several tests commonly used to make important educational decisions. These standardized tests were selected for inclusion in the study because of their strong psychometric properties and their frequent use by school psychologists to determine eligibility for special education services. The four math probes consisted of a single skill that had previously been taught and students were expected to have mastered, a single skill that was to be acquired during the course of the study, multiple skills that were to be acquired over the course of the study, and an error analysis of multiple skill steps ranging from skills that had been mastered to skills that students were to acquire. All skills on each probe were math calculation skills only (e.g., addition, subtraction, multiplication) and did not represent math reasoning skills (e.g., rounding, estimating, story problems). The data was used to explore the efficacy of the MCBM probes to identify students most likely to succeed or fail on three assessment measures commonly used in decision-making processes for students struggling to succeed in school. First, the MCBM scores were analyzed according to level, slope, and the aggregate of level and slope to determine the differential impact of each probe type on their individual correlations between students' MCBM scores and their respective scores on the outcome measures. Second, predictive classification accuracy was employed to explore the degree that various types of MCBM survey-level assessments with students in Grades 2, 4, and 6 identified high- and low-risk students in mathematics based upon the subsequent results of three NRTs with the hypothesis that the error analysis would be the most accurate. Further, predictive classification accuracy was employed to explore the degree that various types of performance indicators (level, slope, or dual discrepancy) identified high- and low-risk students in mathematics based upon

the subsequent results of several subtests from three NRTs with the hypothesis that the dual discrepancy indicator would be the most accurate.

Thus, the current study sought to answer the following research questions:

1. What is the relationship between screening performance indicators (i.e., level, slope, or dual discrepancy) on four types of MCBM probes (a probe representing a taught and mastered skill, a probe representing a single skill to be mastered, a probe representing multiple skills to be mastered, and a probe representing multiple step-by-step skills) and math calculation subtests from three individually administered standardized achievement tests?

2. To what extent does each of the screening performance indicators on each of the four types of MCBM probes accurately predict high- and low-risk student performance status on scores derived from three individually administered standardized achievement tests?

3. What combination of MCBM screening measures best predicts high- and low-risk student performance status on scores derived from three individually administered standardized achievement tests?

CHAPTER III

METHODS

Setting and Participants

The study was conducted in three suburban elementary schools (consisting of Grades kindergarten through 6) located in a Western state. Approximately 1,600 students from kindergarten through sixth grade attended the schools that consisted of 85% Caucasian students, 10% Latino students, 3% African American students, and 2% students from other ethnic and racial backgrounds. Approximately 35% of all students at all of the schools qualified for federal free or reduced lunch programs. Schools 1 and 2 contained three classes of students in Grades 2, 4, and 6. School 3 contained two classes of students in Grades 2, 4, and 6. All schools used Houghton Mifflin Mathematics curriculum; the school district required all teachers to teach from the same curriculum. None of the schools included in the study utilized a tiered instruction program in mathematics instruction.

The MCBM assessments were conducted by trained research assistants in the students' regular classrooms. A total of 6 undergraduate research assistants participated in the training and administration of the MCBM and NRT measures. All research assistants who participated in the administration of measures were required to demonstrate mastery of administration procedures before being allowed to administer measures to research participants. In addition, all research assistants were observed by the primary researchers and/or licensed school psychologists during the first 3-5 administrations of the NRTs to all research participants. No research assistants were

allowed to continue administering NRTs until they had demonstrated mastery of standardized administration procedures both during training sessions and during administration to research participants. Thus, all of the subtests from the NRTs were individually administered by trained research staff in empty school rooms.

Participants were involved in two phases of the study. Participants in phase I included all students within the second, fourth, and sixth grades of the participating schools who were assessed as part of a schoolwide assessment process the school utilized to evaluate student math performance. In total, 685 participants were involved in the first phase of the study (n = 291 from School 1, n = 248 from School 2, n = 146 from School 3). Table 1 contains the demographic information for all phase I participants involved in the study; these data were collected from records obtained through the school district office.

The MCBM data collected during phase I were utilized to calculate risk status cut-points for all phase II participants; performance indicator cut-points were calculated for each classroom from which phase II participants were involved in the study. Specifically, each individual student's risk status was determined using the MCBM data to calculate the specific cut-points per class. For example, growth rates or slope scores were calculated using least squares regression between monitoring scores and calendar days for all phase I participants (Deno, Fuchs, Marston, & Shin, 2001; Good & Shinn, 1990; Hintze & Christ, 2004; McMaster, Fuchs, Fuchs, & Compton, 2005). Phase II participants were identified as low or high risk using local normative criteria by rank ordering the growth rates for within the three grades and using the 33rd percentile rank to judge risk status within each class (Burns & Senesac, 2005). That is, student scores

Table 1

Demographic Information of Phase I Participants

	School 1 (%)	School 2 (%)	School 3 (%)	Total (%)
Low SES	8.8	9.0	26.1	34.7
Ethnic minority	3.1	4.6	13.2	15.0
Special Education	6.2	6.5	10.3	8.4
Male	50.2	48.4	50.0	49.5

below the 33rd percentile rank in each individual's class were classified as high risk and student above the criterion were classified as low risk. Table 2 presents the total number of high- and low-risk students for each school and grade.

A consent form and letter explaining the purpose of the study (including a description of the requirements of student participation) was sent to all parents/legal guardians of phase I students (see Appendix A). Letters and consent forms were written in English and Spanish; parents whose native language is other than English or Spanish were offered an interpreter through the school's existing services.

All students involved in phase I of the study were eligible for participation in phase II, except those students who were actively involved in a determination process for special education services or those who were scheduled to be involved in such within the next 6 months (e.g., students requiring a special education re-evaluation to determine continued eligibility). All potential phase II participants were sent a letter from their

Table 2

Risk Status Percentages of Phase I Participants

	Grade 2		Grade 4		Grade 6		Total	
Source	High	Low	High	Low	High	Low	High	Low
Error Analysis								
Time 1	29	174	31	151	28	152	91	477
Time 2	35	174	32	172	44	158	113	505
Time 3	32	178	32	161	29	144	94	485
Time 4	35	169	35	173	31	171	102	513
Slope Time 3	74	140	69	129	66	132	210	403
Slope Time 4	62	160	104	109	59	156	226	427
DD Time 3	22	179	23	162	12	157	58	500
DD Time 4	16	187	26	177	14	186	57	550
Multiple Skills								
Time 1	30	173	30	152	27	152	87	480
Time 2	34	175	38	166	29	169	103	511
Time 3	32	178	30	162	25	148	89	489
Time 4	29	175	32	176	31	171	93	522
Slope Time 3	76	141	70	134	66	129	214	405
Slope Time 4	98	126	48	164	79	135	226	427
DD Time 3	21	183	17	170	13	152	53	506
DD Time 4	21	182	16	186	21	178	58	547
Single Skill L								
Time 1	32	171	29	153	92	87	154	413
Time 2	33	176	30	174	97	105	162	456
Time 3	31	179	32	160	42	131	107	471
Time 4	37	167	32	176	31	171	101	514
Slope Time 3	73	142	65	140	68	125	207	409
Slope Time 4	72	150	68	148	70	144	211	444
DD Time 3	17	185	16	173	23	143	57	503
DD Time 4	22	180	17	189	30	171	70	540
Single Skill M								
Time 1	433	170	29	155	27	150	91	476

respective principal encouraging them to participate in the study and notifying them that they would be offered a small tangible reinforcer (e.g., a new pencil or piece of candy) for returning the consent form regardless of study participation status (see Appendix B). All phase II participants with informed consent and a complete data set were included in the study. Only participants for whom consent was obtained for participation in phase II were administered the NRTs. Table 3 contains the number of students per grade for whom consent to participate was obtained and denied by school and grade. In addition, a total of 35 forms denying consent and 19 forms providing consent were returned without sufficient identifying data to determine which students returned the forms.

Table 4 contains the following demographics for all phase II participants as reported by their parents/guardians (Appendix C): percentage of students in the low SES range, percentage of students representing ethnic minorities, and percentage of students receiving special education services. All percentages are calculated according to percentage within the entire sample.

Measures

Predictor Variables: MCBM

Four different probes comprised of grade level basic computational mathematics skills were constructed for this study comprising four different sets of problems. The mastered single skill (SSM) probe consisted of problems that students had previously been taught and mastered (as defined by 80% accuracy on 80% of trials as reported by their teachers). The single skill to be learned (SSL) probe consisted of a single math skill (e.g., addition) that students were being taught over the nine weeks that the study

Table 3

Results of Informed Consent Returned

	School 1		School 2		School 3	
	Obtained (*n*)	Denied (*n*)	Obtained (*n*)	Denied (*n*)	Obtained (*n*)	Denied (*n*)
Grade 2	49	3	30	2	16	0
Grade 4	30	9	27	4	17	1
Grade 6	37	11	15	0	4	0

Table 4

Demographic Information of Phase II Participants

	School 1 (%)	School 2 (%)	School 3 (%)	School 4 (%)
Low SES	6.9	1.4	8.3	16.7
Ethnic minority	2.1	2.1	4.2	8.3
Special Education	2.8	3.5	0.7	6.9
Male	20.8	15.3	3.5	39.6

progressed. The multiple skills probe (MS) consisted of several different math skills (e.g., addition and subtraction) that students were reportedly being taught over the 9 weeks of the study. The final probe , error analysis (EA), consisted of multiple step-by-step skills using simple computation facts. Examples of each probe can be seen in Appendices D, E, F, and G.

The specific problems that were included on each of the probe types were created from the information obtained through surveys completed by the majority of teachers whose students participated in the study. After the teachers completed the survey reporting problems students were scheduled to learn over the course of the study, random

numbers were utilized to create each probe so that all problems present on each probe were consistent with teacher feedback. Once probes were created, a sample of approximately 50% of teachers whose students participated in the study were asked if the problems represented the curriculum they were scheduled to teach over the course of the study. The variants of the probes were created by randomly moving problems within each probe such that all probes contained the same problems, but that all problems were presented in a random order.

The number of problems included on each probe was selected such that correctly written answers equaled a minimum number of digits correct to ensure that all students who had mastered the skill would not finish during a 2-minute probe administration. This was determined by consulting with teachers regarding the estimated number of problems students could complete within the allotted time. In addition, a sample of approximately 20 students from a school that was not involved in the research completed a variant of the CBM probes to determine if they were able to complete all problems within 2 minutes.

The type of skill problems presented on each of the four MCBM probes varied. For the EA, single skill, and multiple skill probe types, each problem type was selected to represent specific skills that teachers reported students were actively learning over the course of the study. For the single skill mastery probe, problem types were selected to represent math skills teachers reported that 80% of their students could perform with 90% accuracy on 80% of administrations. Teachers were asked to select specific skills for each of the probes on a survey that all participating teachers received 3 weeks prior to the beginning of the study. The survey consisted of problem types that covered the four

basic computational skills, decimals, and fractions relevant to each grade level as determined by state and district curriculum requirements, content of classroom math text, and teacher input (Shapiro, 1996); these problem types were hierarchically arranged (see Appendix H). Teachers were asked to select specific skills that met the stated skill set criteria from the list of grade level computation math skills (see Appendix H). A skill or combination of skills most commonly identified by all teachers within each grade level across all schools was included in the probe administered to that grade. The four types of MCBM probes and the definition of the skill presented on the probe follows.

Mastered Single Skill Probe (SSM). Problems presented on the mastered single skill probes (SSM) were selected by asking teachers in each grade level to indicate one specific skill their students were expected to have mastered by the time of probe administration. Teachers were told that a mastered skill would be a skill that meets three criteria: (a) the skill has already been taught to students within the last two months, (b) students have been given several opportunities to independently practice the skill, and (c) 80% or more of the students are able to complete this skill above 90% accuracy.

Single Skill to be Learned Probe (SSL). A single math skill (e.g., two-digit addition plus regrouping for second grade students) that had not been taught to students in the current school year was included on the single skill probe (SSL). However, teachers were asked to select skills that they planned to teach and practice during the 9-week curriculum during which the study took place.

Multiple Skills Probe (MS). Teachers were asked to select all skills on the list of grade level computation skills that students should have mastered by the end of the 9-

week curriculum during which the study took place. A combination of three to five skills were randomly presented using randomly generated numbers to generate the problems.

Error analysis of skill steps probe (EA). This type of MCBM probe covered multiple computation skills structured in hierarchical order of skill difficulty. Similar to prior studies, these probes were created to enable the identification of possible areas of intervention for math computation skills using skills analysis (Fuchs et al., 1990). This probe was also designed to minimize mastery of basic facts to focus on mastery of computation steps. To accomplish this, each problem represented a skill step that can be taught within a brief (i.e., 5-minute) lesson using a coach card. However, all problems were comprised of facts that would require minimum finger counting or mental counting; for example, 21 + 13 rather than 45 + 36 were used for a double-digit addition problem with no regrouping.

Criterion Variables: Standardized Norm-Referenced Tests

Three NRTs were selected for use in this study based on robust psychometric properties and were reported to be commonly used to make eligibility determination for special education services (Magyar, Pandolfi, & Peterson, 2007). The three selected NRTs are described below.

Woodcock-Johnson Tests of Achievement–Third Edition (WJ-ACH-III). The WJ-ACH-III is an individually administered, standardized, and norm-referenced achievement test. The battery consists of 22 subtests that cluster into 16 composites. Each subtest yields a standard score and a composite score can be calculated for 16 areas. For the purpose of this study, two subtests were selected as a criterion for comparison against the

experimental probe measures: calculation and math fluency. The calculation subtest requires students to complete math problems arranged hierarchically according to difficulty under nontimed conditions. Students continue attempting to solve problems until they answer six consecutive problems incorrectly. Problems range in difficulty from simple addition to calculus problems. The math fluency subtest requires students to complete as many mathematics problems as they can within 3 minutes. All problems are hierarchically arranged according to difficulty and difficulty levels range from simple addition to simple multiplication (i.e., one-digit by one-digit products); there are no division problems on the math fluency subtest. These two subtests were selected for two reasons: (a) these subtests sampled skills similar to those sampled by the experimental probes, and (b) the subtests selected have strong psychometric properties. Reliability coefficients within the normative sample representative of the age ranges that were used in the study for the math calculation skills cluster score (comprised of the calculation and math fluency subtest scores) ranged from .80-.87; reliability coefficients within the normative sample for the math fluency subtest ranged from .77-.89 across the age ranges for participants within the study (McGrew & Woodcock, 2001). One-year test-retest reliabilities for Calculation and Math Fluency ranged from .81-.83 and from .86-.89, respectively, across the age ranges for participants within the study. Correlations between the calculation subtest of the WJ-ACH-III and the mathematics composite and mathematics computation scores of the Kaufman Test of Educational Achievement (KTEA) were reported as .60 and .67, respectively (McGrew & Woodcock). Correlations between the calculation subtest of the WJ-ACH-III and the mathematics

composite of the Wechsler Individual Achievement Test (WIAT-II) was reported as .69 (McGrew & Woodcock).

Key Math 3. Each student was administered the mathematics calculation subtest of the Key Math 3 (KM 3). The KM 3 is an individually administered, standardized, and norm-referenced achievement test. The battery consists of 13 subtests that cluster into four composite scores: total test, basic concepts, operations, and applications. Each subtest yields a scaled score and an area score can be calculated for each of the four areas named above. For the purpose of this study, two subtests (addition/subtraction and multiplication/division) were selected to be used as a criterion for comparison against the experimental probe measures. All subtests consist of domain-specific items hierarchically arranged according to difficulty and are individually administered under nontimed conditions. In addition, students are required to attempt to answer items until they fail to correctly complete three consecutive items. Problems on the addition/subtraction subtest range in difficulty from counting visual images (i.e., "finger counting") to adding and subtracting fractions with unlike denominators. The multiplication/division subtest ranges in difficulty from the exploration of basic facts (i.e., single-digit products and dividends) to multiplication and division of fractions and mixed numbers. These subtests were selected because they sample skills similar to those sampled by the experimental probes and the subtests selected have strong psychometric properties. Reported split-half reliability coefficients for the included subtests and composite scores across the included grades ranged between $r = .54$ to $r = .92$ for second graders, $r = .56$ to $r = .94$ for fourth graders, and $r = .65$ to $r = .95$ for sixth graders (Connolly, 1997). Correlations between the selected KM 3 subtests and composite score

and the mathematics computation subtest of the Comprehensive Tests of Basic Skills (CTBS) ranged between $r = .56$ to $r = .65$, and ranged between $r = .51$ to $r = .77$ when comparing the selected KM 3 subtests and composite score and the Iowa Test of Basic Skills (ITBS) mathematics concepts and mathematics computation subtest scores.

Wechsler Individual Achievement Test–2nd Edition (WIAT-II). The WIAT-II is an individually administered, standardized, and norm-referenced achievement test. The battery consists of nine subtests that cluster into four composite scores: reading, mathematics, written language, and oral language. Each subtest yields a scaled score and a composite score can be calculated for each of the four areas named above. For the purpose of this study, only the numerical operations subtest of the mathematics composite was selected to be used as a criterion for comparison against the experimental probe measures. This subtest consists of domain-specific items hierarchically arranged according to difficulty and is individually administered under nontimed conditions. In addition, students are required to attempt to answer items until they fail to respond correctly to six consecutive items. This subtest was selected because it samples skills similar to those sampled by the experimental probes and has moderate to strong psychometric properties. Reported split-half reliability coefficients for the included subtests and composite scores across the included grades were $r = .83$ for second grade, $r = .85$ for fourth grade, and $r = .92$ for sixth grade (Wechsler, 2002).

The examiner's manual (2002) reports correlations of $r = .77$ between the numerical operations subtest of the WIAT-II and the arithmetic score of the Wide Range Achievement Test--Three (WRAT3) and correlations of $r = .75$ between the numerical

operations subtest of the WIAT-II and the Differential Ability Scales (DAS) basic number skill score.

Procedures

MCBM Probe Training, Administration, and Scoring Procedures

All MCBM probes were administered and scored by trained research assistants in the presence of the classroom teacher. Prior to the study, training on MCBM administration and scoring procedures was provided to everyone involved in the study. Training included detailed explanations and written instructions of specific MCBM administration and scoring procedures; modeling and practicing of the MCBM administration procedures with feedback was also provided. Research assistants were considered trained on correct administration procedures when they completed all steps accurately. Research assistants were considered trained on correct scoring procedures when they obtained at least 95% interscorer agreement with a primary researcher on two math probes. Inter-scorer agreement was calculated as a percentage by dividing the number of agreements for attempted items (i.e., both scorers agreed that the student correctly or incorrectly completed each digit) by the total number of agreements plus disagreements multiplied by 100. The mean interscorer agreement across all research assistants was 100%.

The four probe types described previously (i.e., mastered skill, single skill, multiple skill, and error analysis of skill steps) were group administered in the participants' respective classrooms during a nine week period in the spring of the

academic school year. Each student was administered all MCBM probe types in each session on four separate occasions (at the beginning of the study then after 10, 20, and 30 instructional days) in his/her classroom by a trained research assistant. All four probes were administered and collected within approximately one 10-minute session on each testing occasion (see Table 5 for test administration schedule).

Teachers were notified approximately 2 weeks in advance each time students were administered the MCBM probes. All students within each class in the study were administered the math calculation probes at their desks in their classroom during nonlunchtime hours. During an MCBM administration session, participants were provided with a packet of grade-appropriate MCBM probes and a pencil. The order of presentation of the four math calculation probes was counterbalanced across probe type by testing session. Before each testing session, students were told that they would be asked to complete some math problems and that they should try to do their best work (see Appendix I). All students were given the math probes with the problems facing down by the research assistant. The research assistant then instructed the students to turn the packets over, to begin working on the problems starting at the top of the page and to move across the page in a left-to-right, top-to-bottom manner, and began timing the students. Students were provided two minutes to complete each probe with approximately a 30-second break between probes. Upon completion of the last probe, students were asked to return their packets to the research assistant.

After all MCBM administrations, trained research staff scores the probes. The number of digits correct per 2 minutes (DCP2M) on each probe served as the datum

Table 5

Timeline for Administration of Measures

	Instructional days	Tests administered	Measurements
Time 1	0	Error Analysis MCBM Mastered Single Skill MCBM Multiple Skills MCBM Single Skill MCBM	Level
Time 2	10	Error Analysis MCBM Multiple Skills MCBM Single Skill MCBM	Level
Time 3	20	Error Analysis MCBM Multiple Skills MCBM Single Skill MCBM	Level & Slope
Time 4	30	Error Analysis MCBM Multiple Skills MCBM Single Skill MCBM	Level & Slope
Time 5	34	WJ-ACH-III WIAT-II KM 3	Level

reported for each student on each probe (Thurber et al., 2002). That is, any number written as part of the correct answer and in the correct place value was counted as a correct digit. If a student's answer was correct, the student received the full problem value for potential digits correct whether or not his/her work was shown. For problems that are incorrect or incomplete, credit was given for those digits correct in the solution. Finally, "carries" or "borrows" involved in problems requiring regrouping were not counted as digits correct.

Approximately 2 weeks after each administration, teachers were provided a graph that displayed the students' math scores on the MCBM probe from lowest to highest scores with lines marking an instructional standard norm benchmark and the class median score (see Appendix J). The instructional standard applied for math were 10-20 DCP2M for Grades 1-3, and 20-40 DCP2M for Grades 4-6 (as described by Deno, Mirkin, & Chiang, 1982).

Throughout the study, two trained scorers independently scored responses on 30% of each of the four administered probes and NRTs. Inter-scorer agreement was calculated as a percentage by dividing the number of agreements (i.e., both scorers agree that the student correctly or incorrectly responded) by the total number of agreements plus disagreements multiplied by 100. The mean interscorer agreement was 100% on all four probes.

Standardized Test Training, Administration, and Scoring Procedures

All NRTs were administered and scored by trained research assistants. Prior to the beginning of the study, training on NRT administration and scoring procedures for all research assistants was provided. In addition, evaluation of the training provided to the research assistants was conducted using the same procedures described previously regarding the training on MCBM procedures. The mean interscorer agreement between all research assistants and the primary researcher was 100% on all NRTs.

Phase II participants were administered the NRTs within one week of the last probe administration. Teachers were given a list of students selected for participation along with the specified time of test administration approximately 2 weeks before NRT

administration. All students within each class in the study were individually administered the selected sections of the WJ-ACH-III, the KM 3, and the WIAT-II in an empty room during non-lunchtime hours and were administered by trained research assistants.

