Response to Intervention Universal Math Fluency Screenings: Their Predictive Value for Student Performance on National and State Standardized Achievement Tests in Arizona

by

Thomas J. Gambrel

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2014 by the
Graduate Supervisory Committee:

Linda C. Caterino, Chair
Jill Stamm
Samuel DiGangi

ARIZONA STATE UNIVERSITY

December 2014

ABSTRACT

The most recent reauthorizations of No Child Left Behind and the Individuals with Disabilities Education Act served to usher in an age of results and accountability within American education. States were charged with developing more rigorous systems to specifically address areas such as critical academic skill proficiency, empirically validated instruction and intervention, and overall student performance as measured on annual statewide achievement tests. Educational practice has shown that foundational math ability can be easily assessed through student performance on Curriculum-Based Measurements of Math Computational Fluency (CBM-M). Research on the application of CBM-M's predictive validity across specific academic math abilities as measured by state standardized tests is currently limited. In addition, little research is available on the differential effects of ethnic subgroups and gender in this area. This study investigated the effectiveness of using CBM-M measures to predict achievement on high stakes tests, as well as whether or not there are significant differential effects of ethnic subgroups and gender. Study participants included 358 students across six elementary schools in a large suburban school district in Arizona that utilizes the Response to Intervention (RTI) model. Participants' CBM-M scores from the first through third grade years and their third grade standardized achievement test scores were collected. Pearson product-moment and Spearman correlations were used to determine how well CBM-M scores and specific math skills are related. The predictive validity of CBM-M scores from the third-grade school year was also assessed to determine whether the fall, winter, or spring screening was most related to third-grade high-stakes test scores.

DEDICATION

I dedicate this dissertation to my entire family for their tireless love, support, and assistance in the completion of this odyssey. A special dedication goes out to my dear mother, who did not live to see this, but whose love, generosity, and pride in her boy I feel foreverlastingly. I also offer a very special dedication to my beautiful wife, my gift from Heaven, my Ashleigh, for selflessly showering me with the patience and support I've needed throughout the past decade to not only achieve this goal, but also to be able to begin pursuing it in the first place. I love you all dearly! Apparently, it also takes a village to complete a PhD program!

ACKNOWLEDGEMENT

I would like to acknowledge my dissertation committee for their invaluable guidance and support in the completion of this project. I thank Dr. Samuel DiGangi for taking time out of his busy schedule to be a part of my committee and make himself available to provide assistance. I thank Dr. Jill Stamm for being such a wonderful supervisor, teacher, mentor, and friend to me throughout my entire ASU experience. Finally, in deep gratitude I acknowledge Dr. Linda Caterino; from the very first phone interview in my Brooklyn apartment to this moment, she has served as a generous and loving champion for my sake and the sake of all of her students: Thank you Dr. C!

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

Introduction and Literature Review

The past decade in American education has seen a significant shift in emphasis, from a more process oriented perspective to one where results and accountability have become the prevailing standards (Reschly & Bergstrom, 2009). Following the 2002 reauthorization of the No Child Left Behind legislation (NCLB, 2002), states were charged with putting more rigorous systems in place to specifically address critical academic skill development, empirically validated instruction and intervention techniques, and overall student performance as measured on annual statewide achievement tests. Furthermore, the 2004 reauthorization of the Individuals with Disabilities Education Act (IDEA) afforded state departments of education the option of utilizing instructional processes based on "the child's response to scientific, research-based interventions" (IDEIA, 2004; Public Law 108-446) in the diagnosis of specific learning disabilities. This evidence-based process is referred to as Response to Intervention (RTI).

The tool used within RTI systems for screening and progress monitoring is Curriculum-Based Measurement (CBM; Shinn, 2008), and the most commonly used CBM with the greatest amount of research support is the measure of reading fluency (Shinn, 2008; Thurber, Shinn, & Smolkowski, 2002). Reading fluency is commonly believed to be one of the best overall predictors of general reading proficiency during the primary school years (Reschly, Busch, Betts, Deno, & Long, 2009; Shinn, 2008). Other CBMs, such as those for math, have not been researched as thoroughly even though they

are being used more and more as key components of RTI systems (Fuchs, Fuchs, Yazdian, & Powell, 2002; Shinn, 2008; Thurber et al., 2002; Vaughn & Fuchs, 2003).

Response to Intervention

Response to Intervention is a tiered instructional system that typically begins with universal screenings in reading, writing, and math that are meant to assist educators in the identification of students who are at risk for low academic achievement (Anderson, Lai, Alonzo, & Tindal, 2011; Fuchs & Fuchs, 2001; Reschly & Bergstrom, 2009). This, in turn, leads to the implementation of evidence-based interventions designed to provide further support for struggling students while monitoring for progress. Although RTI systems with as many as five tiers are known to exist, the three-tier paradigm is the most common (Barnes & Harlacher, 2008; Reschly & Bergstrom, 2009).

Tier one is universal in scope and consists of quality instruction and behavioral supports in the general education classroom. Tier two is marked by a schedule of small-group, evidence-based interventions for students who are found to be performing significantly below their same-aged peers in one or more academic skill areas (e.g., early reading or math failure). Students who fail to respond to the prescribed interventions are then moved to tier three. Depending on the model of RTI being endorsed by a particular school district, tier three may be comprised of more intensive, individualized interventions (Hughes, 2008), a comprehensive psychoeducational evaluation by a multidisciplinary team (Wodrich & Schmitt, 2006; Wodrich, Spencer, & Daly, 2006), or both (Fuchs, Mock, Morgan, & Young, 2003). In any case, tier three is where a determination is ultimately made regarding eligibility for specialized instruction. In some

models of RTI, being assigned to tier three is, in and of itself, synonymous with special education eligibility (Reschly, 2005; Shinn, 2005).

It should be noted that one of the major challenges with any attempts to define, elaborate on, or even effectively implement RTI systems is a lack of national standardization (Hughes, 2008; Reynolds, 2008). As such, there exists substantial variation in RTI paradigms from state to state and from district to district (Barnes & Harlacher, 2008; Berkeley, Bender, Peaster, & Saunders, 2009). For instance, in some RTI systems tier one includes more than just the expectation of quality instruction and support, and so at-risk students begin receiving small-group supplemental instruction in the classroom immediately (Hughes, 2008). Contrast this with Fuchs and Fuchs (2001), who endorsed an RTI model where tier one had at-risk students being identified through universal screening and then subsequently being monitored for eight weeks in the absence of any interventions in order to find the subset of students that did not respond adequately to the general education curriculum. Additionally, it is important to remember that in the most recent reauthorization of the Individuals with Disabilities Education Act (IDEIA, 2004; Public Law 108-446) states are provided the option of specific learning disability identification through either insufficient progress following the adequate implementation of an RTI system *or* the finding of a significant discrepancy between aptitude and achievement. As a result, considerable variation still exists in exactly how RTI is utilized.