All three tests were administered within approximately one 25-minute session. The order of the presentation of the three NRTs was counterbalanced across all students. Before the testing session, students were told that they would be asked to complete some math problems and that they should try to do their best work. The researcher then administered the standardized tests according to instructions provided in their respective testing manuals. Following each administration, the NRT was scored to obtain raw and standard scores for each subtest following the scoring directions provided in the test manuals.

CHAPTER IV

RESULTS

Overview

For this study, MCBM performance data were measured and analyzed using level score obtained at each time; slope score and dual discrepancy score (i.e., below level and below slope) were calculated for administrations from Time 1 to Time 3 and Time 1 to Time 4. Each of these three performance indicators was calculated based on prior research. Level scores were calculated by scoring the total number of digits correctly recorded by the student in 2 minutes on each math probe (Shinn, 1989). Next, the high risk students were identified as those students whose scores fell below the 16th percentile within their class on the MCBM probes and who were below the preset instructional grade level. The instructional standard applied for math was 10-20 DCP2M for Grades 1-3, and 20-40 DCP2M for Grades 4-6 (as described by Deno et al., 1982).

Growth rates or slope scores were calculated using least squares regression between monitoring scores and calendar days (Deno et al., 2001; Good & Shinn, 1990; Hintze & Christ, 2004; McMaster et al., 2005). Next, the high-risk and low-risk groups were identified using local normative criteria by rank ordering the growth rates for within the three grades and using the 33rd percentile rank to judge risk status (Burns & Senesac, 2005). That is, student scores below the 33rd percentile rank in each individual's class were classified as high risk and student above the criterion were classified as low risk.

For the dual discrepancy criterion suggested by Fuchs, Fuchs, Hosp, and Hamlett (2003), a student was identified as high risk if his or her growth rate fell in the high risk

criterion for slope and the CBM score fell within the high-risk level criterion for level. Alternatively, a student was identified as low risk if scores fell within the low-risk criteria for slope or level or both.

Classification of high- and low-risk status for each student was also calculated for each NRT test. High-risk students were identified as any students whose standard score fell below one standard deviation of the mean. In efforts to improve the psychometric stability of the criterion scores, cluster scores were used for the WJ-ACH-III. The WIAT and KM 3 subtests administered were insufficient to calculate a cluster score.

Descriptive Statistics

Several analyses of student MCBM and standardized subtest scores were conducted. First, descriptive statistics (e.g., means, ranges, and standard deviations) were computed (see Tables 6, 7, and 8). Results indicated that mean scores (DCP2M) on all probe types were highest for sixth-grade students and lowest for second-grade students across all probe administration times.

Second, preliminary analyses between participants at each school were conducted using chi-square tests to determine statistically significant differences between the three schools on gender, SES, ethnicity, and special education status (see Appendix K). Significant differences between schools were noted; however, data were not separated by school during subsequent analyses for several theoretical and empirical reasons. First, in order to provide a more useful metric to school districts, data were collapsed in order to illustrate the potency of using district level data to create local norms and thereby provide an empirical framework for the establishment of local norms for data-drive decision

Table 6

Descriptive Statistics of MCBM Level Scores of Phase II Participants

Population	*n*	Mean	*SD*	Median	Range	Skewedness
Time 1						
Error analysis						
Grade 2	49	15.1	6.7	16.0	31.0	.02
Grade 4	48	29.3	8.2	30.5	37.0	-.83
Grade 6	47	38.5	12.2	36.0	48.0	.13
Multiple skill						
Grade 2	49	19.8	9.8	19.0	41.0	.56
Grade 4	48	21.4	10.3	22.0	55.0	.80
Grade 6	47	25.8	10.2	25.0	47.0	.50
Single skill						
Grade 2	49	18.4	11.6	17.0	31.0	.50
Grade 4	48	30.0	14.7	28.0	61.0	.47
Grade 6	47	45.5	29.4	48.0	97.0	.13
Acquired skill						
Grade 2	49	19.7	11.1	17.0	51.0	.37
Grade 4	48	37.8	15.2	36.0	61.0	.68
Grade 6	47	40.8	16.4	38.0	72.0	.22
Time 2						
Error analysis						
Grade 2	49	15.6	6.4	17.0	28.0	-.59
Grade 4	48	29.7	9.0	29.5	42.0	-.28
Grade 6	47	35.6	14.4	34.0	58.0	-.27
Multiple skill						
Grade 2	49	20.0	10.7	19.0	43.0	.59
Grade 4	48	23.6	10.2	20.5	51.0	1.01
Grade 6	47	29.6	11.3	29.0	53.0	.04
Single skill						
Grade 2	49	19.0	11.3	20.0	49.0	.30
Grade 4	48	37.3	16.1	36.5	68.0	.65
Grade 6	47	48.2	28.1	46.0	109	.35

(*table continues*)

Population	*n*	Mean	*SD*	Median	Range	Skewedness
Time 3						
Error analysis						
Grade 2	49	17.7	5.5	18.0	25.0	-.76
Grade 4	48	30.1	9.1	30.0	37.0	.13
Grade 6	47	42.3	12.0	43.0	54.0	.16
Multiple skill						
Grade 2	49	22.6	10.9	21.0	48.0	.46
Grade 4	48	24.3	9.6	23.0	43.0	.48
Grade 6	47	28.1	12.8	29.0	59.0	.42
Single skill						
Grade 2	49	23.7	10.4	22.0	57.0	.60
Grade 4	48	36.9	16.3	32.0	75.0	.61
Grade 6	47	51.2	23.8	55.0	108.0	.28
Time 4						
Error analysis						
Grade 2	49	18.3	5.9	20.0	27.0	-.88
Grade 4	48	30.8	9.1	31.0	37.0	-.24
Grade 6	47	41.4	12.1	38.0	49.0	.42
Multiple skill						
Grade 2	49	23.6	13.8	23.0	59.0	.50
Grade 4	48	27.1	12.5	26.5	56.0	.18
Grade 6	47	33.0	13.7	35.0	63.0	.59
Single skill						
Grade 2	49	23.5	13.4	21.0	74.0	1.15
Grade 4	48	39.4	18.7	41.0	71.0	.36
Grade 6	47	46.0	30.2	38.0	111.0	.38

making (VanDerHeyden, Witt, & Gilbertson, 2007). Second, data were collapsed across schools for analyses because it provided a more diverse representation of students with regards to ethnicity, SES, and special education status. Finally, data were collapsed because a low number of high-risk students were identified within the sample of phase II participants and separating students according to school would have decreased the power

Table 7

Descriptive Statistics of MCBM Slope Scores of Phase II Participants

Population	*n*	Mean	*SD*	Median	Range	Skewedness
Weeks 1 to 3						
Error analysis						
Grade 2	49	.65	1.38	.50	6.75	.24
Grade 4	48	.17	1.59	.13	9.75	.32
Grade 6	47	.61	2.78	.75	19.50	-3.66
Multiple skill						
Grade 2	49	.83	2.26	.75	13.75	1.05
Grade 4	48	.69	1.66	.63	8.00	.171
Grade 6	47	.48	2.21	.50	10.75	2.21
Single skill						
Grade 2	49	1.29	1.93	1.25	8.50	.33
Grade 4	48	2.49	2.49	1.63	10.00	.28
Grade 6	47	1.72	1.72	.75	7.75	1.08
Weeks 1 to 4						
Error analysis						
Grade 2	49	.49	.90	.30	4.56	.22
Grade 4	48	.25	.73	.28	4.94	-.11
Grade 6	47	.59	1.10	.45	4.98	.86
Multiple skill						
Grade 2	49	.49	.90	.79	5.86	-.46
Grade 4	48	.80	1.16	.72	5.48	.01
Grade 6	47	.87	1.62	1.25	6.67	-.40
Single skill						
Grade 2	49	.83	1.55	.61	8.15	.71
Grade 4	48	1.12	1.62	.88	6.73	.36
Grade 6	47	3.68	3.20	4.03	11.60	.06

of the statistical analyses included in the study.In order to investigate significant differences between grades and math probes within probe type, a series of mixed between within ANOVA were conducted. A three (MCBM probe types) by three (grades)

Table 8

Descriptive Statistics of NRT Scores of Phase II Participants

Population	*n*	Mean	*SD*	Median	Range	Skewedness
WIAT						
Grade 2	49	101.4	13.5	100.0	64	0.547
Grade 4	48	108.8	10.9	109.0	50	0.110
Grade 6	47	103.3	12.9	102.0	60	-0.064
Total	144	104.5	12.8	104.0	70	0.096
WJ-ACH-III						
Grade 2	49	106.6	12.2	108.0	54	-0.636
Grade 4	48	108.8	9.2	109.0	35	0.110
Grade 6	47	105.8	13.2	105.0	64	0.177
Total	144	107	11.6	107.5	66	-0.234
KM Add/Sub						
Grade 2	49	10.0	2.8	10.0	14	-0.226
Grade 4	48	10.6	2.8	11.0	13	0.343
Grade 6	47	9.8	2.5	10.0	9	-0.284
Total	144	10.1	2.7	10.0	15	-0.003
KM Mult/Div						
Grade 4	48	10.6	2.7	11.0	14	-1.065
Grade 6	47	9.8	2.6	10.0	15	-0.639
Total	95	10.2	2.7	10.0	15	-0.805

ANOVA was performed with the three probe types administered across the four administration times serving as the within-subjects variable and the three grades serving as the between-subjects grouping variable. Results of the assumptions of normality of distribution, equality of population variances, and independence of scores were investigated.

As shown in Table 9, the RM ANOVA indicated significant main effects for time and grade and a significant time by grade interaction. As shown in Table 10, independent

Table 9

Results of RM ANOVA for Grade by Probe Type Across All Administration Times

Source	*df*	Mean square	*F*	*p*	Partial η^2
EA	3	0.76	14.75	0.000	0.241
grade	2	25175.67	88.78	0.000	0.557
EA * grade	6	0.88	3.12	0.006	0.063
MS	3	0.74	16.51	0.000	0.263
grade	2	2856.51	6.79	0.002	0.088
MS * grade	6	0.91	2.28	0.037	0.047
SSL	3	0.79	12.10	0.000	0.207
grade	2	34191.98	25.26	0.000	0.264
SSL * grade	6	0.86	3.73	0.001	0.074

Table 10

Post-hoc Comparisons Among Significant RM ANOVA Main Effects

Grades	Administration time	*t*	*df*	*p* (2-tailed)
EA				
2 and 4	Time 1[a]	-9.335	95.00	0.000
	Time 2[b]	-8.815	84.57	0.000
	Time 3[b]	-8.827	82.28	0.000
	Time 4[b]	-8.050	80.57	0.000
4 and 6	Time 1[b]	-4.295	80.56	0.000
	Time 2[b]	-2.386	77.29	0.019
	Time 3[b]	-5.753	80.61	0.000
	Time 4[b]	-4.795	85.49	0.000
2 and 6	Time 1[b]	-11.589	70.94	0.000
	Time 2[b]	-8.728	63.07	0.000
	Time 3[b]	-12.805	63.80	0.000
	Time 4[b]	-11.826	66.32	0.000

(*table continues*)

Grades	Administration time	*t*	*df*	*p* (2-tailed)
MS				
2 and 4	Time 1[b]	-0.761	94.52	0.449
	Time 2[b]	-1.695	94.93	0.093
	Time 3[b]	-0.843	93.95	0.402
	Time 4[b]	-1.316	94.36	0.191
4 and 6	Time 1[a]	-2.079	93.00	0.040
	Time 1[a]	-2.729	93.00	0.008
	Time 3[b]	-1.611	85.40	0.111
	Time 4[a]	-2.182	93.00	0.032
2 and 6	Time 1[a]	-2.905	94.00	0.005
	Time 2[a]	-4.283	94.00	0.000
	Time 3[a]	-2.272	94.00	0.025
	Time 4[a]	-3.337	94.00	0.001
SSL				
2 and 4	Time 1[a]	-4.334	89.47	0.000
	Time 2[a]	-6.495	95.00	0.000
	Time 3[b]	-4.774	79.82	0.000
	Time 4[b]	-4.818	85.25	0.000
4 and 6	Time 1[b]	-3.232	67.19	0.002
	Time 1[b]	-2.302	72.85	0.024
	Time 3[b]	-3.418	81.18	0.001
	Time 4[b]	-1.280	76.48	0.205
2 and 6	Time 1[b]	-5.893	59.51	0.000
	Time 1[b]	-6.610	59.90	0.000
	Time 3[b]	-7.313	62.54	0.000
	Time 4[b]	-4.698	62.98	0.000

[a] equal variances assumed
[b] equal variances not assumed

t tests showed significant differences between Grades 2 and 4, 4 and 6, and 2 and 6 at all times for all three probes except for all four administration times of MS between Grades 2 and 4 and for Time 3 between Grades 4 and 6. Figures 1, 2, and 3 indicate that more digits were correct on average as grade level increased at each of the four times the MCBM probes were administered.

In sum, the results of the RM ANOVA indicated that for the majority of the MCBM probes administered, the mean scores increased across administrations. It also indicated that the mean scores were consistently highest for the sixth grade participants, followed by the fourth-grade participants, with the second-grade participants' scores being the lowest. This indicates that student performance improved over time on each of the MCBM probes. It is worthy to note, however, that although the scores did not always improve from week to week, the final scores were always higher than the original scores.

Concurrent Criterion-Related Validity

A Pearson product-moment correlation was conducted to determine if there were significant relationships between MCBM scores and NRTs scores. Due to stronger psychometric properties relative to individual subtest scores, component scores from the NRTs were included in the analyses when available. Results of this analysis are presented in Table 11.

As noted in Table 12, moderate correlations between MCBM probes and WJ-ACH-III and between MCBM scores and WIAT scores were found across all testing periods in all grades; correlations ranged from $r = .34$ to $r = .72$ and all correlations were

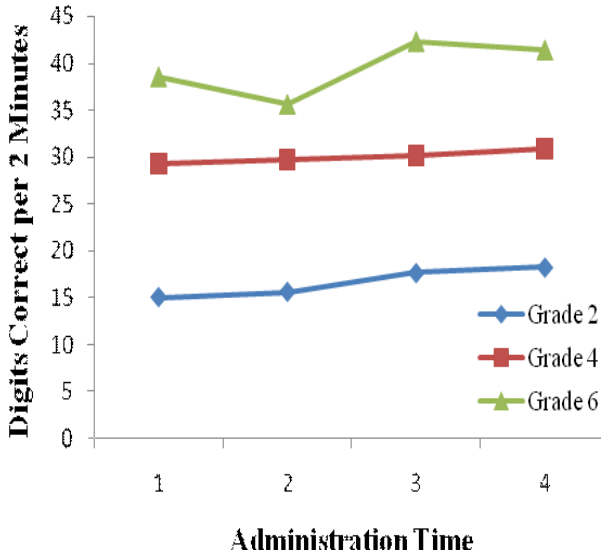


Figure 1. Marginal means for EA by grade.

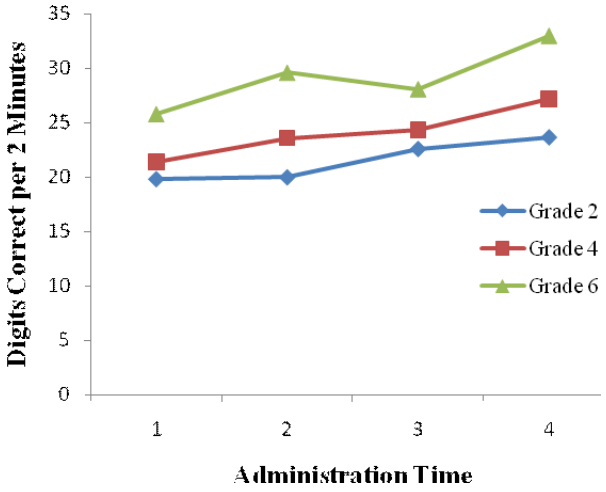


Figure 2. Marginal means for MS by grade.

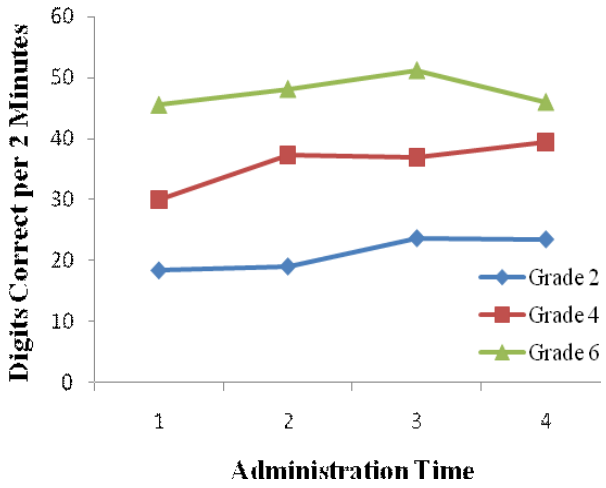


Figure 3. Marginal means for SSL by grade.

statistically significant ($p < .05$). Lower correlations were found between MCBM probes and the KM 3 subtests scores = .16 to .39), with two nonsignificant relationships identified (KM 3 Mult/Div with Time 2 MS and Time 4 MS). Statistically significant correlations were found between all probes and tests for 2nd graders at all administration times ($r = .34$ to $r = .66$). Correlations were generally weaker for fourth-grade scores relative to the other two grades between all NRTs and MCBM probes with inconsistent significant correlations across time ($r = .01$ to $r = .40$) For sixth grade, statistically significant correlations were found between all probes and tests comparisons ($r = .31$ to $r = .62$) with two exceptions: Time 3 EA and SSL compared to KM 3 Add/Sub.

Table 11

Pearson Correlation Coefficients Between NRTs

	WIAT	WJ-ACH-III	KM 3 Add/Sub
All phase II participants (N = 144)[a]			
WJ-ACH-III	.720**		
KM 3 Add/Sub	.716**	6.80**	
KM 3 Mult/Div	.588**	.539**	0.570**
Grade 2 (n = 49)			
WJ-ACH-III	.713**		
KM 3 Add/Sub	.770**	.730**	
Grade 4 (n = 48)			
WJ-ACH-III	.715**		
KM 3 Add/Sub	.740**	.636**	
KM 3 Mult/Div	.466**	.295*	.519**
Grade 6 (n = 47)			
WJ-ACH-III	.745**		
KM 3 Add/Sub	.636**	.685**	
KM 3 Mult/Div	.677**	.721*8	.610**

* correlation significant at $p < .05$.
** correlation significant at $p < .01$.
[a]$N = 144$ for all correlations except those between MCBM probes and KM 3 Mult-Div where $N = 95$.

Predictive Validity

The first research question inquired as to which performance indictor (i.e., level, slope, or dual discrepancy on each of the four MCBM probes (evaluation of mastered skill, single skill, multiple skill, and error analysis of skill steps) consistantly and accurately predicts high-risk and low-risk student performance across all testa and grades and would best differentiate math skill performance level measured by norm-referenced

Table 12

Pearson Correlation Coefficients between NRTs and MCBM Level Scores by Grade

MCBM Probe	WIAT	WJ-ACH-III	KM 3 Add/Sub	KM 3 Mult/Div
All Phase II Participants (N =144) [a]				
Time 1				
EA	.391**	.420**	.272**	.257*
MS	.442**	.588**	.394**	.237*
SSL	.453**	.450**	.297**	.323**
SSM	.503**	.523**	.360**	.302**
Time 2				
EA	.363**	.370**	.250**	.208*
MS	.393**	.536**	.322**	.162
SSL	.393**	.409**	.225**	.296**
Time 3				
EA	.338**	.406**	.270**	.219*
MS	.421**	.622**	.359**	.224*
SSL	.347**	.471**	.282**	.281**
Time 4				
EA	.412**	.409**	.256**	.228*
MS	.393**	.563**	.362**	.197
SSL	.720**	.491**	.290**	.321**
Grade 2 (n = 49)				
Time 1				
EA	.650**	.545**	.641**	
MS	.535**	.587**	.505**	
SSL	.630**	.472**	.504**	
SSM	.606**	.567**	.533**	
Time 2				
EA	.567**	.589**	.544**	
MS	.385**	.532**	.358*	
SSL	.442**	.427**	.442**	
Time 3				
EA	.551**	.658**	.612**	
MS	.488**	.626**	.428**	
SSL	.511**	.603**	.510**	

(*table continues*)

MCBM Probe	WIAT	WJ-ACH-III	KM 3 Add/Sub	KM 3 Mult/Div
Time 4				
EA	.617**	.646**	.664**	
MS	.507**	.594**	.418**	
SSL	.431**	.524**	.420**	
Grade 4 (*n* = 48)				
Time 1				
EA	.166	.452**	.252	.028
MS	.323*	.569**	.377**	.129
SSL	.406**	.528**	.378**	.202
SSM	.405**	.500**	.312*	.103
Time 2				
EA	.334*	.472**	.373**	.187
MS	.263	.482**	.294*	-.023
SSL	.270	.431**	.214	.020
Time 3				
EA	.262	.516**	.403**	.037
MS	.334*	.556**	.311*	.030
SSL	.320*	.495**	.257	.120
Time 4				
EA	.235	.442**	.243	.063
MS	.241	.520**	.255	.012
SSL	.265	.569**	.276	.056
Grade 6 (*n* = 47)				
Time 1				
EA	.612**	.783**	.407**	.602**
MS	.532**	.717**	.385**	.440**
SSL	.584**	.624**	.365*	.539**
SSM	.582**	.767**	.453**	.537**
Time 2				
EA	.373**	.479**	.207	.314*
MS	.587**	.722**	.436**	.442**
SSL	.535**	.588**	.274	.565**
Time 3				
EA	.569**	.761**	.428**	.591**
MS	.497**	.718**	.405**	.442**
SSL	.618**	.670**	.407**	.542**

(*table continues*)

MCBM Probe	WIAT	WJ-ACH-III	KM 3 Add/Sub	KM 3 Mult/Div
Time 4				
EA	.511**	.732**	.388**	.560**
MS	.506**	.671**	.507**	.459**
SSL	.486**	.579**	.343*	.558**

* correlation significant at $p < .05$.
**correlation significant at $p < .01$.
[a] N = 144 for all correlations except KM 3 Mult/Div where $N = 95$.

tests of math achievement. First, students were identified in two categories: high risk low risk. The categories were calculated four ways: NRT, level, slope, and dual discrepancy. The number of students identified for each classification by each of the NRTs is presented in Table 13. The number of students identified as high risk and low risk by the MCBM probes based on each performance indicator (i.e., level, slope, and dual discrepancy) are presented in Table 14, 15, and 16, respectively. These numbers provide an overview of the frequency of students identified as high- and low risk for each measure.

Several analyses that were conducted using this data to answer the first research question will be presented subsequently, including McNemar correlation tasks, binary classification tests, and logistic regression analyses.

McNemar Test

The first analysis conducted was the McNemar test. This statistic was conducted to determine if the proportion of students' risk status on the NRT subtests was significantly different from the prooportion of students' risk status on the MCBM probes (e.g., low risk on NRT, but high risk on MCBM). The results of the McNemar test comparing differences in identified proportions of students within the high- and low-risk

Table 13

Students' Risk Status Determined by NRTs

Test	*N*	High risk	%	Low risk	%
All Phase II Participants (*N* =144) [a]					
WIAT	144	10	6.9	134	93.1
WJ-ACH-III	144	6	4.2	138	95.8
KM 3 Add/Sub	144	25	17.4	119	82.6
KM 3	95	8	8.4	87	91.6
Mult/Div					
Grade 2 (*n* = 49)					
WIAT	49	7	14.3	42	85.7
WJ-ACH-III	49	3	6.1	46	93.9
KM 3 Add/Sub	49	10	20.4	39	79.6
Grade 4 (*n* = 48)					
WIAT	48	0	0.0	48	100.0
WJ-ACH-III	48	0	0.0	48	100.0
KM 3 Add/Sub	48	6	12.5	42	87.5
KM 3	48	3	6.3	45	93.8
Mult/Div					
Grade 6 (*n* = 47)					
WIAT	47	3	6.4	44	93.6
WJ-ACH-III	47	3	6.4	44	9.36
KM 3 Add/Sub	47	5	10.6	42	89.4
KM 3 Mult/Div	47	5	10.6	42	89.4

[a] *N* = 144 for all measures except KM 3 Mult/Div in which *N* = 95.

categories between the NRTs for all grades are presented in Table 17. Generally, all of the NRTs identified a similar proportion of students with only one exception: the KM 3 Add/Sub, which identified a significantly greater proportion in the high-risk range than the WIAT or the WJ-ACH-III.

Table 14

Students' Risk Status Determined by MCBM Level Scores

	Time 1		Time 2		Time 3		Time 4	
Measure	High risk	Low risk	High risk	Low risk	High risk	Low risk	High risk	Low risk
All Phase II participants (N =144)								
EA	19	125	23	121	19	125	19	125
MS	17	127	21	123	18	126	15	129
SSL	20	124	20	124	17	127	20	124
SSM	20	124						
Grade 2 (n = 49)								
EA	7	42	6	43	6	43	8	41
MS	8	41	10	39	4	45	7	42
SSL	8	41	7	42	8	41	8	41
SSM	9	40						
Grade 4 (n = 48)								
EA	6	42	8	40	8	40	8	40
MS	3	45	6	42	6	42	6	42
SSL	4	44	6	42	7	41	7	41
SSM	5	43	6	42				
Grade 6 (n = 47)								
EA	6	41	9	38	5	42	3	44
MS	6	41	5	42	6	41	2	45
SSL	8	39	7	40	6	41	5	42
SSM	6	41						

The results of the McNemar Test for all participants on each NRT compared with performance on MCBM probes (using level, slope, and dual discrepancy performance indicators) are presented in Table 18. First, the level performance indicators resulted in no significant differences in the proportion of students identified on MCBM probes and the WIAT, KM 3 Add/Sub, and KM 3 Mult/Div tests. However, there was a significant

difference (i.e., significantly different proportions) between the MCBM probes and the

Table 15

Students' Risk Status Determined by MCBM Slope Scores

	Time 3		Time 4	
Measure	High risk	Low risk	High risk	Low risk
All Phase II Participants (N =144)				
EA	47	97	43	101
MS	44	100	32	112
SSL	38	106	42	102
Grade 2 (n = 49)				
EA	16	33	18	31
MS	12	37	10	39
SSL	11	38	13	36
Grade 4 (n = 48)				
EA	15	33	6	42
MS	20	28	10	38
SSL	13	35	16	32
Grade 6 (n = 47)				
EA	16	31	19	28
MS	12	35	12	35
SSL	14	33	13	34

WJ-ACH-III such that the level performance indicator identified significantly more high risk students than the WJ-ACH-III for all types of MCBM probes.

Second, using the slope performance indicator resulted in significant differences in the proportion of students identified by the NRTs and those identified by the MCBM probes. In fact, only the slope scores for Time 3 SS and Time 4 MS probes identified a similar proportion of students as high risk compared to the proportion identified by the KM 3 Add/Sub. Otherwise, the slope identification method resulted in significantly

Table 16

Students' Risk Status Determined by MCBM Dual Discrepancy Scores

	Time 3		Time 4	
Measure	High risk	Low risk	High risk	Low risk
All phase II participants (N =144)				
EA	12	132	12	132
MS	11	133	9	135
SSL	8	136	14	130
Grade 2 (n = 49)				
EA	5	44	8	41
MS	3	46	4	45
SSL	1	48	4	45
Grade 4 (n = 48)				
EA	4	44	3	45
MS	4	44	4	44
SSL	4	44	5	43
Grade 6 (n = 47)				
EA	3	44	1	46
MS	4	43	1	46
SSL	3	44	5	42

more students identied as high risk on the MDBM probes than were idenfitied by the NRTs.

Finally, the proportion of students in the risk status using dual discrepancy analysis scores was compared to the proportion of students in the high-risk category using the NRTs. The results indicate that dual discrepancy method was not significantly different in the classification status that was determined using the MCBM probes and that determined using the NRTs. Overall, only the KM 3 Add/Sub was shown to be

Table 17

McNemar Correlation p Values for NRT and NRT

Test 1	Test 2	*n*	*p*
WIAT	WJ-ACH-III	144	0.125
WIAT	KM 3 Add/Sub	144	0.001
WIAT	KM 3 Mult/Div	95	0.063
WJ-ACH-III	KM 3 Add/Sub	144	<0.000
WJ-ACH-III	KM3 Mult/Div	95	0.063
KM 3 Add/Sub	KM 3 Mult/Div	95	0.118

significantly different with the KM 3 Add/Sub identifying significantly more students in the high-risk category than the MCBM probes.

Binary Classification Test

An evaluatino of the binary classification based on the MCBM probes was conducted to determine the accuracy with which each of the MCBM predictor measures identified the same students in the high-risk and low-risk categories as were identified by the NRT criterion measures. Sensitivity, specificity, positive preduction values (PPV) and negative prediction values (NPV) were also calculated using NRT scores as the criterion measure. The degree that a high risk MCBM probe score would accurately predict students who scored one standard deviation below the mean on each of the NRTs was calculated. The higher the sensitivity, the fewer high-risk students in the NRTs are undetected based on MCBM results (few false negatives when the test reports no high

Table 18

McNemar Correlation p Values for NRT and MCBM

MCBM probe	WIAT (N = 144)	WJ-ACH-III (N = 144)	KM 3 Add/Sub (N = 144)	KM 3 Mult/Div (N = 95)
Error analysis				
Time 1	0.064	0.002**	0.327	0.424
Time 2	0.011*	0.000**	0.856	0.064
Time 3	0.064	0.002**	0.307	0.267
Time 4	0.078	0.004**	0.345	0.607
Slope Time 3	0.000**	0.000**	0.004**	0.000**
Slope Time 4	0.000**	0.000**	0.018*	0.003**
DD Time 3	0.815	0.210	0.015*	1.000
DD Time 4	0.815	0.210	0.024*	0.388
Multiple skills				
Time 1		0.143		0.007**
Time 2		0.019*		0.000**
Time 3		0.152		0.012*
Time 4		0.359		0.049*
Slope Time 3		0.000**		0.000**
Slope Time 4		0.000**		0.000**
DD Time 3		1.000		0.302
DD Time 4		1.000		0.581
Single skill L				
Time 1	0.041*	0.003**	0.473	0.481
Time 2	0.064	0.003**	0.500	0.302
Time 3	0.167	0.007**	0.185	0.267
Time 4	0.087	0.007**	0.458	0.454
Slope Time 3	0.000**	0.000**	0.092	0.001**
Slope Time 4	0.000**	0.000**	0.016*	0.000**
DD Time 3	0.774	0.727	0.002**	1.000
DD Time 4	0.523	0.096	0.052	0.791
Single skill M				
Time 1	0.031*	0.001**	0.458	0.581

* $p < .05$.
** $p < .01$.

risk for a person who actually is high risk). The higher the specificity, the fewer low risk students on the NRT are labeled as high risk (few false positives). The PPV determines the likelihood that an identified high-risk student based on MCBM results actually is high risk and would likely need additional support. The NPV determines the likelihood that an identified low-risk student based on MCBM results actually is low risk and would likely not need additional support. High values (e.g., 90% probability) reflect more accurate classification.

NRT by NRT Comparisons

Before computing binary classification tests between NRTs and MCBMs, a series of pairwise comparisons between NRTs was completed to determine the amount of agreement in identified student risk status between each of the NRTs (see Figures 4, 5, 6, 7, 8, and 9). The primary purpose of these comparisons was to establish a rudimentary baseline between the rate of comparison between NRTs in order to better appreciate the rate of agreement between NRTs and MCBMs as identified through subsequent binary classification tests. As can be seen in the following figures, the results of pairwise comparisons between the WIAT, WJ-ACH-III, and KM 3 Mult/Div subtests indicate that these NRTs demonstrated moderate to strong agreement in identified risk status of the same students. However, the KM 3 Add/Sub demonstrated low agreement in identified risk status when compared to KM 3 Mult/Div subtests.

		WIAT		Total
		High Risk	Low Risk	
WJ-	High Risk	6	0	6
	Low Risk	4	134	138
Total		10	134	144

Figure 4. Risk status agreement between WIAT and WJ-ACH-III.

		WIAT		Total
		High Risk	Low Risk	
KM 3	High Risk	7	118	125
	Low Risk	3	16	19
Total		10	134	144

Figure 5. Risk status agreement between WIAT and KM 3 Add/Sub.

		WIAT		Total
		High Risk	Low Risk	
KM 3	High Risk	3	5	8
	Low Risk	0	87	87
Total		3	92	95

Figure 6. Risk status agreement between WIAT and KM 3 Mult/Div.

		WJ-ACH-III		Total
		High Risk	Low Risk	
KM 3	High Risk	4	21	25
	Low Risk	2	117	119
Total		6	138	144

Figure 7. Risk status agreement between WJ-ACH-III and KM 3 Add/Sub.

		WJ-ACH-III		Total
		High Risk	Low Risk	
KM 3	High Risk	3	5	8
	Low Risk	0	87	87
Total		3	92	95

Figure 8. Risk status agreement between WJ-ACH-III and KM 3 Mult/Div.

		KM 3 Add/Sub		Total
		High Risk	Low Risk	
KM 3	High Risk	4	4	8
	Low Risk	11	76	87
Total		15	80	95

Figure 9. Risk status agreement between KM 3 Add/Sub and KM 3 Mult/Div.

Level Performance Indicator by NRT Comparisons

The second series of binary classification tests compared participant risk status determined using level scores and risk status determined using NRT score. Overall, for grades using MCBM level scores, the results indicate that each of the NRTs demonstrated sensitivity less than 30% (range 5-29%) between MCBM probes and the WIAT and WJ-ACH-III. The KM 3 Add/Sub ranged between 25-53% and the KM 3 Mult/Div ranged between 0-36%. Specificity ranged between 84-99%. Negative prediction values were greater than .83 for all test comparisons. In contrast, PPVs were lower and more variable, ranging from .09 - .6. The WJ-ACH-III had the highest PPVs across all probes and KM 3 Add/Sub were the lowest. Overall, the results indicate that the MCBM probes performed poorly at identifying the same students classified into the high-risk range compared to those identified by the NRTs. Early administration times (Time 1 and time 2) on the EA probe, multiple skills probe, and the single skill probe consistently had the highest PPVs compared to other probes for each NRT. Alternatively, the results indicate that the MCBM probes were highly accurate in identifying the same students as low risk compared to those identified by the NRTs. Results of the binary classification test for MCBM level performance indicator and NRTs are presented in Table 19.

Slope Performance Indicator by NRT Comparisons

The third series of binary classification tests compared participant risk status determined using slope score and NRT score. Results of these analyses are presented in

Table 19

Results of Binary Classification Tests for MCBM Level Scores

Test	Sensitivity	Specificity	PPV	NPV
WIAT (*N* = 144)				
EA				
Time 1	0.263	0.960	0.500	0.896
Time 2	0.217	0.959	0.500	0.866
Time 3	0.263	0.960	0.500	0.896
Time 4	0.211	0.952	0.400	0.888
MS				
Time 1	0.294	0.961	0.500	0.910
Time 2	0.286	0.967	0.600	0.888
Time 3	0.105	0.936	0.200	0.873
Time 4	0.200	0.946	0.300	0.910
SSL				
Time 1	0.211	0.952	0.400	0.888
Time 2	0.150	0.944	0.300	0.873
Time 3	0.235	0.953	0.400	0.903
Time 4	0.053	0.920	0.091	0.865
SSM				
Time 1	0.263	0.960	0.500	0.896
WJ-ACH-III (*N* = 144)				
EA				
Time 1	0.211	0.984	0.667	0.891
Time 2	0.217	0.992	0.833	0.870
Time 3	0.211	0.984	0.667	0.891
Time 4	0.158	0.976	0.500	0.884
MS				
Time 1	0.235	0.984	0.667	0.906
Time 2	0.238	0.992	0.833	0.884
Time 3	0.111	0.968	0.333	0.884
Time 4	0.188	0.977	0.500	0.906
SSL				
Time 1	0.150	0.976	0.500	0.877
Time 2	0.150	0.976	0.500	0.877
Time 3	0.235	0.984	0.667	0.906
Time 4	0.100	0.968	0.333	0.870

(*table continues*)

Test	Sensitivity	Specificity	PPV	NPV
SSM				
Time 1	0.200	0.984	0.667	0.884
KM 3 Add/Sub (N = 144)				
EA				
Time 1	0.474	0.872	0.360	0.916
Time 2	0.391	0.868	0.360	0.882
Time 3	0.526	0.880	0.400	0.924
Time 4	0.421	0.864	0.320	0.908
MS				
Time 1	0.529	0.874	0.360	0.933
Time 2	0.476	0.878	0.400	0.908
Time 3	0.278	0.841	0.200	0.891
Time 4	0.400	0.853	0.240	0.924
SSL				
Time 1	0.368	0.856	0.280	0.899
Time 2	0.250	0.839	0.200	0.874
Time 3	0.412	0.858	0.280	0.916
Time 4	0.263	0.864	0.227	0.885
SSM				
Time 1	0.400	0.863	0.320	0.899
KM 3 Mult/Div (N = 95)				
EA				
Time 1	0.250	0.892	0.250	0.892
Time 2	0.176	0.885	0.250	0.831
Time 3	0.231	0.890	0.250	0.880
Time 4	0.091	0.869	0.083	0.880
MS				
Time 1	0.333	0.895	0.250	0.928
Time 2	0.182	0.881	0.167	0.892
Time 3	0.083	0.867	0.083	0.867
Time 4	0.000	0.862	0.000	0.904
SSL				
Time 1	0.167	0.880	0.167	0.880
Time 2	0.231	0.890	0.250	0.880
Time 3	0.231	0.890	0.250	0.880
Time 4	0.182	0.881	0.167	0.892
SSM				
Time 1	0.364	0.905	0.333	0.916

Table 20. Overall, for all participants using MCBM slope scores, the results indicate that each of the probe administrations demonstrated sensitivity less than 15% (range 2-14)) between MCBM probes and the WIAT, WJ-ACH-III, and KM 3 Mult/Div. The KM 3 Add/Sub ranged between 10-26%. Specificity ranged between 82-98% for all NRTs.

Negative prediction values ranged from .64-.80 for all test comparisons. In contrast, PPVs were lower and more variable, but a general bicameral distribution was noted with the WIAT, WJ-ACH-III, and KM 3 Mult/Div comprising the lower range and the KM 3 Add/Sub comprising the upper range. Specifically, the WIAT, WJ-ACH-III, and KM 3 Mult/Div ranged from .13-.36 and the KM 3 Add/Sub ranged from .38-.67. The WJ-ACH-III had the highest PPVs on the single skill and multiple skill probes; the KM 3 Mult/Div had the highest PPVs on the error analysis probe on Time 3, but the KM 3 Add/Sub probe had the highest PPVs on the error analysis probe at Time 4. The WJ-ACH-III had the lowest PPV on the Time 3 error analysis probe; the KM 3 Mult/Div had the lowest PPV on the Time 4 error analysis and Time 4 multiple skills probes; the KM 3 Add/Sub had the lowest PPV on the Time 3 single skill probe; finally, the WIAT had the lowest PPV on the Time 4 single skill probe. Overall, the NPVs were less varied with the KM 3 Add/Sub having the highest NPV on all probes except the Time 4 error analysis (KM 3 Mult/Div had the highest) and the Time 3 single skill (the WJ-ACH-III had the highest). The KM 3 Mult/Div had the lowest NPVs on all probes except both error analysis probes (the WIAT had the lowest NPVs on both).

Generally, the results for slope scores indicate that the MCBM probes performed poorly at identifying the same students classified into the high risk range compared to those identified by the NRTs. The SS probe consistently had the highest PPVs and the

Table 20

Results of Binary Classification Tests for MCBM Slope Scores

Test	Sensitivity	Specificity	PPV	NPV
WIAT (*N* = 144)				
EA				
Time 3	0.043	0.918	0.200	0.664
Time 4	0.047	0.921	0.200	0.694
MS				
Time 3	0.068	0.930	0.300	0.694
Time 4	0.094	0.938	0.300	0.784
SSL				
Time 3	0.079	0.934	0.300	0.739
Time 4	0.095	0.941	0.400	0.716
WJ-ACH-III (*N* = 144)				
EA				
Time 3	0.021	0.948	0.167	0.667
Time 4	0.023	0.950	0.167	0.696
MS				
Time 3	0.045	0.960	0.333	0.696
Time 4	0.063	0.964	0.333	0.783
SSL				
Time 3	0.079	0.972	0.500	0.746
Time 4	0.095	0.980	0.667	0.725
KM 3 Add/Sub (*N* = 144)				
EA				
Time 3	0.191	0.835	0.360	0.681
Time 4	0.186	0.832	0.320	0.706
MS				
Time 3	0.182	0.830	0.320	0.697
Time 4	0.250	0.848	0.320	0.798
SSL				
Time 3	0.158	0.821	0.240	0.731
Time 4	0.262	0.863	0.440	0.739
KM 3 Mult/Div (N = 95)				
EA				
Time 3	0.097	0.922	0.375	0.678
Time 4	0.040	0.900	0.125	0.724

(*table continues*)

Test	Sensitivity	Specificity	PPV	NPV
MS				
Time 3	0.031	0.889	0.125	0.644
Time 4	0.091	0.918	0.250	0.770
SSL				
Time 3	0.074	0.912	0.250	0.713
Time 4	0.138	0.939	0.500	0.713

EA probe consistently had the lowest PPVs. In contrast, however, the results indicate that the MCBM probes were highly accurate in identifying the same students as low risk compared to those identified by the NRTs.

Dual Discrepancy Performance Indicator by NRT Comparisons

The final series of binary classification tests compared participant risk status determined using dual discrepancy scores and NRT scores. Results of these analyses are presented in Table 21.

Overall, for all participants using MCBM dual discrepancy scores, the results indicate that each of the probe administrations demonstrated sensitivity less than 50% (range 0-50%) between MCBM probes and all NRTs. Specificity ranged between 82-98% for all NRTs. Negative prediction values were greater than .90 for all test comparisons.

Positive prediction values were varied across probe types and NRTs. The KM 3 Mult/Div had the highest PPV on the Time 3 EA and Time 4 SSL probes, the WIAT had the highest PPV on the Time 4 EA and MS probes, and the WJ-ACH-III had the highest PPV on the Time 3 MS and SSL probes. The lowest PPVs were recorded for the

Table 21

Results of Binary Classification Tests for MCBM Dual Discrepancy Scores

Test	Sensitivity	Specificity	PPV	NPV
WIAT (*N* = 144)				
EA				
Time 3	0.167	0.939	0.200	0.925
Time 4	0.167	0.939	0.200	0.925
MS				
Time 3	0.091	0.932	0.100	0.925
Time 4	0.222	0.941	0.200	0.948
SSL				
Time 3	0.375	0.949	0.300	0.963
Time 4	0.071	0.931	0.100	0.903
WJ-ACH-III (*N* = 144)				
EA				
Time 3	0.083	0.962	0.167	0.920
Time 4	0.083	0.962	0.167	0.920
MS				
Time 3	0.091	0.962	0.167	0.928
Time 4	0.111	0.963	0.167	0.942
SSL				
Time 3	0.375	0.978	0.500	0.964
Time 4	0.071	0.962	0.167	0.906
KM 3 Add/Sub (*N* = 144)				
EA				
Time 3	0.500	0.856	0.240	0.950
Time 4	0.333	0.8441	0.160	0.933
MS				
Time 3	0.091	0.820	0.040	0.916
Time 4	0.444	0.844	0.160	0.958
SSL				
Time 3	0.375	0.838	0.120	0.958
Time 4	0.429	0.854	0.240	0.933
KM 3 Mult/Div (*N* = 95)				
EA				
Time 3	0.286	0.932	0.250	0.943
Time 4	0.000	0.912	0.000	0.954

(*table continues*)

Test	Sensitivity	Specificity	PPV	NPV
MS				
Time 3	0.000	0.908	0.000	0.908
Time 4	0.000	0.911	0.000	0.943
SSL				
Time 3	0.286	0.932	0.250	0.943
Time 4	0.200	0.929	0.250	0.908

WJ-ACH-III on the Time 3 EA, the KM 3 Mult/Div on the Time 4 EA and both MS probes, the KM 3 Add/Sub on the Time 3 SSL probe, and the WIAT for the Time 4 SSL probe.