States and/or local education agencies (LEAs) that have adopted a particular RTI model either utilize it exclusively for the task of determining special education eligibility or as a pre-referral strategy to be implemented prior to a full psychoeducational

evaluation. Comprehensive RTI systems are currently used as the sole determinant of disability diagnosis in fourteen states, including Colorado, Iowa, and Florida (Reschly & Bergstrom, 2009; Reynolds & Shaywitz, 2009; Zirkel, 2013), while a host of experts endorse the complementary approach in which a shift to tier three leads to full evaluation (Fuchs et al., 2003; Kavale, Kauffman, Bachmeier, & LeFever, 2008; Reynolds, 2008; Reynolds & Shaywitz, 2009; Swanson, 2008; Wodrich & Schmitt, 2006; Wodrich et al., 2006).

RTI: A Brief History

Although the actual term "response to intervention" only emerged as recently as the late 1990s to the early 2000s (Reschly & Bergstrom, 2009), the paradigm itself is quite old (Gresham, 2007; Swanson, 2008) and is being used more and more in schools as a pre-referral strategy. RTI emerged out of the research base on assessment and interventions, which has its roots in fields such as applied behavior analysis, instructional science, and behavioral consultation (Gresham, 2007; Reschly & Bergstrom, 2009). The basic concept, originally meant as a means to further clarify the definition of learning disabilities, has been discussed in the literature for many decades (Swanson, 2008; see also Weiderholt, 1974). The research supporting the core tenets of RTI continued to expand with the development of behavior assessment and intervention techniques in the 1960's and 1970s. Hence, they share several key features such as direct, measurable behavioral observation performed in naturalistic settings and efficient, reliable measurement tools with short retest turn-around capability (Reschly & Bergstrom, 2009). Throughout the past two decades many different studies and meta-analyses have provided consistent empirical support for the benefits of the instructional, behavioral, and

intervention programs found within RTI systems (Kavale, 1990, 2005, 2007; Kavale & Forness, 1999; VanDerHeyden, Witt, & Gilbertson, 2007). For instance in their research, VanDerHeyden, Witt, and Gilbertson (2007) sought to analyze the RTI system as an integrated, dynamic, problem-solving procedure as implemented by real frontline educational professionals rather than paid research associates. They found that properly applied RTI systems resulted in fewer psychoeducational evaluations, decreased disproportionality through increased decision-making accuracy, and substantially lowered district expenditures.

Around the time that RTI began to emerge in the applied literature, U.S. educational policy was being closely reviewed in order to confront persistent nationwide complaints such as low reading achievement levels, low overall academic performance comparative to other countries, a wide variation in special education practices, and the failure to implement evidence-based curricula and interventions (Reschly, 2008). As a result, a variety of research agencies made policy recommendations aimed at increasing accountability and achieving better overall results through the use of empirically supported educational practices (see *A New Era: Revitalizing Special Education for Children and Their Families,* 2002; Bradley, Danielson, & Hallahan, 2002; National Reading Panel, 2000). This perceived need for policy change was instrumental to the 2002 reauthorization of the Elementary and Secondary Education Act now known as No Child Left Behind (NCLB, 2002), which contained several key provisions that helped to promote the advancement of RTI methodologies. These included (a) the frequent assessment of educational outcomes, (b) accountability for the results of those assessments, (c) the required use of empirically validated instruction and intervention

techniques, and (d) methods aimed at the early identification of academic skill deficits (NCLB, 2002; Reschly & Bergstrom, 2009). Subsequently, the 2004 reauthorization of IDEA would end up being consistent with many of these provisions. In this way, NCLB essentially became the system of accountability through which IDEIA would be supported (Tilly, 2008).

As previously mentioned, it was through the 2004 IDEA reauthorization that the detection of a severe discrepancy between aptitude and achievement was no longer mandatory in order to diagnose a specific learning disability. Consequently, RTI was introduced into the legislation as a viable option by which learning disabilities could be diagnosed (Jacob & Hartshorne, 2007; Wodrich & Schmitt, 2006). This development catapulted RTI into the national education consciousness and currently, it is being developed for implementation in all states either as the sole requirement for SLD placements or as a pre-referral intervention system to be used in conjunction with psychoeducational evaluations (Berkeley et al., 2009; Reschly & Bergstrom, 2009).

Key Features of RTI

Response to Intervention is comprised of economical and pragmatic tools that are grounded in the concept of prevention. Preventative practices such as early detection and early intervention are endorsed across virtually all health professions as the most effective way to reduce the prevalence and severity of a given ailment. In the domain of RTI the chief "ailments" to be identified are academic in nature. Regardless of the wide variation in RTI design and utilization, there are still core features consistently found to be present (Gibbons, 2008; Reschly & Bergstrom, 2009; VanDerHayden et al., 2007).

The tools and processes typically employed within RTI systems are as follows: (a) service delivery occurs in a multi-tiered format in which the intensity of intervention is directly proportionate to the needs of the student; (b) educational and behavioral goals and benchmarks are clearly identified; (c) universal screenings in reading, math, and sometimes writing are generally performed three times a year to identify students who may be at risk academically and/or behaviorally; (d) the universal screenings, in combination with *can't do/won't do* assessments (VanDerHeyden & Witt, 2008), help to determine students' needs by identifying gaps between expected and actual performance, (e) actual interventions employed are evidence-based; and (f) the CBM progress monitoring data is frequently compared to appropriate benchmarks and projected goals.

Curriculum-Based Measurement

Curriculum-Based Measurement assessments are a systematic means of quantifying the growth of students' basic academic skill competence in a short amount of time (Deno, 1985). This is done by capturing a particular academic behavior such as basic math computational fluency, oral reading fluency or written expression with brief, standardized measurement tools that are grade-level appropriate and constructed from content rooted in the district's general curriculum (Hintze, 2009; Shinn, 2008). To illustrate by way of an example, one commonly used type of CBM is for math computational fluency. The procedure involves simply giving a student a math probe consisting of a single page of math problems (see Appendix B) where performance is measured in total digits correct per two minutes. This scoring method is distinct from total answers correct in that credit is actually being given for each digit that is in the correct place value. This quick, two-minute assessment produces potentially valuable information about how a student

compares to his or her peers with regard to the specific performance objectives of a particular curriculum. This information can then be utilized in a number of ways, such as assessment of the retention of critical academic skills and level of material mastery, problem identification, differentiated intervention development, and intervention progress monitoring (Hintze, 2009).