Generally, the results for dual discrepancy scores indicate that the MCBM probes performed poorly to modestly at identifying the same students classified into the high risk range compared to those identified by the NRTs. However, it would appear that the dual discrepancy analysis yielded a more accurate identification procedure than the other performance indicators. The SSL probe consistently had the highest PPVs and the MS probe consistently had the lowest PPVs. Again, the results indicate that the MCBM probes were highly accurate in identifying the same students as low risk compared to those identified by the NRTs.

Logistic Regression Analyses for a Multiple Test Model

A second question important to the current study inquired about combinations of MCBM tests as a screening strategy. To explore which combination of MCBM screening measures best predicts low student performance status on individually administered standardized achievement tests, a series of logistic regression analyses (LRA) was

performed; the NRT risk status was the DV and the MCBM risk status (as determined by level, slope, and dual discrepancy) on each of the administration times of each of the MCBM probes were the predictor variables. The variables were entered using a forward stepwise procedure.

LRAs Between NRTs and Level Performance Indicators

The first series of LRAs were run with all phase II participants included; results are found in Table 22. The first LRA compared all phase II participants' risk status as determined by the WIAT with that determined by each of the MCBM probes using the level method. A total of 144 cases were analyzed and the full model containing two steps significantly predicted risk status, $\chi^2 = 4.28$, df = 1, $p = .039$. The model accounted for between 11.2-28.3% of the variance in risk status. Overall, 93.8% of the cases' risk statuses were correctly predicted with 100% of the low-risk and 40% of the high-risk cases correctly predicted. Table 23 gives coefficients and the Wald statistic and associated degrees of freedom and probability values for each of the predictor variables. This shows that only Time 1 SSM and Time 1 EA probes reliably predicted risk status.

The second LRA compared all phase II participants' risk status determined by the WJ-ACH-III with that determined by each of the MCBM using the level method. A total of 144 cases were analyzed and the full model containing two steps significantly predicted risk status, $\chi^2 = 4.45$, df = 1, $p = .035$; see Table 22 for results. The model accounted for between 12.8% and 43.6% of the variance in risk status. Overall, 95.8% of the cases' risk statuses were correctly predicted with 100% of the low-risk and 0% of the high-risk cases correctly predicted. Table 23 gives coefficients and the Wald statistic and

Table 22

Results of the LRAs with NRT and MCBM Level Scores

	Omnibus model results					Prediction accuracy			
Analysis	Steps	χ^2	df	*p*	Accounted variance %	Overall %	Low risk %	High risk %	Contributing variable(s)
WIAT	2	4.28	1	.039	11.2 – 28.3	93.1	100	0	Time 1 SSM, Time 1 EA
WJ-ACH-III	2	4.45	1	.035	12.8 – 43.6	95.8	100	0	Time 2 EA, Time 2 MS
KM 3 Add/Sub	2	5.96	1	.015	13.5 – 22.4	82.6	100	0	Time 1 MS, Time 3 EA
KM 3 Mult/Div	1	6.88	1	.009	7.0 – 15.9	91.6	100	0	Time 3 EA

Note. $N = 144$ for all measures except KM 3 Mult/Div where $N = 95$.

Table 23

Predictor Variable Statistics Included in LRAs with NRT and MCBM Level Scores

Variables	Coefficients	Wald stat.	df	*p*
WIAT				
Time 1 SSM	-2.186	8.635	1	0.003
Time 1 EA	-1.624	4.547	1	0.033
WJ-ACH-III				
Time 2 MS	-2.549	4.198	1	0.040
Time 2 EA	-2.338	3.519	1	0.061
KM 3 Add/Sub				
Time 3 EA	-1.657	8.118	1	0.004
Time 1 MS	-1.535	6.322	1	0.012
KM 3 Mult/Div				
Time 3 EA	-2.159	7.474	1	0.006

Note. $N = 144$ for all measures except KM 3 Mult/Div where $N = 95$.

associated degrees of freedom and probability values for each of the predictor variables. This shows that only the Time 1 EA and Time 2 MS probes reliably predicted risk status.

The third LRA compared all phase II participants' risk status determined by the KM 3 Add/Sub with that determined by each of the MCBM using the level method (see Table 22). A total of 144 cases were analyzed and the full model containing two steps significantly predicted risk status, $\chi^2 = 5.96$, df = 1, p,= .015. The model accounted for between 13.5-22.4% of the variance in risk status. Overall, 84.0% of the cases' risk statuses were correctly predicted with 100% of the low-risk and 0% of the high-risk cases correctly predicted. Table 23 gives coefficients and the Wald statistic and associated degrees of freedom and probability values for each of the predictor variables. This shows that only the Time 1 MS and Time 3 EA probes reliably predicted risk status.

The fourth LRA compared all fourth- and sixth-grade phase II participants' risk status determined by the KM 3 Mult/Div with that determined by each of the MCBM using the level method. Ninety-five cases were analyzed and the full model containing one step significantly predicted risk status, $\chi^2 = 6.88$, df = 1, $p = .009$; see Table 22). The model accounted for between 7.0-15.9% of the variance in risk status. Overall, 91.6% of the cases' risk statuses were correctly predicted with 100% of the low-risk and 0% of the high-risk cases correctly predicted. Table 23 gives coefficients and the Wald statistic and associated degrees of freedom and probability values for each of the predictor variables. This shows that only the Time 3 EA probe predicted risk status.

LRAs Between NRTs and Slope Performance Indicators

The next series of analyses were run with all phase II participants included. The LRAs compared all phase II participants' risk status as determined by each of the NRTs with that determined by each of the MCBM slope scores. In total, 144 cases were analyzed (except for the analysis with the KM 3 Mult/Div, which contained 95 cases). The full models for the LRAs with the WIAT, KM 3 Add/Sub, and KM 3 Mult/Div were not found to be significant. This indicates that no single MCBM probe provided adequate statistical value to account for a significant portion of the variance within student performance. Thus, no probe could be identified as particularly capable of identifying students as high risk relative to another. Table 24 presents the results of this analysis.

The LRA between student risk status determined by slope scores and the WJ-ACH-III was the only model that was significant. The full model containing one step significantly predicted risk status, $\chi^2 = 3.78$, df = 1, $p = .05$. The model accounted for between 2.6-8.8% of the variance in risk status. Overall, 95.8% of the cases' risk statuses were correctly predicted with 100% of the low-risk and 0% of the high-risk cases correctly predicted. Only the Time 4 SSL reliably predicted risk status (coefficient = -1.661, Wald = 3.508, df = 1, $p = .061$).

LRAs Between NRTs and Dual Discrepancy Performance Indicators

The last series of analyses were run with all phase II participants included. The LRAs in this series compared all phase II participants' risk status as determined by the NRTs with that determined by each of the MCBM probes using the dual discrepancy

method. The first analysis used the WIAT as the outcome variable. In total, 144 cases were analyzed and the full model containing one step significantly predicted risk status, Table 24

Results of the LRAs with NRT and MCBM Slope Scores

	Omnibus model results					Prediction accuracy			
Analysis	Steps	χ^2	df	*p*	Accounted variance %	Overall %	Low risk %	High risk %	Contributing variable(s)
WIAT	0				93.1				None
WJ-ACH-III	1	3.78	1	.05	2.6 – 8.8	95.8	100	0	Time 4 SSL
KM 3 Add/Sub	0				82.6				None
KM 3 Mult/Div	0				91.6				None

Note. $N = 144$ for all measures except KM 3 Mult/Div where $N = 95$.

$\chi^2 = 6.88$, df = 1, $p = .009$; results are presented in Table 25. The model accounted for between 4.7-11.8% of the variance in risk status. Overall, 93.1% of the cases' risk statuses were correctly predicted with 100% of the low-risk, and 0% of the high-risk cases correctly predicted. Table 26 gives coefficients and the Wald statistic and associated degrees of freedom and probability values for the predictor variable. This shows that only Time 3 SSL probe reliably predicted risk status.

The next LRA compared all phase II participants' risk status determined by the WJ-ACH-III with that determined by each of the MCBM using the dual discrepancy method; see Table 25 for results. In total, 144 cases were analyzed and the full model containing one step significantly predicted risk status, $\chi^2 = 10.48$, df = 1, $p = .001$. The

Table 25

Results of the LRAs with NRT and MCBM Dual Discrepancy Scores

	Omnibus model results					Prediction accuracy			
Analysis	Steps	χ^2	df	*p*	Accounted variance %	Overall %	Low risk %	High risk %	Contributing variable(s)
WIAT	1	6.88	1	0.009	4.7 – 11.8	93.1	100	0	Time 3 SSL
WJ-ACH-III	1	10.48	1	0.001	7.0 – 24.0	95.8	100	0	Time 3 SSL
KM 3 Add/Sub	1	7.51	1	0.006	5.1 – 8.4	82.6	100	0	Time 3 EA
KM 3 Mult/Div	1	2.71	1	0.099	2.8 – 6.4	91.6	100	0	Time 3 EA

Note. $N = 144$ for all measures except KM3 Mult/Div where $N = 95$.

Table 26

Predictor Variable Statistics Included in LRAs with NRT and MCBM Dual Discrepancy Scores

Analysis	Variable	Coefficients	Wald stat.	df	*p*
WIAT	Time 3 SSL	-2.403	8.443	1	.004
WJ-ACH-III	Time 3 SSL	-3.281	12.314	1	.000
KM 3 Add/Sub	Time 3 EA	-1.783	8.052	1	.005
KM 3 Mult/Div	Time 3 EA	-1.699	3.283	1	.070

Note. $N = 144$ for all measures except KM 3 Mult/Div where $N = 95$.

model accounted for between 7.0-24.0% of the variance in risk status. Overall, 95.8% of the cases' risk statuses were correctly predicted with 100% of the low-risk and 0% of the high-risk cases correctly predicted. Table 26 gives coefficients and the Wald statistic and associated degrees of freedom and probability values for the predictor variable. This shows that only Time 3 SSL probe reliably predicted risk status.

The next LRA compared all phase II participants' risk status determined by the KM 3 Add/Sub with that determined by each of the MCBM using the dual discrepancy method. In total, 144 cases were analyzed and the full model containing one step significantly predicted risk status, $\chi^2 = 7.51$, df = 1, $p = .006$. Table 25 presents the results of this analysis. The model accounted for between 5.1-8.4% of the variance in risk status. Overall, 82.6% of the cases' risk statuses were correctly predicted with 95.0% of the low-risk and 24.0% of the high-risk cases correctly predicted. Table 26 gives coefficients and the Wald statistic and associated degrees of freedom and probability values for each of the predictor variables. This shows that only the Time 3 EA probe reliably predicted risk status.

The final LRA compared all fourth- and sixth-grade phase II participants' risk status determined by the KM 3 Mult/Div with that determined by each of the MCBM using the dual discrepancy method. In total, 95 cases were analyzed; the full model containing one step did not significantly predict risk status, $\chi^2 = 2.71$, df = 1, $p = .099$. Table 25 presents the results of this analysis. The model accounted for between 2.8-6.4% of the variance in risk status. Overall, 91.6% of the cases' risk statuses were correctly predicted with 100% of the low-risk and 0% of the high-risk cases correctly predicted. Table 26 gives coefficients and the Wald statistic and associated degrees of freedom and probability values for each of the predictor variables. This

shows that only the Time 3 EA was included in the model, but this probe did not reliably predicted risk status.

Examination of Individual Changes in Risk Status Across NRTs and MCBM Probes

Given that few students were identified as high risk in this study on the NRTs, the data were further explored by examining individual differences across test scores for an individual who was classified as high risk on any one of the three NRTs. Figure 10 presents a graph of the number of NRTs that agreed on high risk status on all of the 31 students identified by at least one NRT. As the graph illustrates, on 4 students were identified as high risk on all NRTs and only 3 were identified as high risk on three NRTs, the majority of students identified as high risk on any of the NRTs were identified on only one NRT (n = 14) and two NRTs (n = 10).

Figure 11 presents the percentage of MCBM probes that corresponded with the NRTs' high-risk classification according to the number of NRTs agreeing on risk status per student. For example, of the 14 students identified as high risk by only one NRT, 23.6% of level probes, 22.6% of slope probes, and 9.5% of dual discrepancy probes also indicated that student was at high risk. These data indicate that as the number of NRTs indiciating high-risk status increases, so does the overall percentage of MCBM probes indicating high-risk status. The graph also illustrates the fact that the level method resulted in higher percentages of high-risk status across probes when one and three NRTs agree on high-risk status and that the slope performance indicator had the highest percentage of probes agreeing with NRT risk-status when two and four NRTs agreed on student-risk status.

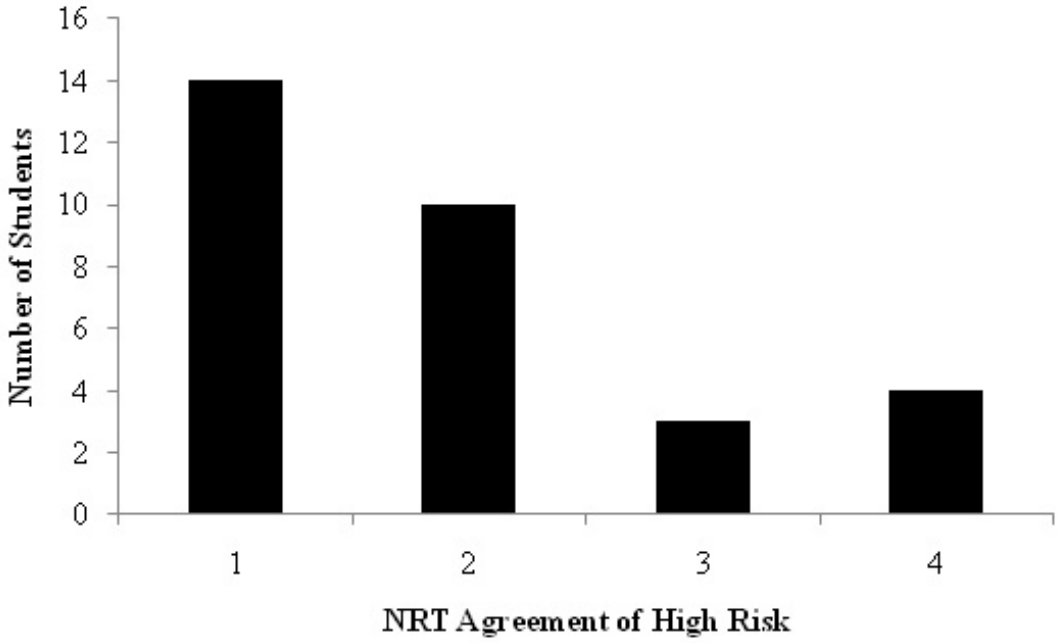


Figure 10. Number of NRTs identifying the same student as high risk.

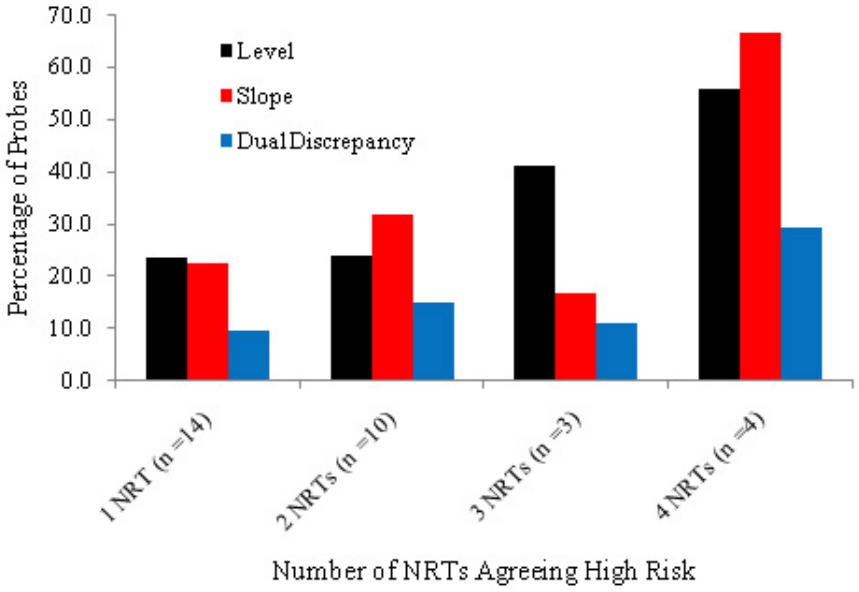


Figure 11. Percentage of agreement of high risk status between MCBM performance indicators and NRTs.

Figure 12 provides a graphical representation of the number of students identified as high risk during times 1, 2, 3, and 4 for the EA, MS, SSL, and SSM probes using the level performance indicator in comparison with the number of students identified as high risk on the NRTs. Figures 13 and 14 provide a graphical representation of the number of students identified as high risk during weeks 3 and 4 for the EA, MS, and SSL probes using the slope and dual discrepancy performance indicators, respectively, in comparison with the number of students identified as high risk on the NRTs. As illustrated in these figures, KM 3 Add/Sub identified more kids than any other test using any measurement strategy. These figures also indicate that the greatest number of students was identified as high risk using the slope method, followed by the level method, and finally, the dual discrepancy method. Further, the dual discrepancy method of risk identification resulted in the greatest similarity among all MCBM probes compared to the NRTs.

Figures 15, 16, and 17 illustrate the amount of agreement between high-risk status determined using the NRTs and that determined using the level, slope, and dual discrepancy performance indicators, respectively, for MCBM probes. Figure 15, for example, displays the number of students that matched with 0, 1, 2, 3, or all 4 NRT tests for three MCBM measures based on the level criterion. As can be seen in the following three figures, much variability exists among the measures. However, the SSL probe showed the greatest number of NRT tests matching the MCBM high-risk classification across all three identification strategies. Alternatively, Time 4 EA had the least matches with only 0 and 1 NRT tests matching the MCBM high risk classification.

Figures 18, 19, and 20 represent the number of students identified as high risk by at least one of the Time 3 and Time 4 MCBM probes but were *not* identified as high risk

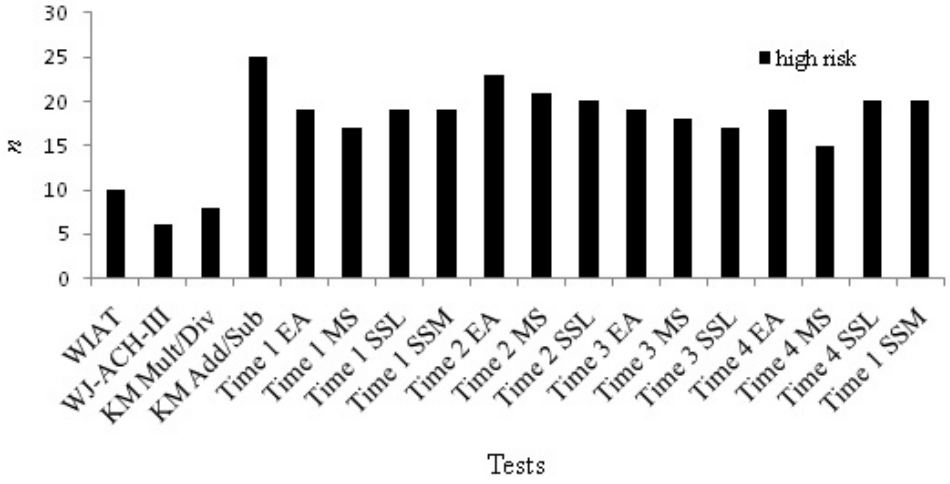


Figure 12. Frequency of high risk status determined by NRTs and level scores.

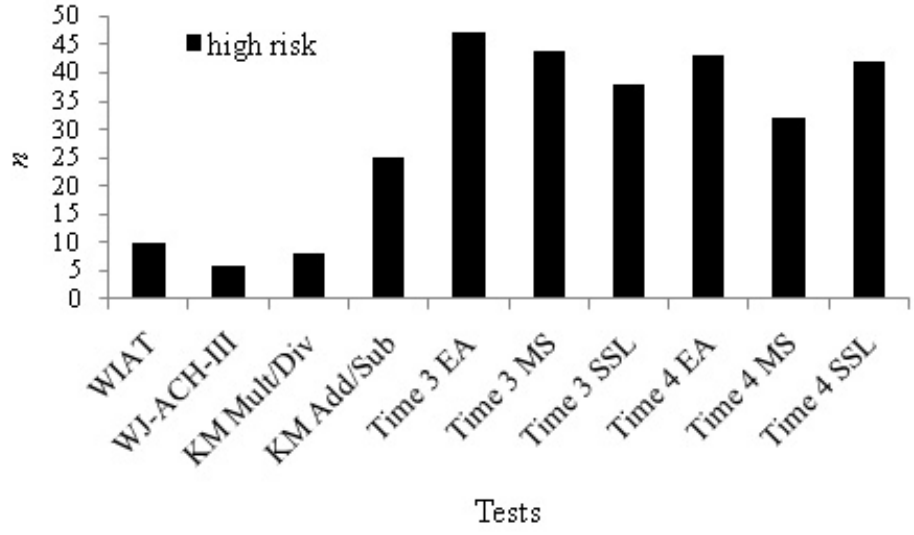


Figure 13. Frequency of high risk status determined by NRT and slope scores.

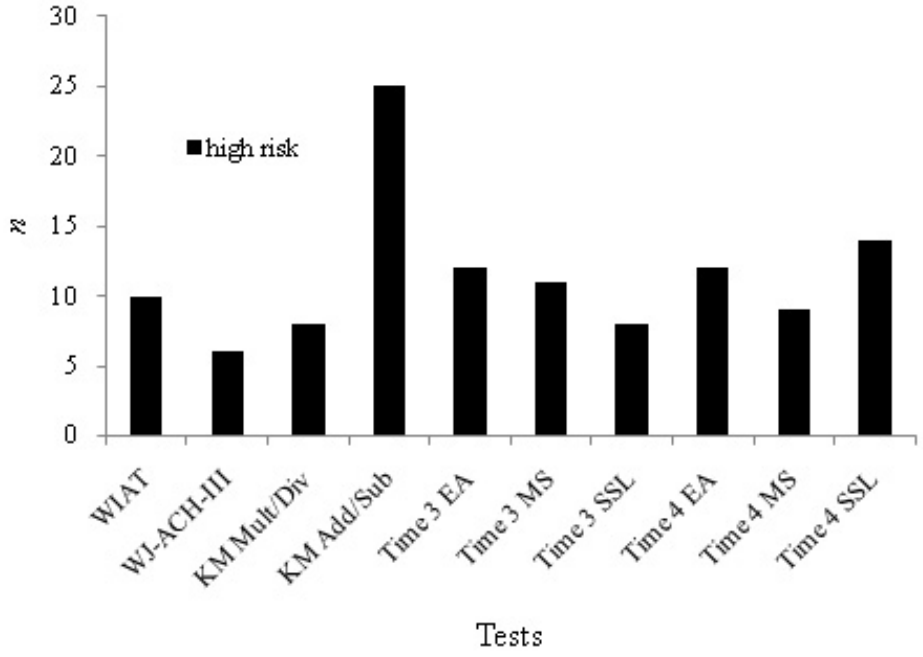


Figure 14. Frequency of high risk status determined by NRT and dual discrepancy scores.