In an early research review, Marston (1989) focused on CBM as a much needed alternative to traditional models of academic assessment and decision-making due to problematic issues on both the technical and social policy levels. Marston argued that among the thousands of standardized psychometric instruments available, there were many published reports of substantial challenges with regard to technical adequacy, especially when considering their use on children with disabilities and the unacceptable risk of potentially flawed methods of special education placement. He further argued that issues of time, expense, and a demonstrated lack of consistent decision-making criteria from school to school subverted best efforts to reliably and validly meet students' educational needs. Marston went on to make a case for CBM as a technically sound measure of student performance that is directly related to curricula, is sensitive to improvement in academic achievement over time, and is a reliable, valid measure of basic skill content areas.

As the hinge on which formal evaluative decisions pivot, the information provided by CBM is essential to any effective RTI service delivery. In this context, CBMs produce much of the "data" in data-based decision making, which is the hallmark of all RTI models. They are a particularly useful method of assessment because they are brief, sensitive to short-term growth, and are able to be repeated with regular frequency

(Hintze, 2009). In addition, they now have some thirty years of research supporting their validity (Deno, 1985; Fletcher, Denton, & Francis, 2005; Reschly, Busch, Betts, Deno, & Long, 2009; Shinn, 2008). There is also a substantial amount of research suggesting that student achievement outcomes improve when teachers use CBM data to shape differentiated instructional strategies (Stecker & Fuchs, 2000; Stecker, Fuchs, & Fuchs, 2005). This is much easier done with CBMs because of their formative nature; that is, they allow teachers to better adjust any modifications to instruction more fluidly throughout the year in order to ensure greater academic success in a shorter period of time (Fore, Boone, Lawson, & Martin, 2007).

Educators and other direct stakeholders such as administrators and school psychologists need this type of practical, efficient, and continuous measurement capability (Jiban & Deno, 2007) to better equip them in their efforts to use evidence-based practice that ensures the biggest possible effect size with the smallest possible intrusion on instruction time. This is particularly relevant within current education policy and law, as NCLB, the IDEA, and the President's Commission on Excellence in Special Education (United States Department of Education Office of Special Education and Rehabilitative Services, 2002) all include CBM as an important part of broader evidence-based preventative practices. In addition, Ysseldyke and his colleagues included CBM as a crucial competency for best practices in *School Psychology: A Blueprint for Training and Practice III* (Ysseldyke, Burns, Dawson, Kelley, Morrison, Ortiz, Rosenfeld, & Telzrow, 2006; 2008) when they emphasized accountability through data-based decision making and the application of scientific methodologies.

Curriculum-Based Measurement for Universal Screening

In the current age of accountability and inadequate educational budgets, cost-effective and efficient formative assessment methods such as CBMs are crucial tools to have in order to better instruct, screen, intervene, and monitor according to students' instructional needs (Erickson, Ysseldyke, Thurlow, & Elliot, 1998). This is particularly relevant when it comes to national and statewide tests of academic achievement. Having a fast and durable way to assess basic academic skills that can also be predictive of student performance on high-stakes standardized tests is important for several reasons (Helwig, Anderson, & Tindal, 2002), not the least of which being that such standardized tests are large in scale, time-consuming, and are usually only administered once a year. Curriculum-Based Measurements afford educators the ability to take academic snapshots of their students' conceptual understanding of key skill areas throughout the school year, thereby guiding efforts to properly differentiate instruction, implement interventions or re-teach a topic altogether if needed. Further, CBM data can provide additional evidence for a school's adequate yearly progress (Helwig et al., 2002). In general, CBM is considered a key part of "general school improvement efforts" and as such, it is vitally important that it be closely aligned with curriculum and instruction if it is expected to lead to positive educational outcomes (Elliott, Huai, & Roach, 2007; Shinn, 2008).

Curriculum-based measurements generally take on several key roles in decision making processes regarding academic progress and competence, such as (a) universal screenings conducted in order to identify which students are in need of intervention support; (b) *can't do/won't do* assessments, which are meant to tease out the subset of students who fall below benchmark due to motivational factors; (c) identifying an

accurate instructional level for students who are in need of supports; (d) intervention progress monitoring for the purpose of increasing, decreasing or changing a given intervention strategy; (e) curriculum evaluation, and (f) evaluation of instructional quality (Deno, 2003; Shinn, 2008). However, the most common use of CBM is for universal screening purposes (Ikeda et al., 2008).

When considering CBM as part of an RTI system, universal screening is a key feature of tier one and represents an initial, proactive effort to identify and track students who are at risk of academic failure (Ikeda et al., 2008). This stands in contrast to the traditional "wait-to-fail" model in which struggling students do not begin receiving intervention until the point of significant distress with regard to academic health (Deno, 2003; Vaughn & Fuchs, 2003). During universal screening, all students are assessed in one or more academic skill areas and data is generated that indicates the presence or absence of a problem as dictated by local, empirically derived standards known as benchmarks (Glover & Albers, 2007; Shinn, 2008). In the event that a problem is detected, further investigation is then warranted to see if the performance deficit amounts to a legitimate educational need. If such a need is identified, the student is then moved to tier two and provided with the necessary supports.

Curriculum-Based Measurements in Reading

There are, of course, several academic skill areas that can be assessed during universal screening (e.g., reading, written expression, math, spelling), but the most commonly researched and utilized CBM is the oral reading fluency assessment (ORF; Jiban & Deno, 2007; Poncy, Skinner, & Axtell, 2005; Reschly et al., 2009; Shinn, 2008; Thurber et al., 2002). A preponderance of the research on ORF supports it as a valid and

reliable indicator of generalized reading skills and a predictor of future reading proficiency (Poncy et al., 2005; Reschly et al., 2009; Shinn, 2008).

In a meta-analytic study investigating the magnitude and variability of ORF CBM reliability estimates obtained in 28 studies from 1993 to 2008, Yeo (2011) found a high mean estimated average alternate-form reliability (r = .89), which was close to the mean alternate-form reliability estimate found in Marston's (1989) review (r = .90). Similarly, Wayman and colleagues (2007) reviewed research studies on the technical adequacy of reading CBM conducted from 1981 to 2005. They reported that a majority of findings showed strong correlations between ORF, basic reading proficiency, and reading comprehension. They further reported that ORF has received consistent support as being more strongly related to reading comprehension than are measures designed to specifically assess comprehension. The only exception was found in the first grade, where a considerable floor-effect was detected, and in the intermediate grades (e.g., 5th and 6th), where correlations tended to decrease. In general though, correlations between ORF and state standardized tests ranged from .60 to .80 across studies (Wayman, Wallace, Tichá, & Espin, 2007).

Shapiro and colleagues (2006) also reviewed literature examining the relationship between ORF measures and state assessment outcomes in eight different states and found that, on average, the reported correlations fell within the .60 to .75 range, which suggests a strong link between ORF and statewide reading goals. Furthermore, their own analysis on the relationship between ORF and student outcomes on Pennsylvania state assessments (Shapiro et al., 2006) revealed similar results, with reported correlations in the .62 to .69 range. Paleologos and Brabham (2011) analyzed the relationship between

DIBELS ORF scores and performance on the nationally normed Stanford-10 and found correlation coefficients ranging from .23 to .60 for proficient readers and coefficients up to .65 for non-proficient readers, which suggests that CBMs may be especially useful in identifying struggling students. Overall, such comprehensive support clearly establishes ORF measures as important instructional tools for educators to have at their disposal (Reschly et al., 2009).

Although CBM research in the state of Arizona is generally limited, Wilson (2005) conducted a study in the Arizona's Instrument to Measure Standards (AIMS) technical report that analyzed the correlation of DIBELS oral reading fluency benchmark scores to AIMS test performance on a sample of third-grade students (N = 241). The results indicated a strong positive linear relationship with an obtained correlation of .74 for the overall group, which was also consistent with related reading CBM research. Devena (2013) also conducted a study on a sample of third-graders (N = 321) from four schools to determine if ORF was effective in predicting high-stakes reading test scores in Arizona. One result from that study showed medium to large correlations equal to or higher than .34, indicating a positive linear relationship between reading CBM and the reading portions of the AIMS/Stanford-10 Dual Purpose Assessment (DPA) with the strongest correlations occurring with the spring screenings. Meanwhile, such an overall consensus on determining a similarly valid and efficient task to measure math proficiency is considerably less robust (Jiban & Deno, 2007).

Curriculum-Based Measurement in Mathematics

Mathematics CBM assessments are typically comprised of three core skill areas of focus: (a) early numeracy, (b) computation, and (c) applications. Of these, computational

fluency CBMs, which involve working math problems within a two to four minute time limit, receive the most research attention (Foegen, Jiban, & Deno, 2007). Thurber, Shinn, and Smolkowski (2002) conducted a confirmatory factor analytic study on 207 fourth grade students to investigate what constructs CBM-Ms actually measure validly relative to a range of other mathematics measures. Their findings identified a two-factor model of mathematics featuring computation (execution of concepts, strategies, and facts) and applications (applied word problems, measurement, etc.) as distinct yet related constructs. As expected, they found that CBMs constructed with computation items correlated more highly with the computation factor (median coefficients of .82), although they also correlated, albeit lower, with applications (median coefficients of .44). Their results also showed CBM-M as a measure of computation with high alternate form reliability (correlations between .90 and .92). This is consistent with other research supporting the reliability and validity of math CBM data in general (Burns, 2004; VanDerHeyden & Burns, 2005).

In a recent review on the critical elements needed for effective practices in RTI for mathematics, Lembke, Hampton, and Beyers (2012) reported that for students in earlier primary grades (e.g., first through third), math CBMs are typically based on counting, numeracy, and simple operations skills, while CBMs for the upper grades have more advanced concepts, such as algebraic components, integrated into their content. The former method of CBM-M development referred to utilizes *robust indicators,* which is when items are constructed with material more representative of the core mathematics areas of proficiency (Foegen, Jiban, & Deno, 2007). With this approach, measures do not so much represent a particular curriculum per se, but are characterized by select

proficiency criteria. This can be contrasted with the *curriculum sampling* method used more on CBM-M development for intermediate to older grades, in which the measure is made from a representative sample of what is typically found in the year's mathematics curriculum. As such, research on CBM-M for earlier primary grades has been entirely focused on robust indicators of numeration (Foegen, Jiban, & Deno). Thus, when analyzing the relationship between early math computational fluency and core areas of mathematical proficiency it would be expected that there would be significant associations, but that the relationship would be strongest for numeration.

In a review of research, Foegen, Jiban, and Deno (2007) identified some 578 articles that were generally related to CBM, which they then reduced to the 160 articles that featured empirical research results. Out of these, only 29 (18%) addressed mathematics measures. They reported that alternate-form reliability estimates for 3rd through 5th grade students ranged from .72 to .93, with a majority of those estimates coming in above .80. They also found that internal consistency estimates were greater than .90 in all of the studies reviewed, and that validity coefficients were found to range from moderate (r =.35) to strong (r =.87). Overall, their findings indicated that technical adequacy tended to be stronger for the 4th and 5th grade than it was for the 3rd grade (Graney, Missall, Martínez, & Bergstrom, 2009). Further, they reported that in the studies they reviewed, the relationships between CBM-M and state achievement tests tended to fall in the .50 to .70 range, which is more modest than for CBM-R (.60 to .80), but very similar to correlations reported for commercially available achievement tests of mathematics (Foegen et al., 2007; Salvia, Ysseldyke, & Bolt, 2007).

As more and more states continue to develop and adopt RTI models (Berkeley et al., 2009; Reschly & Bergstrom, 2009), expanding the research base on CBM in math continues to be an urgent need in the effort to provide formative assessments that have stronger reliability and validity for this vital academic skill area (Foegen et al., 2007; Vaughn & Fuchs, 2003) as the search for "technically and theoretically appropriate measures, largely resolved in CBM of reading, remains active in mathematics" (Foegen et al., 2007, p. 137).

Math Curriculum-Based Measures and Screening Time of Year Growth Rates

In general, the type of mathematics CBM used for fall, winter, and spring screenings can vary between single or combined computation items that are timed from 2 to 4 minutes and scored by total digits correct. On these differing types of CBM-M forms, research has shown that they have demonstrated an average test-retest reliability of .87, as well as alternate form reliability coefficients from .66 to .91(Thurber et al., 2002; Tindal, Marston, & Deno, 1983). Further, Marston (1989) has reported concurrent validity correlations in the .42 to .45 range. Researchers have looked at within-year growth patterns on math CBMs associated with universal screenings throughout the school year and whether they demonstrate consistent rates of improvement across the benchmark assessments. Results showed that the observed trends were somewhat inconsistent, but that students tended to show more growth from the winter screening to spring screening as contrasted with the period from fall to winter (Graney, Missall, Martínez, & Bergstrom, 2009), implying general growth patterns leading to a stronger overall spring showing. This is consistent with similar studies on time of year growth rates for reading CBM, which has shown that longer intervals between ORF screenings

and standardized tests tend to produce weaker correlations, with the winter and spring administration typically demonstrating the strongest relationship (Baker, Smolkowski, Katz, Fien, Seeley, Kame'enui, & Beck, 2008; Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008; Shapiro, Keller, Lutz, Santoro, & Hintze, 2006; Wanzek, Roberts, Linan-Thompson, Vaughn, Woodruff, & Murray, 2010).