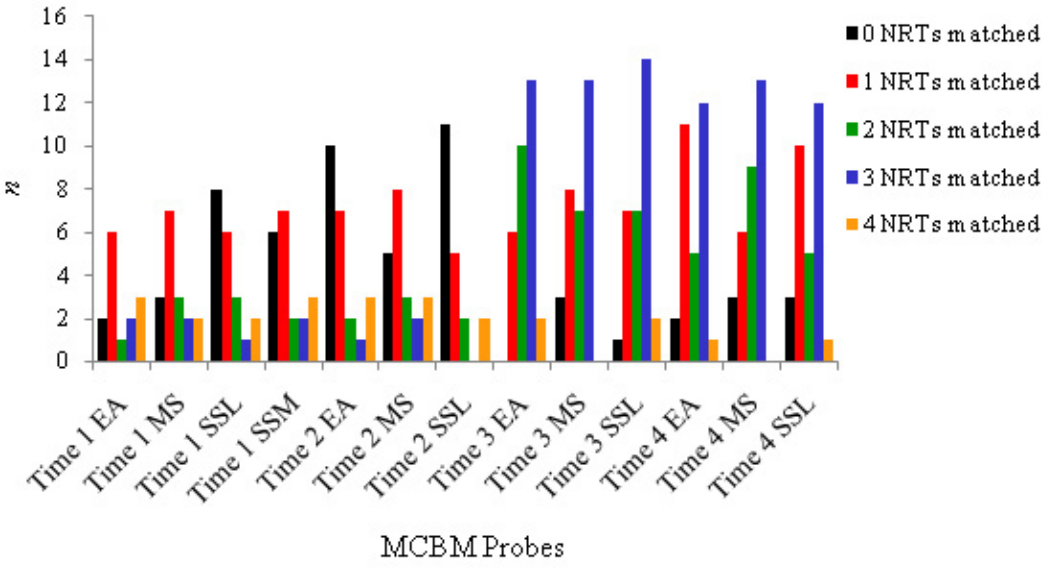


Figure 15. Agreement of high risk students identified by NRTs and level scores.

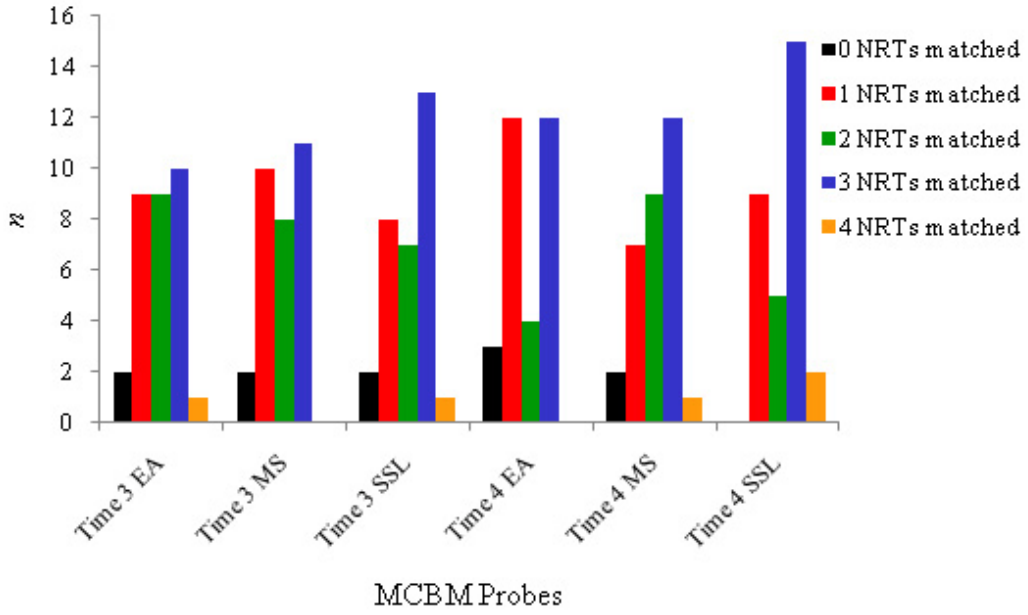


Figure 16. Agreement of high risk students identified by NRTs and slope scores.

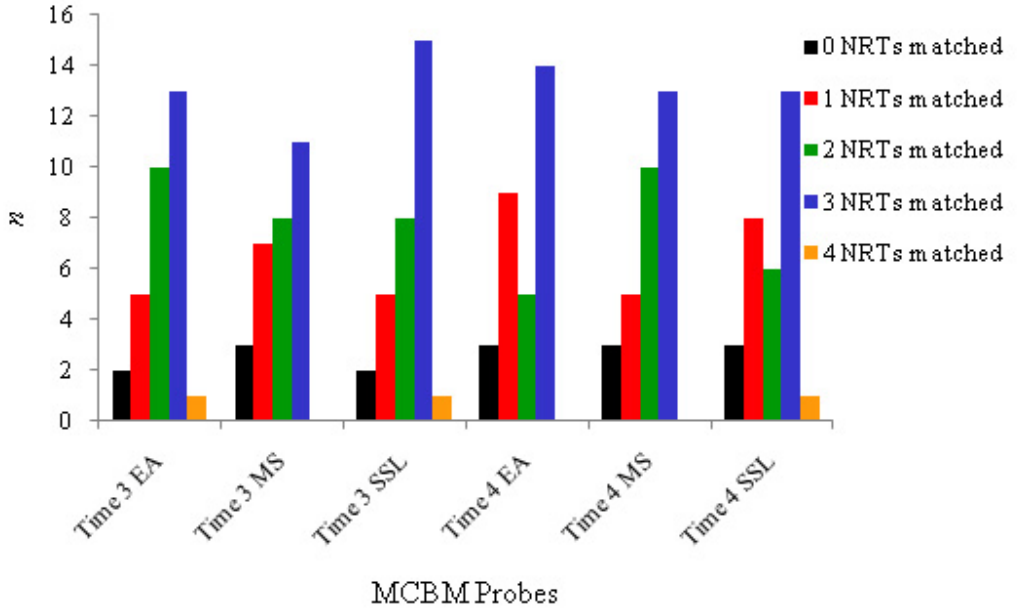


Figure 17. Agreement of high risk students identified by NRTs and dual discrepancy scores.

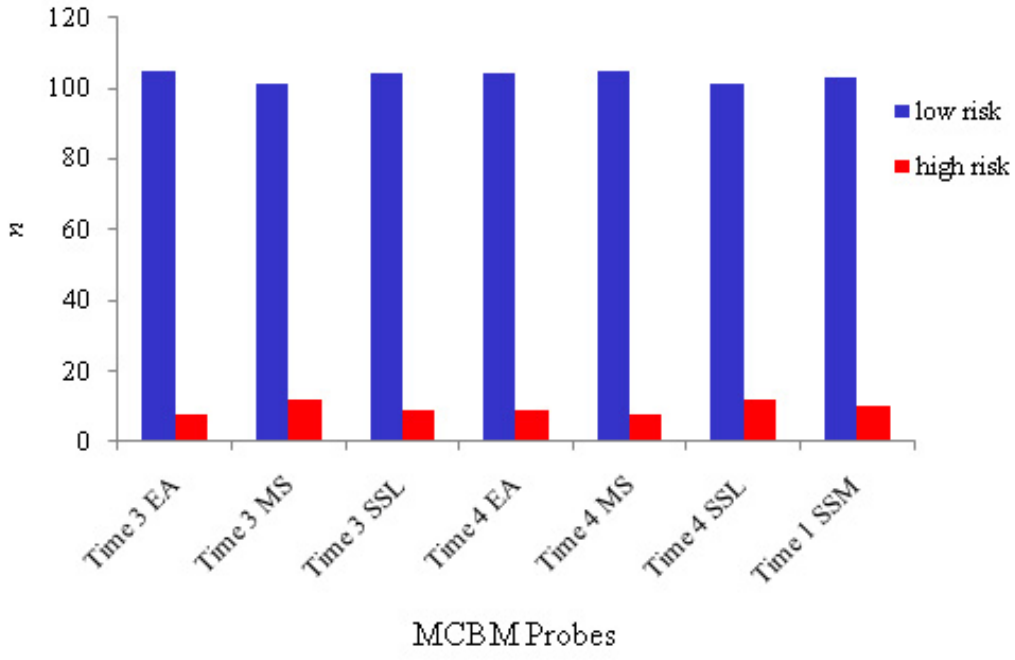


Figure 18. Number of students identified low risk by all NRTs and high and low risk on at least one MCBM level score.

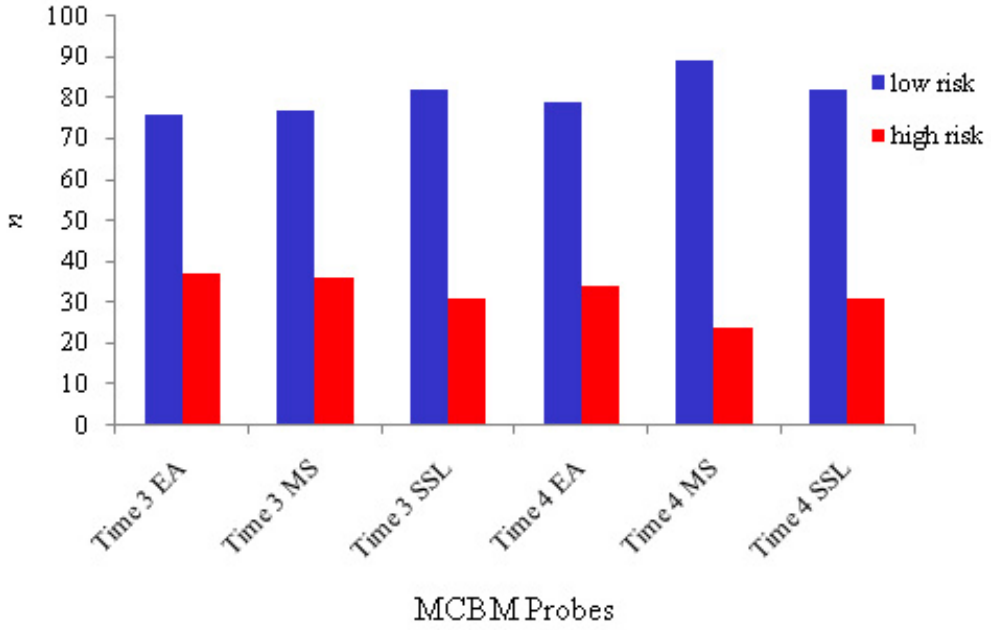


Figure 19. Number of students identified low risk by all NRTs and high and low risk on at least one MCBM slope score.

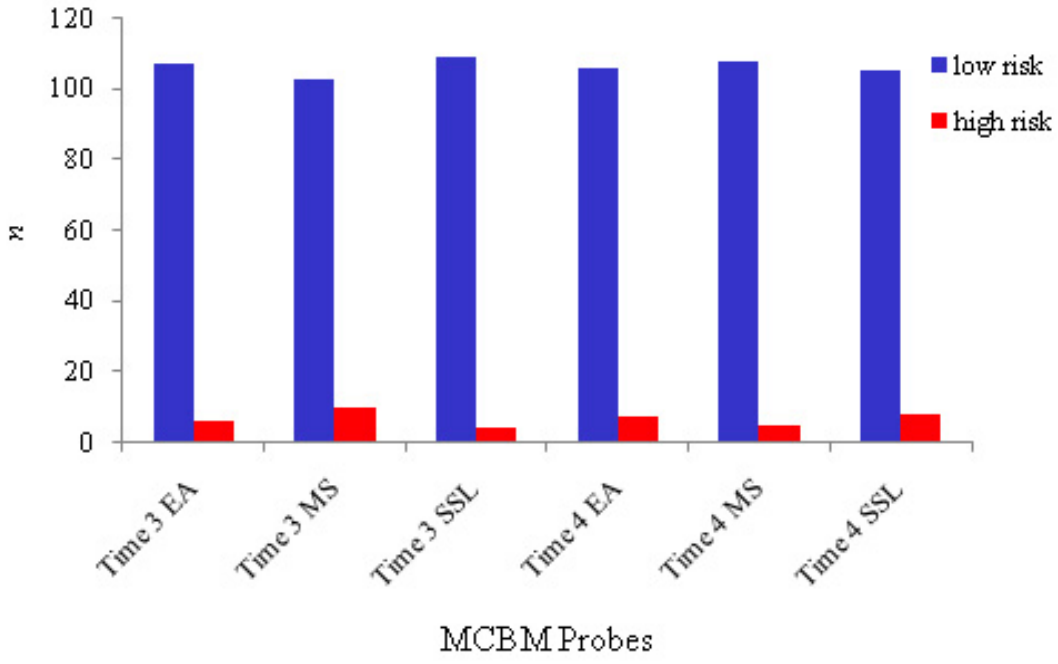


Figure 20. Number of students identified low risk by all NRTs and high and low risk on at least one MCBM dual discrepancy score.

on any NRTs. Overall, the graphs show that the slope performance indicator tended to overidentify kids on all probes relative to level and dual discrepancy procedures. The level strategy identified a range of 8 to 12 students for level, the dual discrepancy strategy identified a range of 4 to 10 students, and the slope strategy identified a range of 24 to 37 students.

Finally, Figure 21 presents the number of students identified as high risk by the Time 3 and Time 4 MCBM probes who were also identified as low risk on all NRTs. Again, these graphs indicate that the slope strategy identified more students as high risk that were not concurrently identified as high risk on any NRT. The dual discrepancy method resulted in the greatest similarity of agreement on high risk status between MCBMs and NRTs.

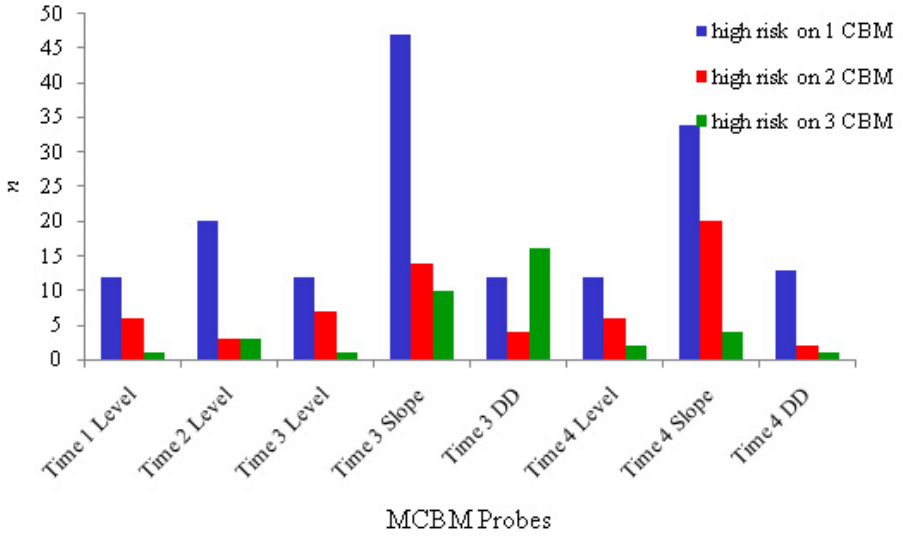


Figure 21. Number of students identified high risk on MCBM and low risk on all NRTs.

CHAPTER V

DISCUSSION

Overview

Research into the utility and effectiveness of MCBM strategies to improve academic assessment and intervention has gained substantial attention in the past decade, but the majority of this research has focused on reading, with little attention provided to mathematics (Badian, 1999; Daly & McCurdy, 2002). Research in the area of mathematics is warranted given estimates that math disabilities occur almost as frequently in children as reading disabilities occur (Jitendra et al., 2005; Mazzocco & Myers, 2003; Shalev, 2004). Further, multiple studies have indicated that frequent progress monitoring has the potential to identify academic problems when they first emerge in order to correct identified deficits and to avoid any serious delays in academic progress (Hintze et al., 2002), but few researchers have investigated the utility of CBM procedures to adequately identify high risk students in the area of mathematics (Shinn, 1989, 1998).

This study sought to extend the current literature on CBM screening assessments by examining the relationship between CBM and NRT performances as well as the decision utility of CBM for predicting high- and low-risk students on standardized tests beyond first grade. This question remains an important and valid area for empirical study because of the need for increased efficiency and effectiveness in alternative methods for identifying students in need of intense specialized instruction (Stecker & Fuchs, 2000). The need for accurate and cost-effective screening methods used to identify children who are at high risk for failure on educational measures in mathematics remains potent in today's educational systems (Fletcher et al., 2005). However, current research on the effectiveness of these methods within mathematics remains sparse (Badian, 1999; Daly & McCurdy, 2002). In this way, the present study sought to contribute to the current empirical literature by evaluating the effectiveness of CBM screening

tests to identify students at high risk for failure on mathematics calculation subtests of three individually administered NRTs.

Specifically, student performance on four different types of CBM consisting of computation skills to be taught over 9 weeks were compared to student performance on three different NRTs (Fuchs et al., 2007). Further, performance was monitored and examined over time; this differs from most research in that the majority of previous research on the utility of CBM assessments has been conducted on CBM performance at one point in time to determine its utility in identifying high- and low-risk students. The present study focused on which types of screening performance indicators (i.e., level, slope, or dual discrepancy) on various types of survey level MCBM probes administered over a 9-week period consistently and accurately predicted low student performance on several subtests of three different NRTs. Incorporating three NRTs as criterion variables and using a multiple-week system of CBM administration furthers the empirical literature and differs from previously discussed research (e.g., Foegen & Deno, 2001; Helwig et al., 2002).

Another important contribution of this study was the inclusion of outcomes related to three different types of NRTs that are often used as part of the evaluation process for learning disabilities in mathematics computation (Magyar et al., 2007). Although similar to the study described previously by Helwig and colleagues (2002), the current study expanded their contributions by using three NRT measures as criterion variables instead of two NRT measures. Further, the tests utilized in their study were not nationally normed, standardized tests commonly used to make important educational decisions. Overall, the results of the Helwig and colleagues' study showed that MCBM measures had moderate to strong relationships with the NRT measures (range = .53 to .72), suggesting that similar construct or computational skills are measured on

both NRTs and CBM probes. The results of this study also showed that MCBM measures had moderate relationships with the WJ-ACH-III and WIAT subtests with somewhat lower relationships with the KM 3 Add/Sub and KM 3 Mult/Div subtests. In particular, the multiple skill and the mastered single skill MCBM measures had slightly stronger relationships with the NRTs. This indicates that student performance on the MCBM probes generally corresponded to student performance on the WIAT and WJ-ACH-III, but corresponded less well on the KM 3 subtests. Moreover, time of administration of the MCBM probes appeared not to have a strong impact on the strength of the correlation between the predictor and the criterion measures. Given that NRTs had more types of problems represented on their tests, required generalization of skills, and possibly had poorer content overlap with local curriculum, it was not surprising that the correlations between MCBM and NRTs were moderate. In general, these findings were consistent with previous research comparing multiple MCBM computational probes on skills taught during the entire school year with a statewide math competency test and SAT scores (Fuchs et al., 2007; Helwig et al.; Magyar et al.; Shapiro et al., 2006).

Among the statistical analyses used in the study, the McNemar test was conducted to identify if two measures, when compared together, identify a similar proportion of students in the same category (i.e., high and low risk). The results of the pairwise analysis of the NRT by NRT McNemar correlations indicated that each of the NRTs identified a similar proportion of students as high risk except the KM 3 Add/Sub, which identified a significantly higher proportion of students than the other three. Additional analysis of the data indicated that the MCBM probes consistently identified a similar proportion of students in the high- and low-risk categories as did the NRTs across the majority of administrations and for the majority of the probe types using the level and dual discrepancy classification strategies, but not using the slope performance indicator (which resulted in a higher proportion of identified students overall). This appeared to provide some confidence that the predictor probes could be utilized as a screening assessment tool to accurately identify a similar proportion of students in the high- and low-risk categories as the NRTs did. However, the results of the McNemar test provided information solely upon the similarities of the proportions of students identified into each risk category. These results did not provide information on the accuracy with which the same students were identified into each category by both the predictor and criterion measures. Thus, although correlations derived from the McNemar tests were encouraging, binary classification tests were conducted to determine the accuracy with which each MCBM probe identified the same students as the NRTs did.

A series of binary classification tests were used to examine the predictive power (or value) estimates of all MCBM performance indicators across the three NRTs. For screening purposes, negative predictive power estimates should be higher than positive predictive power,

thus indicating that students who are not exhibiting a serious problem are accurately identified across tests. Strong negative predictive power estimates were found across all MCBM probes. Thus, the likelihood of students in the low-risk group that would score in the high-risk range of the NRTs is very low; hence, teachers or psychologists could place strong credibility in the results of the low-risk group as being truly low risk. However, differences in negative predictive power estimates were found across the three types of performance indicators. Specifically, dual discrepancy had greater negative predictive power than level and slope, and level had greater negative predictive power than slope. No consistent differences across type of MCBM probes were observed.

In comparison to negative predicative power, positive predictive power estimates (i.e., students identified as high risk on MCBM probes and NRTs) were much weaker. Thus, the credibility of students identified as high risk on the MCBM probes is rather limited. However, because the intent of the MCBM probes was not to serve as a diagnostic indicator, but rather as a screening device, the low positive predictive power is more tolerable (although far from ideal). Further, evidence exists supporting a high specificity and high negative prediction values in spite of relatively lower sensitivity and positive predictive values when the purpose of the measure in question is solely intended as a screening device rather than a diagnostic instrument (Gray, Tonge, Sweeney, & Einfeld, 2008; McFarlane, McKenzie, Van Hoof, & Browne, 2008; O'Donnell et al., 2008; Roberts, Stuart, & Lam, 2008; Shean & Baldwin, 2008).