Curriculum-Based Measurement and Standardized State Testing

Pursuant to ARS 15-741, the Arizona State Board of Education is mandated to implement academic content standards and to measure student achievement against those standards annually. Curriculum-Based Measurements are often referred to as being analogous to taking a students' "academic temperature". If this is so, then the annual standardized tests used by states to measure achievement standards can be thought of as the "full physical examination". One such test utilized in the state of Arizona is the Stanford-10 (Harcourt Educational Measurement, 2003), which is currently one of the assessments being used for meeting national and state standards in academics as implemented by the No Child Left Behind Act (Statistics Solutions, 2012). The second test is the Arizona Instrument to Measure Standards (AIMS), which is given in reading, writing, math, and science. The math portion of the AIMS is comprised of several skill-specific content standards: Number and Operations, Data Analysis, Probability and Discrete Mathematics, Patterns, Algebra and Functions, Geometry and Measurement, and Structure and Logic. Not only are students required to pass the AIMS in order to graduate from high school, state legislation was recently enacted (ARS, 15-101) mandating a passing score on the third-grade reading portion of the AIMS as a prerequisite to fourth-grade advancement.

As reported earlier, recent research comparing the relationship of reading CBMs to the AIMS and the Stanford-10 (Devena, 2013; Wilson, 2005) produced results consistent with good overall predictive capability. However, there has not been any such research on the relationship between math CBMs and the Arizona high-stakes tests.

Although research support for a comparably effective math-screening tool is not nearly as extensive as it is for reading (Fuchs, Fuchs, Yazdian, & Powell, 2002; Shapiro et al., 2006; Shinn, 2008; Thurber et al., 2002), some of the available results are encouraging (Clarke, Smolkowski, Baker, Fien, Doabler, & Chard, 2011; Jiban & Deno, 2007). For instance, Shapiro and colleagues (2006) examined the relationship between CBMs of reading and math computation and Pennsylvania's state achievement test, as well as with the Stanford-9 norm-referenced achievement test. The researchers utilized a stratified random sample of second through fifth grade students drawn from six elementary schools in a Pennsylvania school district. For the math analyses ($N = 475$) they found moderate significant correlations between CBMs of math computational fluency and student outcomes on Pennsylvania's state achievement test, with average reported coefficients in the .50 to .53 range. Further, a regression analysis went on to show that the winter screening served as the strongest predictor to test scores, while also indicating that the fall screening was the weakest overall predictor. The outcome of analyses of CBM-M with the Stanford-9 produced moderate to strong correlations that ranged from .45 to .72.

Clarke and colleagues (2011) conducted research on the efficacy of an early mathematics intervention program and produced results supporting the moderate concurrent and predictive validity of CBM-Ms. These results are consistent with findings

from several other studies on math CBMs that likewise show a moderate correlation to various types of standardized achievement tests (Foegen et al., 2007; Jiban & Deno, 2007; Thurber et al., 2002). Jiban and Deno (2007) investigated the technical adequacy of grade level CBM-M and CBM-R toward predicting outcomes on third (n = 35) and fifth (n = 49) grade standardized math tests in Minnesota. Results showed weak reliability of a one-minute CBM-M at predicting third grade standardized test performance, while demonstrating a greater reliability of scores (.65 to .86) for the fifth grade. When scores from two one-minute CBM-M administrations were aggregated, moderate correlations were obtained for both grade levels. Meanwhile, regression analyses showed that the math and reading CBMs each made significant unique contributions toward explaining performance but that together they explained 52% of the variance in fifth grade test performance and 27% for third grade.

In a 2002 study Helwig, Anderson, and Tindal examined the predictive value of a noncomputational conceptual math CBM task on the performance of eighth-grade general education students (N = 171) from eight western school districts on a computer adaptive state math achievement test. The researchers correlated the eighth grade CBM-M scores with scores on a high-stakes test and found that the math probes used in their research were successful to 87% accuracy at predicting which students would meet the state math standards, with correlations ranging from .61 to 80. While this kind of result is promising, a substantial challenge still remains in that conceptual, as opposed to computational, CBMs tend to be much more involved, requiring much more time and energy to score and analyze. For this reason, more research is needed to identify and support the use of faster and easier to use math CBMs.

Curriculum-Based Measurement and Differential Effects across Subgroups

The previously noted changes in federal special education policy and law (IDEA, 2004; NCLB, 2002) and the resulting shift in America's educational focus from a process oriented approach to one of accountability, as measured by high-stakes testing, has helped to drive the increased use of CBM as a tool to monitor student progress on state academic standards. Despite this fact, a decade later there is still relatively little available research on the differential effects of CBM across gender and ethnic subgroups (Adkins, 2013), with most of the research being focused on oral reading fluency, neglecting math almost entirely. Further, there is even less research still on ethnic and gender differences related to performance on state standardized tests of mathematics. As a result, the data made available for this study was utilized to investigate relevant differential effects across gender and ethnicity.

Gender Differences

One area worthy of investigation is the purported gap in math achievement between males and females (Beal, 1999) and how this presents on CBM-M performance data (MacMillan, 2001). While it has been historically reported in the research (Beal) that boys perform better than girls in math and science, Sadker (1999) observed that this gap has been rapidly decreasing over the past 30 years. Further, Cole (1997) conducted a meta-analysis on several national and international research studies covering math achievement in grades four through 12 and found that gender differences in math performance are trivial in grades four through six, but that on average, males begin to perform significantly better than females between the 8^{th} and 12^{th} grades. These results are also supported by Leahey and Guo (2001), who found that while boys and girls start

on an even plane during the early elementary school years, boys exhibit more accelerated math achievement in the middle school years. In order to see how findings like these might be supported or refuted by CBM-M performance data, MacMillan (2001) investigated scores from nearly 1500 second to seventh-grade students using Many-Faceted Rasch Measurement (Linacre, 1994), which is a model of analysis specifically designed to assist with performance assessments and paired comparisons. MacMillan was able to determine that there were no significant gender differences across grade levels on CBM-M math scores in the sample. However, these results were obtained over a decade ago and, just as with ethnicity, there remains the need for further research on the relationship between gender and both curriculum-based measures and standardized achievement testing in math.

Tsui (2007) reported that from the period of 1990 to 2003, boys and girls from the fourth, eighth, and twelfth grades were found to perform comparably in mathematics overall. This trend remained consistent into the 2012 surveys, where the only overall statistically significant math achievement gain reported was for 13 year-old female students (National Center for Educational Statistics, 2013). Other studies have reported that girls tend to achieve higher classroom grades in math, while boys tend to obtain higher scores on standardized math tests (Arroyo, Burleson, Tai, Muldner, & Woolf, 2013; Hyde, Lindberg, Linn, Ellis, & Williams, 2008). Despite a respectable amount of evidence to the contrary, stereotypes persist that girls have inferior mathematical ability compared to boys (Hyde, et al., 2008). Historically, these beliefs have been attributed to gender differences purported to favor males in the area of basic spatial reasoning, including the ability to perform mental rotations (Casey, Nuttall, Pezaris, & Benbow,

1995), as well as an affective component related to reports that as they progress through school, girls develop increasingly more negative attitudes toward mathematics in general (Hyde et al., 2008; Royer & Walles, 2007). While there is no lack of research examining gender differences in mathematics, there is very little literature addressing these differences with regard to CBM-M and high stakes state testing in the new era of RTI since the reauthorization of IDEA.

Ethnic Differences

The work of 21st Century educators, administrators, and school psychologists is largely based in public schools that are becoming increasingly more diverse (Espinosa, 2005; Kranzler, Flores, & Coady, 2010; Ortiz, 2006; Sullivan & Kucera, 2011). For instance, during the period between 1972 and 2007, Hispanics represented the fastest growing minority presence in American schools, and by 2007 a little more than 20% of students in the 5 to 17-year-old age group had a non-English home language (Kranzler et al., 2010). This trend in population shift is uniquely represented in the state of Arizona. Demographic information from the 2013-2014 academic year shows that Hispanic students currently represent up to 44% of the total statewide school-age population (Arizona Department of Education, 2014), which is in line with recent U.S. Census Bureau estimates that Hispanics made up some 30 percent of the overall state population in 2012 and that, overall, some 43 percent of the Arizona population considered themselves to be members of an ethnic minority group (Nintzel, 2013). While student populations become more ethnically diverse, the teaching force remains predominantly middle-class and White (Whitebook, 2003). This growing discrepancy between the student body and school personnel underscores the need for all educators to be

knowledgeable about ethnic differences in educational attainment (Espinosa, 2005), including testing outcomes. This fact may be especially relevant in the state of Arizona moving forward, as reports indicate that the growth of the Hispanic population alone has doubled since the early 1990s. In fact, it is reported that Hispanics will be a minority-majority population within the next 15 to 20 years (Nintzel, 2013).

When considering this ongoing population shift, it is important to take into account the well-documented ethnic differences in academic performance, in general (James, Jurich, & Estes, 2001; Morgan & Mehta, 2004; Nyberg, McMillan, O'Neill-Rood, & Florence, 1997), with the reported history of lower academic functioning among some minority students (Dozier & Barnes, 1997), as well as the over-identification of ethnic minorities in special education in the Unites States (Donovan & Cross, 2002; Elliot & Fuchs, 1997; Scott, Boynton-Hauerwas, & Brown, 2014). Research has shown that ethnic minorities obtain significantly lower scores on standardized achievement tests (Adkins, 2013; Sattler, 2008). A major factor considered to contribute to this trend has been test content bias (Bell, Lentz, & Graden, 1992); past research has identified a substantial number of standardized tests as assessing content that did not match classroom curriculum content (Bell, Lentz, & Graden, 1992; Good & Salvia, 1988). Adkins (2013) investigated the potential presence of bias in ORF probes as it pertained to their predictive relationship with computer-based standardized state testing. Results of that study detected the presence of racial predictive bias among ORF probes in the prediction of reading comprehension testing outcomes. Conversely, there is also literature that supports CBM within a culturally responsive RTI system as being a potentially valuable tool to utilize in order to generate data that is more in step with students' acquisition of

current curricula regardless of demographic distinctions such as gender and ethnicity (Hernández-Finch, 2012; Hosp & Madyun, 2007; Klingner & Edwards, 2006; Stecker, Fuchs, & Fuchs, 2005; VanDerHeyden, Witt, & Gilbertson, 2007). In a 2012 review of studies addressing RTI's effect on culturally and linguistically diverse students, Hernández-Finch concluded that although more research is needed to strengthen the research base on culturally responsive RTI, there are methods of promise emerging, such as appropriately researched CBM tools that are more aligned to local norms. More relevant to the current study, VanDerHeyden and colleagues (2007) examined the technical merits of the System to Enhance Educational Performance (STEEP) CBMs for both reading and math on the identification of students for special education across ethnicity and gender in a Southern Arizona school district. They analyzed implementation of STEEP in five of the district schools across two successive years and found that ethnically diverse students were not disproportionately identified for special education when compared to school baseline data. Further, they found no statistically significant gender differences with regard to special education identification.

The research literature addressing the presence of bias in reading CBM has had widely varying results (Hosp, Hosp, & Dole, 2011). However, there remains the need for further research as it pertains to ethnicity, CBM-M (Scott et al., 2014), and standardized state testing in math. In considering the current state of Response to Intervention and culturally diverse students, Hernández-Finch (2012) called for further research beyond the subject of reading to be conducted more frequently. Furthermore, Batsche (2007) spoke to the importance of disaggregating data in order to properly investigate different

predictable outcomes in educational research based on groupings, such as ethnicity, and other categories that are featured in No Child Left Behind.

The High Value of Math Literacy

The assessment of mathematics proficiency is especially important because an increasing amount of research indicates that mathematics is a core cognitive competency that should rank at least as highly as reading in level of importance (Clements & Sarama, 2008), as basic mathematics proficiency has proved to be integral to many daily life skills (e.g., consumer behaviors, household budgeting, and technical work demands) and has become a benchmark for obtaining a high school diploma (Minskoff & Allsopp, 2003). In a 2004 study, Duncan, Claessens, and Engel were able to show positive correlations not only between preschool reading skills and later elementary reading ability, but also between preschool mathematics skills and later math ability. More importantly though, they were able to demonstrate that while early reading skills only predicted later reading, early math skills were predictive of both later math *and* later reading abilities (Clements & Sarama, 2008). Further, research has also indicated that poor development of math skills in the primary grades is predictive of significant math difficulties in secondary school, and that this contributes to an increased risk of negative outcomes in adulthood (Delazer, Girelli, Grana, & Domahs, 2003; Jordan, Kaplan, Locuniak, & Ramineni, 2007; Mazzocco & Thompson, 2005). Results such as these are congruent with Fogen and colleagues (2007), who noted that the scope and sequence of mathematics curricula are more extensive and complex when compared to reading, requiring the ongoing mastery of key competencies (e.g., numerical operations, geometry, measurement, probability, algebraic functions, etc.) within and across all grade levels (Fogen, et al.).