Differences in PPVs were found across the three types of performance indicators; these differences also varied from differences found among negative predictive estimates. Specifically, level had greater positive predictive power than both slope and dual discrepancy; additionally, positive predictive power estimates for the slope performance indicator were much weaker than those for the dual discrepancy strategy. A few consistent differences across type of MCBM probes were also observed. For level, positive predictive estimates for error analysis and mastered

single skill probes were slightly greater compared to single skill and multiple skill probes across time and NRTs. For slope, the SSL probe was consistently greater than other probes on all NRTs at Time 4.

Given the results of the statistical analyses, lower positive predictive power may be due to a higher number of high risk students identified by the MCBM probes than were identified by the NRTs (overidentification). As a result, more referrals for further intervention or evaluation of student performance than would be necessary (i.e., students identified as high risk on MCBM probes who would not likely qualify as high risk on NRTs) would be generated using only the MCBM procedures. However, it would be reasonably accurate to presume that students identified as low risk by the MCBM probes would truly be identified as low risk by the NRTs due to the strength of specificity identified between MCBM probes and NRTs. Thus, students not referred for further evaluation would likely not require further assessment. The implications of this finding provide rationale to use the MCBM probes as a screening instrument to rule out students who likely would not benefit from additional assessment. Given the relatively small numbers of students identified as high risk, this would greatly reduce the overall number of students to be considered for further evaluation of risk status using more timely and expensive methods (Fletcher et al., 2005; Stecker & Fuchs, 2000).

The low sensitivity estimates across all probes is a great concern, however, given that this finding indicates that a substantial portion of students that are identified as high risk on the NRTs were not identified as high risk by MCBM probes. As a result, some students who need help would not get it and may continue to experience academic difficulties in math. Greater sensitivity estimates were obtained between pairwise comparisons of NRTs than were obtained between individual MCBM and individual NRT scores. To obtain higher sensitivity–and, thereby, to decrease the number of false negatives–a less conservative cut-off point may be needed than the one used in this study. The disadvantage of doing so, however, would be that the number of false

positives would also increase. This poses a real concern to school personnel because of the high cost and resources needed to provide additional intervention or assessment to students who do not really need additional help (Stecker & Fuchs, 2000). Overall, the classification accuracy of low-risk students generated by the MCBM probes was strong relative to all of the NRTs; this strength was also observed across MCBM probe types and administration times. However, the results of the study indicate that the MCBM probes are not very accurate in identifying the same students in the high-risk category as were identified by the NRTs. Given the relationship of balance between sensitivity and specificity, it is clear that the MCBM probes perform fairly poorly at identifying high- risk students. However, it is important to appreciate that the construct measured in the study was limited to math calculation skills. In order to increase the construct validity of the comparisons between the MCBM probes and the subtests from the NRTs included in the study, only math calculation subtests from each of the NRTs were administered to the participants. As a result, only one subtest cluster score was derivable from only one NRT battery (the Mathematics Calculation Cluster score on the WJ-ACH-III). Thus, all other analyses compared student performance on the NRTs to student performance on individual subtests of the KM 3 and the WIAT. The psychometric properties of the NRT subtests are weaker than those of the cluster scores that are derived when mathematics reasoning skills are also administered. Further, these subtests are rarely used in isolation to make educational placement decisions and, therefore, do not adequately represent a comprehensive skill evaluation that may be used for this purpose. However, as previously described, the purpose of the study was to determine the ability of mathematics calculation CBM probes to accurately identify students struggling to master math calculation skills only (as opposed to specifically finding students likely to fail more global measures of academic performance).

To further determine how well MCBM measures predict outcomes on standardized and other important educational measures, prior studies have found more encouraging results for the

use of MCBM computation as a moderate predictor of outcomes on state assessments when using ROC curves to establish cut points. For example, Shapiro and colleagues (2006) found that both specificity and sensitivity were found to be above .6 when comparing the predictive validity of low achievers identified using scores from multiple math probes with low achievers based on a statewide curriculum-based test and the SAT-9 scores. These findings are much more acceptable for screening purposes than the findings in this study. However, it is also critical to note that the rate of agreement between pairwise comparisons of the NRTs indicated that although student performance on the NRTs was utilized primarily as a "gold standard" against which to compare student performance on the MCBM probes, disagreement between risk status of individual students was also present–particularly for comparisons involving the KM 3. Thus, while inclusion of nationally norm-referenced standardized academic achievement tests is likely the best comparison against which to judge the current academic ability of individual students, it must also be appreciated that some disagreement between these NRTs exists. This also provides an appreciation for the limitations and challenges inherent in attempting to identify the same kids as high risk using MCBM probes.

Finally, a series of logistic regressions were used to further examine the degree to which one or more MCBM screening measures predicted student risk status based on individually administered standardized achievement test scores. Comparisons of the predictor variables (SSL, SSM, MS, and EA) and each of the NRT scores indicated that only one combination of MCBM scores was a significant predictor of performance on any NRT in any type of screening measure (slope, dual discrepancy, or level). Specifically SSL and MS improved prediction on the KM 3 Add/Sub assessment for the dual discrepancy performance indicator as compared to either alone. Otherwise only one MCBM served as a significant predictor of risk status on any NRT score. Unfortunately, no clear pattern of

results was observed to indicate which MCBM was the most consistent predictor across all NRTs. In addition, it would appear that time of administration tends to have an unknown impact on this outcome. For example, regarding the level performance indicator, one MCBM and/or performance indicator was identified as a significant predictor for students' WIAT scores each time, but the MCBM probes identified differed. A MCBM probe predicted risk status on the WJ-ACH-III and KM 3 Add/Sub on three of the four MCBM administration times, but no MCBM probe significantly predicted student performance on the KM 3 Mult/Div score. For the dual discrepancy performance indicator, although more data points were used to calculate slope in Time 4, MCBM probes (specifically, SSL and MS) significantly predicted student risk status on only one NRT score (KM 3 Add/Sub). Time 3 fared better, but again, no MCBM probe was consistent across all NRT tests (SSL on the WJ-ACH-III and WIAT, but EA on the KM 3). Finally, for the slope performance indicator, no predictor variable was significantly different than the model with only a NRT constant alone with the exception of one case (SSL predicted risk status on the WJ-ACH-III). These results indicate that two of the three MCBM predictors (MS and EA) predicted the outcome (high risk on an NRT test) better than no predictor variables for the slope performance indicator. However, a range of 81.9-95.8% of all cases was correctly predicted across all comparison models examined in this study. Moreover, all models were much better at predicting students who were not in the high-risk category than at predicting students who are in the high-risk category. In general, the error analysis probes (identified as a significant predictor for seven comparisons) were most commonly identified as having the greatest explanatory power in the overall models; the SSL and MS probes were each identified as a significant predictor in four

comparisons. However, it is important to appreciate the inherent limitations of using so many variables that are so highly intercorrelated.

Implications for Future Research

In sum, when comparing the ability of the MCBM probes to match the risk status of each individual child as determined by each of the NRTs, a pattern emerged in which the MCBM probes could be used to accurately discriminate the students in the low-risk category, but not those in the high-risk category across all probe types and administration times. In fact, little differences between probes, administration times, or combination of probes were noted when considering all students together–regardless of which MCBM performance indicator was employed (although the level method generated the greatest support and the slope method generated the least support).

Given that math has poorer validity than reading (Shinn, 1989, 1998), it was not surprising that the results of this study were not as strong as previous validity studies on reading; however, these results were lower than was expected at the outset of the study. Several factors may have influenced these results that require additional research. First, the results of this study may differ from those of prior studies because the current study employed a 9-week math curriculum on probes rather than using a semester or academic year curriculum (Good & Jefferson, 1998; Shapiro et al., 2006). Because the MCBM probes utilized in this study represented a relatively shorter breadth of curriculum, it is possible that the MCBM probes did not measure precisely the same thing as the NRTs did. Specifically, the breadth of curriculum assessed on the NRTs was simply not adequately reflected in the MCBM probes. Second, earlier research in reading has

suggested that using any less than three MCBM probes in identifying students at high risk versus low risk results in unacceptable error rates (Hintze, Owen, Shapiro, & Daly, 2000). Third, some research has indicated that multiple skill math MCBM probes likely represent a different construct of math achievement than single skill math MCBM probes (Hintze et al., 2002). Specifically, it has been theorized that single skill MCBM probes

measure solely an underlying premise for any particular skill, while multiple skill probes represent a more global outcome measure (Hintze et al, 2002).

Finally, a review of individual status for each student who had been identified as high risk on any of the tests used in this study indicated a few interesting observations. First, when comparing multiple MCBM probes to the four NRT measures used in the study, it is important to note that the majority of students identified as high risk were identified as such by only one NRT. Of the 31 students identified by the NRTs as high risk, only 4 students were identified by all four NRTs as high risk and only 7 students were identified as high risk by three or more NRTs. This low agreement in status rating implies that the NRTs used may also not consistently identify the same child; as a result, although the NRTs were included in the study in an effort to provide a "gold standard" against which the results of the MCBM probes could be compared, the standard set by the NRTs was not as firm as was hypothesized at the outset of the study. Thus, an alternative method for identifying students who are truly at high risk for failure on important educational measures may need to be further investigated.

In general, the majority of students' performance on all MCBM probes improved over 9 weeks (with only one exception: sixth-grade student performance increased from Time 1 to Time 3 on the SSL probe, but decreased on Time 4). Thus, it was expected that level scores would be more strongly correlated with NRT scores and, thereby, more accurately identify student risk status on NRT scores at later times when students have been taught skills; it was further anticipated that incorporating the slope performance indicator would improve agreement on risk status between MCBM and NRT scores. Interestingly, however, in this study, students obtained a higher slope than on a probe

covering grade level material over the course of a school year or semester (Fuchs et al., 2007). In fact, in a previous study, researchers found that average student growth rate for first-grade students was .37 DCPM per calendar day throughout the school year on weekly grade-level curriculum probes. These results may suggest that assessing performance covering a small amount of material that is covered within a short time period may potentially be more sensitive to detecting poor or adequate response to intervention within a few weeks and may also have utility as an indicator of instructional modifications. However, a survey of growth on skills taught over a 9-week period as compared to the school year may result in poor results for predicting performance on NRTs due to limited breadth of content or poor content overlap with the NRT measures. Additional research is needed to determine if the type of survey level MCBM employed in the current study may be useful for identifying a student's individual level on a small group of skills. Additionally, it may be possible to compare slope scores of students identified as requiring specialized instruction with those of average students to determine adequate slope scores to be used for shorter term responsiveness to intervention (RTI) decisions.

One interesting finding was that a high percentage of students in Phase I performed below instructional level on the teacher-selected mastered skill MCBM probe. Given that the teacher selected a skill that had been recently taught and was estimated by the teacher to be well taught, it was expected that approximately 10% of the students would perform below the instructional level. However, percentages ranged between 47-61% across all students. Similar to research indicating poor relationships between teacher rating or referral and high-risk students detected using standardized testing

(VanDerHeyden, Witt, Naquin, & Noell, 2001; Mazzocco & Myers, 2003), teachers may need data to refine their judgment of student performance. Additional research may be needed to determine if this type of assessment may be useful to instructional planning purposes or for identifying classes that would benefit from classwide interventions (VanDerHeyden et al., 2007).

Several additional implications for future research can be generated from the current study. First, a larger sample size with a greater number of students identified as high risk would facilitate greater statistical power. Additionally, this study contained a low representation of students from diverse ethnic and SES groups. Finally, implementing follow-up procedures to identify whether or not the number of students originally identified by the MCBM probes could be reduced to a number more similar to that identified by the NRTs may delineate the utility of the MCBM probes as part of a more comprehensive or sophisticated system to identify high-risk students. One potential method would be to conduct a motivational analysis to determine if MCBM scores improved if students were more motivated to improve their performance.

Limitations of the Study

Several important limitations of the present study may have impacted the results generated. First, the sample of high-risk students identified for inclusion in the present study was smaller than anticipated at the outset of the study. There are several important considerations impacting the small sample size. First, students were selected for inclusion based upon receiving informed consent from parents/guardians and a complete data set (i.e., each student was present for all MCBM probe administrations). The return

of informed consent rate was 32%. However, only 65% of students for whom informed consent was obtained had a complete data set. The reason for this attrition was due to constraints imposed upon the researchers from the institutional review board (IRB) of the local school district in which the research was conducted. Specifically, the district's IRB mandated that all students participating in the research program could be removed on only one occasion throughout the study; as a result, students were removed for administration of the NRTs only. Thus, if students were absent during the classwide MCBM probe administrations, the researchers were not permitted to remove students for "make-up" MCBM administrations.

A second, and related, limitation is that the sample of fourth-grade phase II participants did not contain any students identified as high risk by the WIAT or the WJ-ACH-III. Finally, the participants represented a limited number of grade levels with few ethnically diverse and low SES students. Thus, external validity is somewhat threatened because of the extent that the current findings generalize to other school populations is unknown.

A third limitation is the psychometric limitations of the subtests administered from the NRTs. In this study, only subtests consisting primarily of calculation problems were administered rather than a sufficient number of subtests (including those with additional math skills–e.g., word problems, application skills) to create a cluster or battery composite score. This is problematic because students' educational status is more often determined using composite scores rather than subtest or cluster scores. Therefore, the comparisons in this study should be somewhat cautiously interpreted. Moreover, the effects of different MCBM measures on different tests are hard to determine based on

these results. This study examined MCBM data obtained on biweekly probes that assessed performance on curriculum to be covered within a 9-week quarter term, but previous research used content taught throughout the academic year. This is both a limitation and a need for additional comparisons to identify accurate screeners for low achievers in math.

Finally, it is important to note that all performance indicator cut-point scores were locally derived and calculated using a local norm base. Thus, it is important to appreciate that observed differences in risk status on the level and slope performance indicators are the direct result of the cut-points established using the data obtained from the sample within the study and were selected based upon the findings of the literature review (Burns & Senesac, 2005; VanDerHayden, Witt, & Gilbertson, 2007). Thus, using a different set of cut-points for inclusion in the high- and low-risk categories on the performance indicators may have significantly impacted the outcomes of the analyses included in the study. Hence, using ROC curves to derive specific cut-points based upon local normative data may prove to be a more effective method of identifying students in the high- and low-risk range.

Conclusions

Overall, the results of the study indicate that the MCBM probes could be accurately and reliably used to identify students in the low-risk category for failure on important educational measures for second-, fourth-, and sixth-grade students. However, the MCBM probes' ability to successfully identify students in the high-risk category was substantially lower. Further, it would appear that administration time had little impact on

the predictive power of the MCBM probe to identify students as high or low risk. In general, the types of MCBM screening measures used in this study were more problematic for detecting students who are having math difficulties based on standardized tests (sensitivity) than for detecting students who are adequately performing math computation skills on standardized tests (specificity).

Given that the MCBM probes were designed only as a screening strategy of identification and not as a diagnostic instrument, it is important to note that the MCBM probes overidentified students as high risk rather than underidentified them. Thus, it is more acceptable and carries greater clinical utility to have too many students identified as high risk rather than too few students because those students identified as high risk will subsequently be subjected to additional services or diagnostic measures. If, however, the MCBM probes fail to identify students who are truly high risk, those who are erroneously identified as low risk (i.e., those who truly may benefit from further assessment and/or intervention) will not receive it. This poses a much greater threat to the clinical utility of any screening instrument because it undermines its power to adequately perform the task for which it was designed. Although the MCBM probes identified more students than the NRTs did, the MCBM probes generally identified only slightly more students as high risk than the NRTs. However, given the low overall proportion of students identified as high risk, the results of the specificity deteriorate with even small differences.

REFERENCES

Augustyniak, K., Murphy, J., & Phillips, D. K. (2005). Psychological perspectives in assessing mathematics learning needs. *Journal of Instructional Psychology, 32*(4), 277-286.

Badian, N. (1999). Persistent arithmetic, reading or arithmetic, or reading disability. *Annals of Dyslexia, 49*, 45-70.

Burns, M. K., & Senesac, B. V. (2005). Comparison of dual discrepancy criteria to assess response to intervention. *Journal of School Psychology, 43*, 393-406.

Connolly, A. J. (1997). *Key math--Revised, normative update*. Circle Pines, MN: American Guidance Services.

Daly, E. J., & McCurdy, M. (2002). Getting it right so they can get it right: An overview of the special series. *School Psychology Review, 31,* 453-458.

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

Deno, S. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*(3), 184-192.

Deno, S., Fuchs, L. S., Marston, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30*, 507-524.

Deno, S., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*, 36-45.

Fletcher, J. M., Denton, C., & Francis, D. J. (2005). Validity of alternative approaches for the identification of learning disabilities: Operationalizing unexpected underachievement. *Journal of Learning Disabilities, 38*, 545-552.

Foegen, A., & Deno, S. L. (2001). Identifying growth indicators for low-achieving students in middle school mathematics. *The Journal of Special Education, 35*(1), 4-16.

Fuchs, L. S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities Research & Practice, 18*(3), 172-186.

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

Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disability. *Learning Disability Research and Practice, 13,* 204-219.

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

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989a). Monitoring reading growth using student recalls: Effects of two teacher feedback systems. *Journal of Educational Research, 83*(2), 103-110.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989b). Effects of instrumental use of curriculum-based measurement to enhance instructional programs. *RASE: Remedial & Special Education*, 10(2), 43-52.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989c). Effects of alternative goal structures within curriculum-based measurement. *Exceptional Children, 55*(5), 429-438.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Allinder, R. M. (1991). Effects of expert system advice with curriculum-based measurement on teacher planning and student achievement in spelling. *School Psychology Review, 20*(1), 49-66.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement, using a reading maze task. *Exceptional Children, 58*(5), 436-450.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1990). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review, 19*, 6-22.

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

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

Good, R. H., & Shinn, M. R. (1990). Forecasting accuracy of slope estimates for reading curriculum-based measurement: Empirical evidence. *Behavioral Assessment, 12*(2), 179-193.

Gray, K. M., Tonge, B. J., Sweeney, D. J., & Einfeld, S. L. (2008). Screening for autism in young children with developmental delay: An evaluation of the Developmental Behaviour Checklist, Early screen. *Journal of Autism and Developmental Disorders, 38*(6), 1003-1010.

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

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

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

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

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*(3), 358-371.

Klingner, J. K., Artiles, A. J., & Barletta, L. M. (2006). English language learners who struggle with reading: Language acquisition or LD? *Journal of Learning Disabilities, 39*(2), 108-128.

Magyar, C. I., Pandolfi, V., & Peterson, C. R. (2007). In J. W. Jacobson, J. A. Mulick, & J. Rojahn (Eds.), *Handbook of intellectual and developmental disabilities* (pp. 333-351). New York: Springer.

Manzo, K. K., & Galley, M. (2003). Math climbs, reading flat on ’03 NAEP. *Education Week, 23*(12), 1.

Marston, D. B. (1989). Curriculum-based measurement: What it is and why we do it? In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford.

Mazzocco, M. M. M. (2005). Challenges for identifying target skills for math disability screening and intervention. *Journal of Learning Disabilities, 38*(4), 318-323.

Mazzocco, M. M. M., & Myers, G. F. (2003). Complexities in identifying and defining mathematics learning disability in the primary school-age years. *Annals of Dyslexia, 53*, 218-253.

McFarlane, A. C., McKenzie, D. P., Van Hoof, M., & Browne, D. (2008). Somatic and psychological dimensions of screening for psychiatric morbidity: A community validation of the SPHERE questionnaire. *Journal of Psychosomatic Research, 65*(4), 337-345.

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

McMaster, K. L., Fuchs, D., Fuchs, L. S., & Compton, D. L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional Children, 71*(4), 445-463.

O'Donnell, M. L., Creamer, M. C., Parslow, R., Elliott, P., Holmes, A. C. N., Ellen, S., et al. (2008). A predictive screening index for posttraumatic stress disorder and depression following traumatic injury. *Journal of Consulting and Clinical Psychology, 76*(6), 923-932.

Roberts, N., Stuart, H., & Lam, M. (2008). High school mental health survey: Assessment of a mental health screen. *The Canadian Journal of Psychiatry/La Revue Canadienne de Psychiatrie, 53*(5), 314-322.

Saffer, N. (1999). Core subjects and your career. *Occupational Outlook Quarterly, 43*(2), 26-40.

Sattler, J. M. (1988). *Assessment of children* (3rd ed.). San Diego, CA: Author.

Shalev, R. S. (2004). Developmental dyscalculia. *Journal of Child Neurology, 19*(10), 765-771.

Shapiro, E. S. (1996). *Academic skills problems: Direct assessment and intervention* (2nd ed.). New York: Guilford.

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

Shean, G., & Baldwin, G. (2008). Sensitivity and specificity of depression questionnaires in a college-age sample. *Journal of Genetic Psychology, 169*(3), 281-288.

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

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

Shinn, M. R., & Marston, D. (1985). Differentiating mildly handicapped, low-achieving, and regular education students: A curriculum-based approach. *RASE: Remedial & Special Education, 6*(2), 31-38.

Speece, D. L., & Case, L. P. (2001). Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology, 93*(4), 735-749.

Stecker, P. M., & Fuchs, L. S. (2000). Effective superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research & Practice, 15,* 128-134.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42*(8), 795-819.

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*(4), 493-513.

Tindal, G. A., & Marston, D. B. (1990). *Classroom-based assessment: Evaluating instructional outcomes*. Columbus, OH: Merrill.

VanDerHeyden, A. M., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. *School Psychology Review, 30*(3), 363-382.

VanDerHeyden, A. M., Witt, J. C., & Gilbertson, D. (2007). A multi-year evaluation of the effects of a response to intervention (RTI) model on identification of children for special education. *Journal of School Psychology, 45*(2), 225-256.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children--Revised.* New York: The Psychological Corporation.

Wechsler, D. (2002). *Examiner's manual: Wechsler Individual Achievement Test* (2nd ed.). New York: Pearson.

Wesson, C., Deno, S., & Mirkin, P. (1988). A causal analysis of the relationships among ongoing curriculum-based measurement and evaluation, the structure of instruction, and student achievement. *Journal of Special Education, 22*(3), 330-343.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psychoeducational Battery—Revised.* Chicago: Riverside.