Taking into consideration the increasing technological demands of the 21st Century, attaining competency in mathematics is more critical than ever, and has serious impact potential well beyond the classroom (Baglici et al., 2010; Mazzocco & Thompson, 2005). Today, there is an ever-widening range of occupations requiring some level of math literacy, which has been shown to increase the likelihood of successful employment (Mazzocco & Thompson, 2005; Patton, Cronin, Bassett, & Koppel, 1997; Saffer, 1999). Moreover, the National Science Board (2003) has reported math and science proficiency as necessary skills in careers that currently have the highest rate of growth.

In a 2000 review of national math proficiency test scores, the National Center for Educational Statistics (NCES) found that just 22% of U.S. fourth-grade students scored at or above the proficient level (Manzo & Galley, 2003). This figure went up to 31% in 2003, and jumped to 36% in 2005. In the most recent report, an estimated 40% of fourth-graders scored at or above the proficient level (NCES, 2005; 2013). Although this trend is promising and demonstrates a level of consistent improvement, American student progress in mathematics remains sluggish compared to the rest of the industrialized world. These results still leave much to be desired, and poor mathematics achievement remains a national concern (Clarke et al., 2011; Jordan et al., 2007; National Mathematics Advisory Panel, 2008). Recent reports from the Trends in International Mathematics and Science Study (TIMSS, 2012) and the tri-annual Program for International Student Assessment (PISA), released by the Organization for Economic Cooperation and Development (OECD, 2010) illustrate this concern at the international level.

The TIMSS focuses on mathematics achievement among fourth- and eighth-grade students from participating countries (Mullis, Martin, Foy, & Arora, 2012). Overall, U.S.

fourth-grade students ranked 11th out of 50 participating nations while U.S. eighth-graders ranked 9th out of 42 nations (TIMSS, 2012). The PISA (OECD, 2010), on the other hand, assesses the state of students' acquired knowledge in the areas of reading, mathematics, and science as they near the end of compulsory schooling; in 2009 this program analyzed the results of an international sample of 15-year-old students from 64 countries. While U.S. students performed within the average range in reading and science, they were below average in math literacy, ranking just 31st in mathematics performance. On a scale of six possible levels of mathematics proficiency, with level six being the highest, only 27% of U.S. students scored at or above level four (Fleischman, Hopstock, Pelczar, & Shelley, 2010; OECD, 2010), which correlates to students being capable of completing higher order tasks such as carrying out sequential processes, as well as problem solving using visual and spatial reasoning in novel situations. This stands in contrast to the reported 32% of students from the other OECD countries that scored at or above level four (Fleischman et al., 2010, p. 20). Following the release of the 2010 PISA report, U.S. Secretary of Education Arne Duncan commented that, "being average in reading and science—and below average in math—is not nearly good enough in a knowledge economy where scientific and technological literacy is so central to sustaining innovation and international competitiveness" (United States Department of Education, 2010, p. 1). In another major study, Peterson, Woessmann, Hanushek, and Lastra-Anadón (2011) of the Harvard Kennedy School analyzed international math performance, but with an additional state by state breakdown. They found that Arizona students only scored at 26.3 percent proficiency, below the U.S. average of 32.2 percent.

In a recent report submitted to congress, Schacht (2009) outlined the need to further bolster advancements in areas such as math and science in order for the U.S. to remain globally competitive. In fact, it has been estimated that technological progress is responsible for up to one-half of the economic growth in the U.S. and that failing to further cultivate this progress would contribute to increased competitive pressures in the international marketplace (Schacht, 2009). Peterson and colleagues (2011) calculated that if current trends persist it could cost the U.S. up to 75 trillion dollars over the next 80 years. As such, the proper development of technological prowess necessitates a sharp focus on what is, arguably, a key source of human intellectual capital at the primary and secondary school levels, and this is underlined by a clear need for continued improvement of student math performance in the United States.

To this end, in addition to mandates such as NCLB (2002) the Bush administration also created the National Mathematics Advisory Panel (NMAP, 2006) for the sole purpose of making expert, research-based recommendations aimed at overhauling the current delivery system in mathematics education so that the U.S. will not, "…relinquish its leadership in the 21st Century" (NMAP, 2008, p. xi). However, they ultimately concluded that, "international and domestic comparisons show that American students have not been succeeding in the mathematical part of their education at anything like a level expected of an international leader" (NMAP, 2008, p. xii). Of the many recommendations that they submitted in their final report, the NMAP endorsed the regular use of reliable and valid CBMs to monitor math achievement, particularly in the primary grades.

Study Purpose

There is evidence that CBMs are reliable and efficient tools with which to assess key academic skill areas (VanDerHeyden et al., 2007), thus allowing educators to closely monitor student progress and implement interventions when needed. Further, CBMs are less expensive, can be administered more frequently than annual standardized tests, and can be directly related to school and district curriculum (Elliott, et al., 2007); this is critical in order to help close the gap between actual and expected math literacy when needed, which in turn, would help to improve overall student performance on annual state standardized testing. More evidence is needed regarding the concurrent and predictive validity of CBMs in comparison to standardized achievement testing. The purpose of this study is to expand the research base on the validity of math CBM assessments by investigating universal screening measures of computational fluency and how well they predict performance on state standardized tests of academic achievement in Arizona, as well as to investigate whether there are any significant differential effects for ethnic or gender differences and how the screening time of year relates to student success on high stakes tests.

Research Questions and Hypotheses

The research questions and hypotheses for this study are as follows:

Research Question 1: What is the relationship between general CBM-M computational math fluency screening scores and general performance on standardized mathematics tests given in Arizona?

Hypothesis 1a: Following from the available literature, it is hypothesized that there will be a significant correlation with moderate to strong effect sizes between general

CBM-M computational math fluency screening scores and general performance on the math standard score from Arizona's Instrument to Measure Standards.

Hypothesis 1b: It is further hypothesized that there will be a significant correlation with moderate to strong effect sizes between general CBM-M computational math fluency screening scores and general performance on the Stanford-10 mathematics scores.