Yell, M. L., & Drasgow, E. (2007). Assessment for eligibility under IDEIA and the 2006 regulatins. *Assessment for Effective Intervention, 32*(4), 202-213.

APPENDICES

Appendix A:

Informed Consent

Date Created: Nov. 12, 2007; Page 1 of 2
USU IRB Approved: 11/12/2007
Approval terminates: 11/11/2008
Protocol Number 1894
IRB Password Protected per IRB Administrator

INFORMED CONSENT

Tips on Interventions for Parents and Students

Introduction

We would like your permission to include your child in a new study with the Utah State University (USU) Psychology Department that will help to find simple and quick ways to keep track of students' progress in the math program at your child's school. This will help the teachers determine if any children need extra help as they continue to learn new math skills. Your child would be working with Kyle Max Hancock, a doctoral student and certified school psychologist, under the supervision of Dr. Donna Gilbertson. If you agree to allow your child to participate, your child will be asked to complete several brief math tests. Your child will miss about 30 minutes of class time to take the tests, but we will work with teachers to make sure that your child misses the least amount of class possible. We would also like you to tell us a little bit about your child by filling out a brief survey (attached). There will be about 150 to 200 students involved in this research.

Procedures

If you give your permission for your child to participate in this study, your child will complete three short math tests. These math tests consist of grade-level math curriculum (e.g., addition, subtraction, multiplication, and/or division) and will be given to each participant individually by an adult. Your child will simply be asked to work on these math problems and to do his/her best work. None of the math tests will be counted toward your child's grade; they will be used to better understand how we may help all of the students learn better.

Risks

There is minimal risk associated with the programs being used in this study or the surveys we are asking you and your child to complete. It is possible that participating in this research could cause you or child some discomfort due to responding to the questionnaires or participating in the math tests. To avoid stress, you can skip questions that you do not want to answer, but it will help the researchers if most questions can be answered.

New Findings

During the course of this research study, you will be informed of any significant new findings (either good or bad), such as changes in the risks or benefits resulting from participation in the research, or new alternatives to participation that might cause you to change your mind about continuing in the study. If new information is obtained that is relevant or useful to you, or if the procedures and/or methods change at any time throughout this study, your consent to continue participating in this study will be obtained again.

Benefits
We hope to be able to obtain important information about how to best help children who are struggling to learn new math skills. We also hope to provide your child's school with a better and easier way to determine which kids would benefit from extra help or support in math.

Explanation & offer to answer questions
We have explained this research study to you and answered your questions. If you have other questions or research-related problems, you may reach Dr. Gilbertson at (435) 797-2034 or Kyle Hancock at (801) 402-2208.

Voluntary Nature of Participation and Right to Withdraw without Consequence
Participation in this research is entirely voluntary. You and your child may refuse to participate or withdraw from the study at any time without consequence.

Confidentiality
Information about you and your child will be kept confidential and will be available only to the researchers directly involved in the project. Your child will be assigned a code number and only this number will be used when the data is stored in the computer to protect privacy. Public presentations of the results of this study will in no way identify you or your child. All data will be kept in a locked filing cabinet which will be accessible only by Mr. Hancock and Dr. Gilbertson in a locked office at USU.

IRB Approval Statement
The Institutional Review Board (IRB) for the protection of human participants at USU has approved this research; you may contact them at (435) 797-1821 if you have more questions. The Davis School District Research and Assessment Department has also approved this study; you can contact them at (801) 402-5305.

Copy of Consent
This package contains two copies of this Informed Consent Form. Please sign both and keep one copy. Please return one signed copy with the survey you fill out.

Investigator Statement: "I certify that the research study has been explained to the individual, by me or my research staff, and that the individual understands the nature and purpose, the possible risks and benefits associated with taking part in this research study. Any questions that have been raised have been answered."

Donna M. Gilbertson, Ph.D.	Kyle Max Hancock, M.S.
Principal Investigator	Graduate Researcher
(435) 797-2034	(435) 755-3402

Signature of Parent / Guardian **(please check one and sign if agreeing to participate)**

____Yes, I am willing to have my child participate in this study.

____ NO, I do NOT want to participate in this study and I do not want my child to participate

Paren/Legal Guardian's signature	Date

Child/Youth Assent: I understand that my parent(s)/guardian is/are aware of this research study and that permission has been given for me to participate. I understand that it is up to me to participate even if my parents say yes. If I do not want to be in this study, I do not have to and no one will be upset if I don't want to participate or if I change my mind later and want to stop. I can ask any questions that I have about this study now or later. By signing below, I agree to participate.

_____ _____

Name Date

Appendix B:

Letter from Principal to Parents

Tips on Interventions for Parents and Students

Dear Parent:

I am writing this letter to all parents to encourage them to provide permission for their children to participate in a project being conducted at (NAME OF SCHOOL). As you are aware, our school is committed to providing a setting that is productive, safe, and fosters excellence in academic achievement. This project is meant to help provide feedback to teachers and students regarding the progress that students are making in their mathematics skills.

As a first step in this study, Dr. Donna Gilbertson and Kyle Max Hancock from Utah State University have worked with the school to collect information on the current mathematics achievement level of our students. To achieve this goal, the researchers have trained and assisted teachers to administer several two-minute timed math quizzes in their classrooms. Following these short quizzes, approximately 50 students from each grade are needed to participate in the second step of the study. The researchers are sending all parents this letter asking for parents to give permission for their child to participate in the study. All children who return this packet to their teacher will receive a small reward (e.g., pencil or piece of candy) for doing so *regardless of whether or not consent is provided.* Parents are completely free in the choice they make to provide or deny permission.

The second step of the study will involve administering three other math tests to those students whose parents provided permission to participate in the study. The information obtained from these assessments will not be associated with your child's identifying information and will not influence his/her grade in any way. The information will be used to determine the extent to which the two-minute math quizzes previously administered helps to identify those students who are most likely to benefit from additional mathematics intervention.

Please read the attached Informed Consent Form for more details on this project. If you have more questions on the study you can call the researchers (their numbers are on the bottom of the consent form). Also, feel free to give us a call if you have any questions or concerns.

Thank you for your support of this project.

Sincerely,

Principal, SCHOOL

Appendix C:

Phase II Student Demographics Data Sheet

Parent Packet
Student Demographics Sheet

Parent Information

1) Your gender (Check one): _____ male _____ female

2) Relationship to child (Check one):

____ biological parent ____ adoptive parent ____ legal guardian ____ other _____

3) Highest level of education completed (Check one):

_____ did not complete high school
_____ completed high school
_____ completed some college
_____ completed college degree
_____ completed graduate/postgraduate education

4) Your native language (Check one):

[ ] English [ ] Spanish [ ] other _____

Child Information

1) Child's age: _____ Birth date: ___/___/___

2) Child's grade level: ____

3) Child's gender (Check one): _____ male _____ female

4) Child's ethnicity (Check one):

_____ Latino/a _____ African American _____ Caucasian

_____ Asian _____ Native American _____ Other _____

5) Child's native language: [ ] English [ ] Spanish [ ] other _____

6) Has your child received English as a second language services?

[ ] None [ ] Receiving classes now [ ] Not currently but has in the past

6) Has your child been retained or attended a grade more than one year?

[ ] Yes [ ] No If yes, what grade did your child attend for a second year? _____

8) Has your child ever been diagnosed with any psychological and/or behavioral disorders?
_____ no _____ yes (Please specify which ones :_____)

Appendix D:

Mastered Single Skill Probe (SSM)

Mastered Single Skill Probe

1) 3 - 0	2) 6 - 5	3) 8 - 5	4) 3 - 3	5) 4 - 2
6) 9 - 3	7) 6 - 1	8) 5 - 4	9) 8 - 1	10) 5 - 2
11) 7 - 7	12) 3 - 2	13) 7 - 1	14) 4 - 1	15) 7 - 3
16) 9 - 3	17) 4 - 4	18) 8 - 0	19) 8 - 2	20) 9 - 9
21) 6 - 0	22) 9 - 4	23) 2 - 0	24) 9 - 6	

Appendix E:

Single Skill to be Learned Probe (SSL)

Single Skill to be Learned Probe

1) 3 + 0	2) 6 + 5	3) 8 + 5	4) 3 + 3	5) 4 + 2
6) 9 + 3	7) 6 + 1	8) 5 + 4	9) 8 + 1	10) 5 + 2
11) 7 + 7	12) 3 + 2	13) 7 + 1	14) 4 + 1	15) 7 + 3
16) 9 + 3	17) 4 + 4	18) 8 + 0	19) 8 + 2	20) 9 + 9
21) 6 + 0	22) 9 + 4	23) 2 + 0	24) 9 + 6	

Appendix F:

Multiple Skill Probe (MS)

Mastered Skill Probe

1) 347 +123	2) 215 + 495	3) 54 - 12	4) 367 - 121	5) 51 - 22
6) 40 - 11	7) 651 - 215	8) 932 -167	9) 301 - 177	10) 3 4 x 2
11) 12 x 24	12) 67 x 2	13) 503 x 5	14) 234 x 5	15) 20 x 52
16) 36 x 12	17) 37 x 25	18) 2 / 46	19) 5 / 155	20) 2 / 89
21) 8 / 204	22) 11/ 253	23) 3.4 + 5.2	24) 2.4 x 1.5	

Appendix G:

Error Analysis Probe (EA)

Error Analysis Probe

1) 347 +123	2) 215 + 495	3) 54 - 12	4) 367 - 121	5) 51 - 22
6) 40 - 11	7) 651 - 215	8) 932 -167	9) 301 - 177	10) 3 4 x 2
11) 12 x 24	12) 67 x 2	13) 503 x 5	14) 234 x 5	15) 20 x 52
16) 36 x 12	17) 37 x 25	18) 2 / 46	19) 5 / 155	20) 2 / 89
21) 8 / 204	22) 11/ 253	23) 3.4 + 5.2	24) 2.4 x 1.5	

Appendix H:

Teacher Problem Selection Sheet

Teacher Problem Selection Sheet

Teacher: _____ Grade level: ____

Below is a list of basic skills and mastered in first through sixth grade. Please check off appropriate grade level skills that have been taught this year to all students in your grade level. Only add a check if the skill is expected to have been mastered by students who are progressing as expected within your curriculum.

Addition	
	Two 1-digit numbers: sums to 5
	Two digit numbers: sums to 10
	Two 1-digit numbers: sums to 18
	1- to 2-digit number plus 1- to 2-digit number: no regrouping
	Two 2-digit numbers: no regrouping
	Two 3-digit numbers: no regrouping
	Two 2-digit numbers: regrouping
	2- to 3-digit number plus 2- to 3-digit number: regrouping from 1's & 10's columns
	3-digit number plus 3-digit number: regrouping from 1's & 10's columns
Subtraction	
	Two 1-digit numbers: 0 to 5
	Two 1-digit numbers: 0 to 9
	2-digit number from a 2-digit number: no regrouping
	3 digit number from a 3-digit number: no regrouping
	2-digit number from a 2-digit number: regrouping
	3-digit number from a 3-digit number: regrouping from 1's & 10's columns
Multiplication	
	Multiplication facts: 0 to 9
	Multiplication facts: 0 to 5
	2-digit number times 1-digit number: no regrouping
	2-digit number times 1-digit number: regrouping
	3-digit number times 1-digit number: no regrouping
	3-digit number times 1-digit number: regrouping
	2-digit number times 2-digit number: no regrouping
	2-digit number times 2-digit number: regrouping
	3-digit number times 2-digit number: no regrouping
	3-digit number times 2-digit number: regrouping
	3-digit number times 3-digit number: no regrouping
	3-digit number times 3-digit number: regrouping

Division	
	Division facts: 0 to 9
	2-digit number divided by 1-digit number: no remainder
	2-digit number divided by 1-digit number: remainder
	3-digit number divided by 1-digit number: no remainder
	3-digit number divided by 1-digit number: remainder
	3-digit number divided by 2-digit number: no remainder
	3-digit number divided by 2-digit number: remainder
Fifth and Sixth:	
	Fractions addition
	Fractions subtraction
	Decimals addition
	Decimals subtraction
	Decimals multiplication
	Decimals division

Appendix I:

Administrator Scripted Instructions

for 2-Minute Math Probe

Administrator Scripted Instructions for 2-Minute Math Probe

1. Write the teacher's name on the board, the date and NAME under it

2. Say **"Take out a sharpened pencil. I will be passing out a piece of paper face own. Please do not turn the paper face up until you are instructed to do so."**

3. Pass out papers face-down.

4. **"Please write your teacher's name, date and your first and last name on the back of your paper."** Pause briefly to allow students to write their names.

5. Say "**This is a math worksheet. The worksheet contains different kinds of the problems such as addition, subtraction, and multiplication**. **When I say 'start,' turn them over and begin answering the problems. Start on the first problem on the left on the top row** (point). **Work across and then go to the next row. Raise your hand if you have a question. "**

6. Set timer for two minutes. **"Start."** Begin timer.

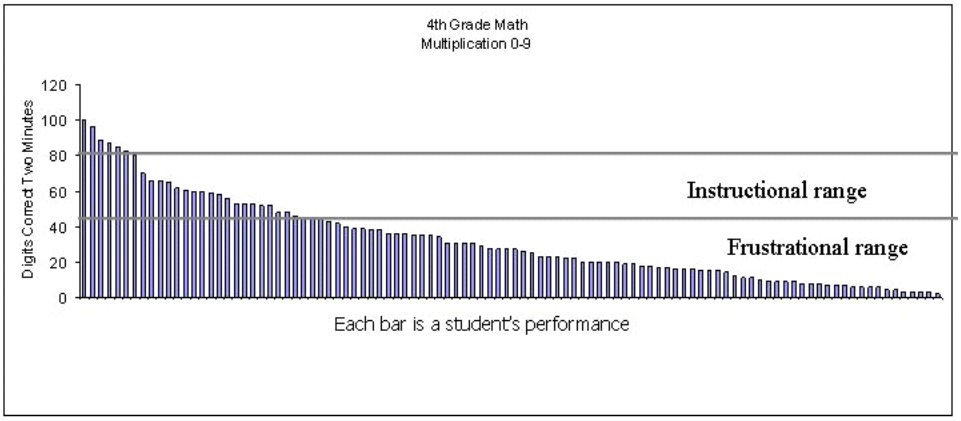
7. Monitor student performance to ensure that students work the problems in rows and do not skip around.

8. When timer rings, say, **"Stop. Raise your papers and put your pencils down."**

9. Collect math sheets and give to data entry person at your school.

Appendix J:

Example of Graph Given to Teachers for Feedback


4th Grade Math
Multiplication 0-9
Digits Correct Two Minutes
120
100
80
60
40
20
0
Instructional range
Frustrational range
Each bar is a student's performance

Appendix K:

Results of Preliminary Statistical Analyses

on Schools Included in the Study

Preliminary analyses between participants at each school were conducted using Chi-square tests to determine statistically significant differences between the three schools on gender, SES, ethnicity, and special education status. No significant differences in sex of students, χ^2 [2, N = 144] = 3.127, p = 0.209, or in number of student receiving special education services, χ^2 [2, N = 144] = 1.612, p = 0.447, were identified among the three schools. However, there was a significant difference in SES of participants among the three schools: χ^2 (2, N = 144) = 32.897, $p < 0.001$. Follow-up analyses indicated that there was no difference between School 1 and School 2, χ^2 (1, N = 124) = 2.376, p = 0.123, but there was a significant difference between School 1 and School 3, χ^2 (1, N = 98) = 20.353, $p < 0.001$, and a significant difference between School 2 and School 3, χ^2 (1, N = 66) = 25.832, $p < 0.001$, with School 3 having the lower SES.

There was also a significant difference in ethnic diversity of participants among the three schools: χ^2 (10, N = 144) = 33.998, $p < 0.001$. Follow-up analyses indicated that there was no difference between School 1 and School 2, χ^2 (3, N = 124) = 4.735, p = 0.192, but there was a significant difference between School 1 and School 3, χ^2 (4, N = 98) = 17.981, $p < 0.01$ and a significant difference between School 2 and School 3, χ^2 (5, N = 66) = 15.994, $p < 0.01$. Overall, School 3 had more diverse students than both School 2 and School 1.

In sum, School 1 and School 2 were similar to each other regarding their demographics. School 3 was typically different; specifically, School 3 had a higher frequency of students in the high risk range of SES and a greater representation of ethnic diversity than both of the other schools.

In sum, School 1 and School 2 were similar to each other regarding their

demographics. School 3 was typically different; specifically, School 3 had a higher frequency of students in the high risk range of SES and a greater representation of ethnic diversity than both of the other schools.

Significant Differences between Schools on Initial Scores

A series of one-way analyses of variance (ANOVA) were conducted to evaluate the differences between phase II participant performance on the NRTs and MCBM probes. The independent variable, school attended, contained three levels: School 1, School 2, and School 3. The dependent variables were the DCP2M on the MCBM probes and the standard scores on the selected subtests of the WJ-ACH-III, WIAT, and KM 3.

The first set of ANOVAs was conducted with all participants. Table 27 presents the results of the ANOVAs. The ANOVA was significant for all of the tests except the Time 1 EA, and Time 2 SSL. The tests of homogeneity of variance were violated for the Time 1 EA (p = .034), Time 1 SSL (p = .010), Time 1 SSM (p = .003), Time 2 EA (p = .023), Time 2 SSL (p = .036), Time 3 SSL (p = .015), and Time 4 SSL (p = .008). Thus, the Dunnet's C post-hoc comparisons were used to identify significant differences between the groups. Table 28 presents the results of the post-hoc comparisons. In sum, mean scores for School 3 were significantly lower than mean scores for School 2 on all MCBM probes; further, mean scores for School 3 were also lower than School 1 on the Time 1 SSM and SSL, Time 3 EA, and Time 4 SSL probes. School 1 mean scores were

Table 27

Results of One-Way ANOVA on NRTs and MCBM Probes for All Participants

Source	*F*	*p*	η^2
NRTs			
WIAT	0.138	0.871	0.002
WJ-ACH-III	4.302	0.015	0.058
KM 3 Add/Sub	2.362	0.098	0.032
KM 3 Mult/Div	0.687	0.505	0.015
MCBM			
Time 1			
EA	2.562	0.081	0.035
MS	7.630	0.001	0.098
SSL	4.025	0.020	0.054
SSM	5.858	0.004	0.077
Time 2			
EA	3.272	0.041	0.044
MS	6.062	0.003	0.079
SSL	3.022	0.052	0.041
Time 3			
EA	4.780	0.010	0.063
MS	6.799	0.002	0.088
SSL	6.443	0.002	0.084
Time 4			
EA	4.675	0.011	0.062
MS	6.501	0.002	0.084
SSL	9.577	0.000	0.120

Table 28

Results of Post-hoc Comparisons for Significant

Results on ANOVA of All Participants

Source	School	*p*
WJ-ACH-III[a]		
	School 2 > School 1	0.011
	School 3 > School 1	0.704
	School 2 > School 3	0.405
SSM[b]		
	School 2 > School 1	>0.05
	School 1 > School 3	<0.05
	School 2 > School 3	<0.05
Time 1MS[a]		
	School 2 > School 1	0.003
	School 1 > School 3	0.603
	School 2 > School 3	0.004
Time 1SSL[b]		
	School 2 > School 1	ns
	School 1 > School 3	<0.05
	School 2 > School 3	<0.05
Time 2 EA[b]		
	School 2 > School 1	ns
	School 1 > School 3	ns
	School 2 > School 3	<0.05
Time 2 MS[a]		
	School 2 > School 1	0.016
	School 1 > School 3	0.439
	School 2 > School 3	0.007
Time 3 EA[a]		
	School 2 > School 1	0.442
	School 1 > School 3	0.047
	School 2 > School 3	0.007

(*table continues*)

Source	School	*p*
Time 3 MS[a]		
	School 2 > School 1	0.003
	School 1 > School 3	0.811
	School 2 > School 3	0.012
Time 3 SSL[b]		
	School 2 > School 1	ns
	School 1 > School 3	ns
	School 2 > School 3	<0.05
Time 4 EA[a]		
	School 2 > School 1	0.345
	School 1 > School 3	0.070
	School 2 > School 3	0.007
Time 4 MS[a]		
	School 2 > School 1	0.010
	School 1 > School 3	0.468
	School 2 > School 3	0.006
Time 4 SSL[b]		
	School 2 > School 1	<0.05
	School 1 > School 3	<0.05
	School 2 > School 3	<0.05

[a] Tukey's HSD
[b] Dunnett's C

significantly lower than School 2 on all four MS probe administrations and the Time 4 SSL probe administration.

Another series of ANOVAs was run with only the grade 2 participants included in the analyses. Table 29 presents the results of the ANOVA for the grade 2 participants. The tests of homogeneity of variance were not violated in all cases except for Time 4 EA ($p = .013$).

Table 29

Results of ANOVA for Phase II Participants and Measures by Grade Level

	Grade 2 (df = 2)			Grade 4 (df = 2)			Grade 6 (df = 2)		
Source	*F*	*p*	η^2	*F*	*p*	η^2	*F*	*p*	η^2
NRTs									
WIAT	3.022	0.058	0.116	0.968	0.387	0.041	1.478	0.239	0.063
WJ-ACH-III	0.094	0.910	0.004	2.405	0.102	0.097	3.776	0.031	0.147
KM 3 Add/Sub	1.465	0.242	0.060	0.854	0.433	0.037	3.326	0.045	0.131
KM 3 Mult/Div				0.266	0.768	0.012	0.560	0.575	0.025
MCBM									
SSM	2.754	0.074	0.107	7.352	0.002	0.246	3.756	0.031	0.146
Time 1 EA	7.586	0.001	0.248	8.081	0.001	0.264	2.049	0.141	0.085
Time 1 MS	0.685	0.509	0.029	18.537	0.000	0.452	4.145	0.022	0.159
Time 1 SSL	7.386	0.002	0.243	9.037	0.001	0.287	3.144	0.053	0.125
Time 2 EA	3.417	0.041	0.129	5.620	0.007	0.200	3.263	0.048	0.129
Time 2 MS	0.170	0.844	0.007	9.037	0.001	0.287	5.093	0.010	0.188
Time 2 SSL	0.593	0.557	0.025	6.887	0.002	0.234	0.508	0.605	0.023
Time 3 EA	0.804	0.454	0.034	3.034	0.058	0.119	3.874	0.028	0.150
Time 3 MS	0.920	0.406	0.038	10.879	0.000	0.326	3.438	0.041	0.135
Time 3 SSL	0.599	0.554	0.025	4.177	0.022	0.157	3.334	0.045	0.132
Time 4 EA	0.283	0.754	0.012	2.139	0.130	0.087	2.177	0.125	0.090
Time 4 MS	0.351	0.706	0.015	5.805	0.006	0.205	5.433	0.008	0.198
Time 4 SSL	0.295	0.746	0.013	6.235	0.004	0.217	5.217	0.009	0.192

No significant differences were found for any NRTs. For MCBM probes, the ANOVA was significant for the Time 1 EA, Time 1 SSL, and Time 2 EA probe. Follow-up tests for significant results were conducted using the Tukey's HSD test (see Table 30). In general, the results indicate that the means of the test scores were generally lower for School 1 than for School 2 and School 3 on the Time 1 EA and Time 1 SSL.

Table 30

Results of Tukey's HSD Follow-up Tests for Significant ANOVA

Source	Schools	*p*	Order
Grade 2			
Time 1 EA			
	School 1 > School 2	0.004	School 1 > School 3 > School 2
	School 3 > School 1	0.882	
	School 3 > School 2	0.003	
Time 1 SSL			
	School 1 > School 2	0.001	School 1 > School 3 > School 2
	School 1 > School 3	0.749	
	School 3 > School 2	0.022	
Time 2 EA			
	School 1 > School 2	0.077	School 1 > School 3 > School 2
	School 3 > School 2	0.058	
	School 1 > School 3	0.909	
Grade 4			
Time 1 EA [a]			
	School 2 > School 1	< 0.05	School 2 > School 1 > School 3
	School 3 > School 1	> 0.05	
	School 2 > School 3	> 0.05	
Time 1MS			
	School 2 > School 1	0.000	School 2 > School 3 > School 1
	School 1 > School 3	0.795	
	School 2 > School 3	0.001	
Time 1 SSL			
	School 2 > School 1	0.008	School 2 > School 1 > School 3
	School 1 > School 3	0.163	
	School 2 > School 3	0.001	
Time 1 SSM			
	School 2 > School 1	0.021	School 2 > School 1 > School 3
	School 1 > School 3	0.195	
	School 2 > School 3	0.004	
Time 2 EA			
	School 2 > School 1	0.023	School 2 > School 1 > School 3
	School 1 > School 3	0.533	
	School 2 > School 3	0.024	

(*table continues*)

Source	Schools	*p*	Order
Time 2 MS			
	School 2 > School 1	0.002	School 2 > School 1 > School 3
	School 1 > School 3	0.438	
	School 2 > School 3	0.004	
Time 2 SSL			
	School 2 > School 1	0.003	School 2 > School 1 > School 3
	School 1 > School 3	0.931	
	School 2 > School 3	0.046	
Time 3 EA			
	School 2 > School 1	0.049	School 2 > School 3 > School 3
	School 3 > School 1	0.956	
	School 2 > School 3	0.456	
Time 3 MS			
	School 2 > School 1	0.000	School 2 > School 1 > School 3
	School 1 > School 3	0.999	
	School 2 > School 3	0.022	
Time 3 SSL			
	School 2 > School 1	0.041	School 2 > School 1 > School 3
	School 1 > School 3	0.767	
	School 2 > School 3	0.080	
Time 4 MS			
	School 2 > School 1	0.013	School 2 > School 1 > School 3
	School 1 > School 3	0.696	
	School 2 > School 3	0.033	
Time 4 SSL			
	School 2 > School 1	0.006	School 2 > School 1 > School 3
	School 1 > School 3	0.931	
	School 2 > School 3	0.058	
Grade 6			
WJ-ACH-III			
	School 2 > School 1	0.024	School 2 > School 1
	School 3 > School 1	0.989	
	School 2 > School 3	0.549	
KM 3 Add/Sub			
	School 2 > School 1	0.035	School 2 > School 1
	School 3 > School 1	0.899	
	School 2 > School 3	0.769	
Time 1 MS			
	School 2 > School 1	0.017	School 2 > School 3 > School 1
	School 3 > School 1	0.994	
	School 2 > School 3	0.496	

(*table continues*)

Source	Schools	*p*	Order
Time 1 SSM			
	School 2 > School 1	0.041	School 2 > School 1 > School 3
	School 1 > School 3	0.728	
	School 2 > School 3	0.176	
Time 2 EA			
	School 2 > School 1	0.043	School 2 > School 3 > School 1
	School 3 > School 1	0.648	
	School 2 > School 3	0.973	
Time 2 MS			
	School 2 > School 1	0.008	School 2 > School 3 > School 1
	School 3 > School 1	0.613	
	School 2 > School 3	0.894	
Time 3 EA			
	School 2 > School 1	0.022	School 2 > School 3 > School 1
	School 3 > School 1	1.000	
	School 2 > School 3	0.468	
Time 3 MS			
	School 2 > School 1	0.081	School 3 > School 2 > School 1
	School 3 > School 1	0.221	
	School 3 > School 2	0.785	
Time 3 SSL			
	School 2 > School 1	0.051	School 2 > School 1 > School 3
	School 1 > School 3	0.828	
	School 2 > School 3	0.252	
Time 4 MS			
	School 2 > School 1	0.006	School 2 > School 3 > School 1
	School 3 > School 1	0.549	
	School 2 > School 3	0.919	
Time 4 SSL			
	School 2 > School 1	0.015	School 2 > School 1 > School 3
	School 1 > School 3	0.570	
	School 2 > School 3	0.076	

[a] Dunnett's C used instead of Tukey's HSD because of violation of assumptions of homogeneity.

For the ANOVAs with the grade 4 participants only, the tests of homogeneity of variance were not violated in all cases except for TIME 2 EA ($p = .026$). As shown in Table 29, no significant differences were found for any NRTs. The ANOVA was significant for the all MCBM tests with the exception of Time 3 and Time 4 EA. Post-

hoc comparisons using the Tukey's HSD and Dunnett's C (for Time 2 EA) are presented in Table 30. Results indicate that School 2 mean scores were greater than School 1 on all MCBM probes and greater than School 3 on all tests with the exception of Time 1 EA, Time 3 EA, Time 3 SSL and Time 4 SSL. No significant differences were found between School 1 and School 3 on any of the tests.

Finally, a series of ANOVAs were run with grade 6 participants included. The tests of homogeneity of variance were not violated in any case. As shown in Table 29, the ANOVA for the NRTs was significant for the WJ-ACH-III and KM 3 Add/Sub. For the MCBM probes, all but the following probes were significant: Time 1 EA and Time 4 EA and Time 1 and Time 3 SSL. Post-hoc comparisons using the Tukey's HSD test are presented in Table 30. Results indicate that School 2 mean scores were significantly greater than School 1 on all but two probes (Time 3 SSL and Time 3 MS). No significant differences were found between School 1 and School 3 or between School 2 and School 3 on any of the tests.

In sum, results of the ANOVA analyses indicated that students at School 2 typically displayed the highest performance on all tests followed by School 1 although only statistically significant differences were found between School 2 and School 1. Students from School 3 achieved the lowest scores on all measures except for the WJ-ACH-III. Further, student performance differed significantly by school on the majority of the math tests administered. This indicates that the mean scores for all participants at each of the school were often quite different from each other.

When only the second-grade participants were compared among schools, however, only three MCBM probes were found to have significant differences. This

indicated that the second grade participants tended to score similarly on all of the probes regardless of the school they attended. Further, among the statistically significant differences that were identified, School 1 maintained the highest mean scores followed by School 3–with the exception of the Time 2 EA probe in which School 3 had the highest mean scores. Students from School 2 maintained the lowest mean scores on all of the statistically significant probes. In addition, the significant differences that were found were often between School 2 and the other schools, but the mean scores of School 1 and School 3 often did not differ significantly from each other.

When only the fourth-grade participants were compared among schools, the mean scores from different schools varied significantly on many different measures. In general, the mean scores were typically highest for School 2; in addition, the significant differences that were found were often between School 2 and the other schools, but the mean scores of School 1 and School 3 often did not differ significantly from each other.

When only the sixth-grade participants were compared, the mean scores for School 2 were typically significantly different from those of School 1, but no other pairwise comparisons were found to be significantly different. Similar to the results for the fourth grade students, the majority of measures were also found to contain statistically significant differences among the schools.

CURRICULUM VITAE

Kyle Max Hancock, PhD

CONTACT INFORMATION

Home: 530 S. 200 W.
Wellsville, UT 84339
kyle,hancock@usu.edu
435-245-6691 (home)
435-757-4078 (cell)

INTERNSHIP

Aug. 2008 - Aug. 2009 — Primary Children's Medical Center
Salt Lake City, UT

Training and experience with outpatient and inpatient behavioral health including assessment; individual, family, and group therapy; consultation and liaison; and coordination of psychological land psychiatric care for children, adolescents, and their families in hospital and outpatient behavioral health clinics operating within a multidisciplinary system. Primary presenting problems ranged from medically ill children (e.g., patients with issues associated with organ transplant, diabetes, and medical trauma) to internalizing and externalizing problems resulting from various disorders of childhood and adolescence (e.g., autism, bipolar disorder, anxiety, depression, and conduct disorder).

Primary supervisors: Merrill Kingston, PhD
Matthew D. Christensen, PhD

EDUCATION HISTORY

PhD 2009 — Combined Clinical, Counseling, and School Psychology
(APA Accredited)
Utah State University, Logan, UT

Dissertation: The Utility of Mathematics Curriculum-Based Measurement to Predict Student Risk Status on Standardized Academic Achievement Measures.

Chair: Donna M. Gilbertson, PhD

MS School Psychology (NASP Approved)
2006 Utah State University, Logan, UT
Thesis: Social Interactions and Bullying in Withdrawn Children: An Evaluation of Generalization Strategies Within a Social Skills Training Intervention
Chair: Donna M. Gilbertson, PhD

BS Psychology (major), Family and Human Development (minor)
Utah State University, Logan, UT

CERTIFICATION

2007 **School Psychology Educator License**
Utah State Office of Education

PRACTICUM EXPERIENCE

June 2007 - May 2008 **Student Therapist**, Advanced Pediatric/Clinical Child Practicum
Budge Clinic, Logan Regional Hospital, Logan, UT
Responsibilities: Consultation and collaboration with medical professionals in a primary care pediatric medical clinic, including providing individual and family therapy for children and adolescents; therapy primarily consisted of behavior management and cognitive-behavioral interventions for youth with psychological and medical problems (e.g, encopresis, trauma, disruptive behavior disorders, anxiety, depression).
Supervisor: Gretchen Gimpel Peacock, PhD

June 2006 - Aug. 2007 **Student Therapist**, Advanced Clinical Child Practicum
USU Psychology Community Clinic,
Utah State University, Logan, UT
Responsibilities: Individual and family therapy in an outpatient psychology clinic, including assessment, diagnosis, formulation, and implementation of behavioral and cognitive-behavioral interventions to children and adolescents with diverse psychological problems (e.g., disruptive behavior disorders, anxiety, depression) and learning difficulties (e.g., learning disabilities, mental retardation).
Supervisor: Gretchen Gimpel Peacock, PhD

Aug. 2005 - May 2006 **Student Therapist**, Counseling Practicum
Counseling Center, Utah State University, Logan, UT
Responsibilities: Individual therapy in a college counseling center, including assessment, diagnosis, formulation, and implementation of cognitive-behavioral interventions to adolescents and adults with a variety of presenting problems, including alcohol/substance abuse, interpersonal violence/abuse, marital/relationship problems, sexual identity issues, depression, and anxiety.
Supervisor: David W. Bush, PhD

Aug. 2004 - May 2005 **School Psychology Practicum Student**, School Psychology Practicum, Ogden City Schools, Ogden, UT
Responsibilities: Behavioral and academic intervention, consultation, and assessment within a public school; member of multidisciplinary teams; coordinated and provided services to students with disruptive behaviors, learning difficulties, cognitive impairments, and physical disabilities.
Supervisors: Donna Gilbertson, PhD
Cher L. King, PhD

Jan. 2004 - Aug. 2004 **Student Therapist**, Counseling Psychotherapy Practicum
USU Psychology Community Clinic,
Utah State University, Logan, UT
Responsibilities: Individual therapy in an outpatient psychology clinic, including assessment, diagnosis, formulation, and implementation of interventions to adolescents and adults with various presenting problems.
Supervisor: Susan L. Crowley, PhD

SUPERVISED PROFESSIONAL EXPERIENCE

Apr. 2007 - Aug. 2008 **Graduate Assistant Therapist**
Avalon Hills Residential Eating Disorders Program
Petersboro, UT
Responsibilities: Member of multidisciplinary team providing individual, family, group, and equine-assisted therapy in a residential facility for adolescent females with eating disorders and other concomitant diagnoses, including OCD, bipolar disorder, borderline personality disorder, substance abuse, depression, and anxiety; activities include assessment, diagnosis, and implementation of behavioral and cognitive-behavioral interventions.
Supervisors: Dave Christian, PhD
David Stein, PhD

Aug. 2006 - June 2007 **School Psychologist**
Davis School District, Cook Elementary, Syracuse, UT
Responsibilities: Behavioral and academic intervention, consultation, and assessment within a public school; member of multidisciplinary teams; coordinated and provided services to students with disruptive behaviors, learning difficulties, cognitive impairments, and physical disabilities.
Supervisors: Donna Gilbertson, PhD
Lorene Newbold, MS

Jan. 2006 - Present **Psychological Assessment Consultant**
Logan, UT
Consultant to psychologist in private practice, including administration, interpretation, and reporting of psychoeducational assessments.
Supervisor: Melanie Domenech-Rodriguez, PhD

Aug. 2005 - June 2006 **School Psychologist Intern**
Davis School District, Cook Elementary, Syracuse, UT
Responsibilities: Behavioral and academic intervention, consultation, and assessment within a public school; member of multidisciplinary teams; coordinated and provided services to students with disruptive behaviors, learning difficulties, cognitive impairments, and physical disabilities.
Supervisors: Donna Gilbertson, PhD
Lorene Newbold, MS

Jan. 2005 - Aug. 2005 **Psychometrist**
Weber School District, Ogden, UT
Responsibilities: Psychoeducational assessment of children, adolescents, and adults ages 3-21.
Supervisor: Maren McFarland, MS

PROFESSIONAL EXPERIENCE

May 2003 - Feb. 2005 **Development and Family Service Specialist**
Centro de la Familia de Utah (Migrant Head Start), Providence, UT
Responsibilities: Psychoeducational screening of children enrolled in a full-service preschool, including the design and implementation of behavioral and educational interventions for children with varying disabling conditions and their families, serving as the primary liaison between clients and their family and local school districts, participation in interdisciplinary intervention teams designed to provide educational and psychological services to clients and their families.

July 2002 - Feb. 2004 **Children's Skills Development Specialist**
Bear River Mental Health, Tremonton, UT
Responsibilities: Development and delivery of therapy services to children ages 3 to 18, including the instruction of cognitive-behavioral skills to clients with diagnoses ranging from autism to oppositional defiant disorder; interventions applied in individual, family, and group settings.

Aug. 2002 - May 2003 **Peer Advisor**
Undergraduate Advising Office, Utah State University, Logan, UT
Responsibilities: Advising undergraduate students concerning their university education, university relations and public affairs, and presenting at University events.

GRANT

2006 - 2007 **Model School Site Grant Recipient**,
Utah State Office of Education ($10,000)
Principal author on grant acquired from Utah State Office of Education for Cook Elementary to serve as a model site for one year to offer training to other schools across the state on Responsiveness to Intervention techniques.

PUBLICATION

Hancock, K. M., & Van Natter, H. (2006). Implementing a responsiveness-to-intervention approach in Davis School District: Cook Elementary and STEEP. *The Utah Special Education, 27*(2), 14-15.

CONFERENCE PRESENTATIONS

Hancock, K. M., Gilbertson, D., & Adams, E. (2008, June). *Implementing RTI: An empirical data-driven model.* Invited workshop at the Sixth Annual Conference on Effective Practices in Special Education and Rehabilitation: Interventions Across the Lifespan, Logan, UT.

Hancock, K. M. Gilbertson, D., Rosenlof, M., & Toone, S. (2007, May). *Social interactions and bullying in withdrawn children: An evaluation of generalization strategies within a social skills training intervention.* Poster presented at the annual conference, Association of Behavior Analysis, San Diego, CA.

Monson-Ford, B., Sonnek, R., Gilbertson, D., & Hancock, K. M. (2007, March).

Team support for teacher implementation of a class-wide reading intervention to evaluate responsiveness-to-intervention. Poster presented at the annual conference, National Association of School Psychologists, New York City, NY.

Hancock, K. M., Gilbertson, D., Rosenlof, M., & Toone, S. (2007, February). *Social interactions and bullying in withdrawn children: An evaluation of generalization strategies within a social skills training intervention.* Poster presented at the annual conference, Utah Association of School Psychologists, Park City, UT.

Monson-Ford, B., Sonnek, R., Gilbertson, D., & Hancock, K. M. (2007, February). *Team support for teacher implementation of a class-wide reading intervention to evaluate responsiveness-to-intervention.* Poster presented at the annual conference, Utah Association of School Psychologists, Park City, UT.

Hancock, K. M., & Gilbertson, D. M. (2006, July). *School-wide and small group preventative interventions for victims of bullying for teachers.* Invited workshop at the Fourth Annual Conference on Effective Practices in Special Education and Rehabilitation: Interventions Across the Lifespan, Logan. UT.

Hancock, K. M., Gilbertson, D. M., Van Natter, H., Francis, A., & Stokes, N. (2006, June). *A responsiveness to intervention approach: System to enhance educational performance.* Invited workshop at the Utah State Office of Education State Conference on Responsiveness to Intervention, Provo, UT.

Hancock, K. M., & Gilbertson, D. M. (2006, March). *School-wide and small group preventative interventions for victims of bullying for teachers.* Paper presentation at the annual conference, National Association of School Psychologists, San Diego, CA.

CONSULTATION PRESENTATIONS

2007 (August) *Using Data to Guide Educational Practice.* Invited workshop to the related services staff of Davis School District.

2006 - 2007 *STEEP: An empirically supported RTI system.* In accordance with model site grant requirements, multiple training workshops for all interested school districts across the state of Utah, including superintendents' councils, district directors of special education, and school-based pre-referral teams.

2006 (January) *Using Steps to Enhance Educational Performance.* A series of three invited presentations to superintendent's council and director of special education, which enabled implementation of STEEP within the school district.

2006 (May) *Managing Stress Effectively.* Invited presentation to annual

convention for senior citizen health in Brigham City, Utah, as an outreach program of the USU Counseling Center.

RESEARCH EXPERIENCE

2004 - 2005 **Graduate Researcher**, Community/University Research Initiative Grant: School-wide and small group preventative interventions for victims of bullying for teachers.

Responsibilities: Designed assessment procedures and implemented school-wide, class-wide, and small group interventions to decrease the frequency of bully victimization among middle-school students; organized, supervised, and trained undergraduate research assistants; served as primary liaison and coordinator for school staff/administration and research team.

Supervisors: Donna Gilbertson, PhD
Gretchen Gimpel Peacock, PhD
Melanie Domenech-Rodriguez, PhD
Tamara Ferguson, PhD

2003 **Research Assistant**, Investigations of bullying and negative social interactions among elementary students, Department of Psychology, Utah State University, Logan, UT

Responsibilities: Assisted in the design and use of an informal functional assessment protocol for assessing frequency and intensity of bullying behavior and supervised and coordinated undergraduate research team.

Supervisor: Donna Gilbertson, PhD

2002 - 2003 **Research Assistant**, Establishing a letter-naming fluency and a letter-sound fluency CBM probe, Department of Psychology, Utah State University, Logan, UT

Responsibilities: Designed and implemented a curriculum-based measurement protocol in conjunction with a short-term, intensive remediation program; supervised/coordinated undergraduate research team.

Supervisor: Donna Gilbertson, PhD

TEACHING EXPERIENCE

Graduate Instructor **Educational Psychology** (Psy 3660), Utah State University

Semesters taught: Spring 2009, Spring 2007, Spring 2006, Spring 2005.

Graduate Instructor	**History and Systems of Psychology** (Psy 5110), Utah State University Semester taught: Fall 2008
Graduate Instructor	**General Psychology** (Psy 1010), Utah State University Semesters taught: Fall 2007, Summer 2007, Fall 2006, Summer 2006
Graduate Teaching Assistant	**Psychological and Educational Consultation** (Psy 6340), Utah State University Semester taught: Fall 2006
Graduate Instructor	**Social Psychology** (Psy 3510), Utah State University Semesters taught: Fall 2007, Fall 2005
Graduate Instructor	**Psychometrics** (Psy 5330), Utah State University Semester taught: Summer 2005
Graduate Instructor	**Developmental Psychology** (Psy 1100), Utah State University Semester taught: Summer 2004
Graduate Instructor	**Psychology of Human Adjustment** (Psy 1210), Utah State University Semesters taught: Spring 2006, Fall 2005, Spring 2004, Fall 2003
Teaching Assistant	**Abnormal Psychology** (Psy 3210), Utah State University Semester: Spring 2003

LEADERSHIP/VOLUNTEER EXPERIENCES

2006 - Present	**APAGS Campus Representative**, Utah State University Responsibilities: Coordination and communication of advocacy efforts for graduate students in psychology; also responsible for recruiting members into APA and co9mmunicating APAGS information.
2005 - 2006	**Graduate Student Representative**, Psychology Department Utah State University Responsibilities: Representation of the program student body with program faculty, service as liaison between students and faculty, and voting member of program council. Elected to position by graduate student body.

HONORS AND AWARDS

2007 **Psychology Department Travel Scholarship**, Association for Behavior Analysis Annual Convention ($300)

2007 **Graduate Student Senate Travel Scholarship**, Association for Behavior Analysis Annual Convention ($300)

2005 **Rookie of the Year**, Davis School District

2005 **Graduate Student Senate Travel Scholarship**, National Association of School Psychologists Annual Convention ($300)

2003 **Outstanding Achievement Award**, Utah State University

PROFESSIONAL ASSOCIATION MEMBERSHIPS (* = membership by invitation only)

Responsiveness to Intervention Roundtable, Utah State Office of Education*
Utah Bullying Task Force, Utah Personnel Development Center*
American Psychological Association
National Association of School Psychologists
Utah Psychological Association