Research Question 2: What is the relationship between CBM-M computational math fluency screening scores and composite skill areas assessed by Arizona's state math measure including: (a) Number and Operations; (b) Data Analysis, Probability, and Discrete Mathematics; (c) Patterns, Algebra, and Functions; (d) Geometry and Measurement; and (e) Structure and Logic?

Hypothesis 2a: Following from the available literature, it is expected that there will be a moderate relationship between CBM-M computational math fluency screening scores and each of the composite skill areas assessed on the AIMS test math portion.

Hypothesis 2b: It is further hypothesized that the third grade CBM-M screening scores will have the strongest predictive validity and that the Number and Operations strand will, likewise, show the strongest strength of relationship compared to the other strands.

Research Question 3: What is the relationship between general CBM-M computational math fluency screening scores and general performance on standardized Arizona state mathematics tests when students are disaggregated by gender?

Hypothesis 3: Following from the available literature, it is hypothesized that there will not be significant differences in the observed predictive validity of general CBM-M computational math fluency screening scores and general performance on standardized Arizona state mathematics tests when disaggregating by gender across grade levels.

Research Question 4: What is the relationship between general CBM-M computational math fluency screening scores and general performance on standardized Arizona state mathematics tests when students are disaggregated by ethnicity?

Hypothesis 4: Following from the available literature, it is hypothesized that there will not be significant differences in the observed predictive validity of general CBM-M computational math fluency screening scores and general performance on standardized Arizona state mathematics tests when disaggregating by ethnicity.

Research Question 5: How does universal math screening time of year relate to student success on high stakes tests?

Hypothesis 5a: There will be a stronger relationship between CBM-M computational math fluency scores and performance on standardized Arizona state mathematics tests when analyzing the winter and spring screening data per grade level and a weaker relationship with the fall screenings.

Hypothesis 5b: It is further expected that there will be a significantly stronger relationship between third grade spring CBM-M math fluency scores and Arizona state mathematics test scores than with either the fall or winter of the third grade.

Chapter 2

METHOD

Participants

The participants in this study were 410 students selected from 46 third-grade classrooms out of six elementary schools in a large Arizona district. The chosen elementary schools were the only schools in the participating district with the data necessary to conduct the present study. A seventh elementary school had to be removed from consideration after it was discovered that the probes had been scored incorrectly (using total answers correct, rather than total digits correct). The focus of the current study was placed on typical students in the interest of achieving greater generalizability of the results; consequently, all students from the sample designated as being in special education were also excluded from the study because the archival data did not provide information as to specific category of special education eligibility. Thus, a total of 52 cases were removed from the study due to special education eligibility, bringing the final sample set to 358 students. The sample was 52.8% male and 47.2% female. The ethnic background of participants was 70.1% White, 19.3% Hispanic, 3.1% Black, 1.1% American Indian or Alaskan Native, 3.6% Asian, .6% Native Hawaiian or Other Pacific Islander, and 2.2% were listed as Two or More Races. As reported by the participating school district, the 2013-2014 district-wide gender and racial breakdown was 50.7% male, 49.3% female, 67.4% White, 20% Hispanic, 3.8% Black, 2.4% Native American or Alaskan Native, 5.1% Asian, .23% Native Hawaiian or Other Pacific Islander, and 1.2% were listed as Two or More Races. Tables 1 and 2 show the relevant participant and participating school district demographic information.

According to the American Psychological Association (APA, 2003), race is defined as a socially constructed category in which specific identification is determined through stereotypical physical characteristics such as skin color or hair type. In contrast, ethnicity is referred to as "the acceptance of the group mores and practices of one's culture of origin and the concomitant sense of belonging" (APA, 2003, p. 9). Thus, for the purposes of the present study it was determined that the term "ethnicity" was more meaningful with regard to the essential characteristics associated with student cultural identity.

Table 1

Study Participant Demographic Variables

Demographic Variable	*n*	Percent
Gender		
Female	169	47.2
Male	189	52.8
Ethnicity		
White	251	70.1
Hispanic	69	19.3
Black/African American	11	3.1
American Indian/Alaskan Native	4	1.1
Asian	13	3.6
Native Hawaiian/Other Pacific Islander	2	0.6
Two or more races	8	2.2

N = 358

Table 2

Participating School District Demographic Variables

Demographic Variable	*n*	Percent
Gender		
Female	12397	49.3
Male	12774	50.7
Ethnicity		
White	16958	67.4
Hispanic	5022	20.0
Black/African American	951	3.8
American Indian/Alaskan Native	594	2.4

Asian	1290	5.1
Native Hawaiian/Other Pacific Islander	58	0.23
Two or more races	298	1.2

N = 25171

Instruments

System to Enhance Educational Performance (STEEP). The STEEP (Witt, 2002) is a research-based CBM system of academic skill probes designed for universal screening, intervention, and progress monitoring. For the purposes of this study, the STEEP math probes consisted of 1st through 3rd grade-level measures ranging from 40 to 49 items that are focused on fluency in single and double digit addition, subtraction or a combination thereof depending on the grade level. These universal screening measures are two minute long individually administered probes that are scored according to total place-value digits correct.

The instructional standard used for interpreting performance on the math probes follows from Deno and Mirkin (1977) at 20–40 digits correct for grades 1–3 as indicating performance relative to same-aged peers being at or above the 16th percentile (VanDerHeyden et al., 2007). Categorically, results fall into one of three rankings: Frustrational, Instructional, and Mastery. In practice, these benchmark criterion scores indicate whether a student has a substantial need for math supports, may need basic numeration and operations support or is experiencing no difficulties with basic math fluency, respectively. Children who score within the Frustrational range are rescreened with what is referred to as a *can't do/won't do* assessment, where some type of reinforcer is provided in an effort to determine whether a student's below benchmark score was the result of low motivation or a legitimate skill deficit. The student is administered the same math fluency probe that was used in the original screening, except this time the examiner

tells them that if they can improve upon their previous score they will be given a reward (Witt & VanDerHeyeden, 2007). As a general practice, students in the current study were administered the *can't do/won't do* trial when necessary and the higher score was recorded in the data.

The reliability and validity of math CBM data is very well supported in the applied research (see Burns, 2004; VanDerHeyden & Burns, 2005). In their 2005 study on the use of STEEP CBM math data to guide primary level mathematics instruction, VanDerHeyden and Burns found Cohen's *d* coefficients in the .47 to .92 range, indicating moderate to strong effect sizes (VanDerHeyden & Burns, 2005). Table 3 shows the math computational fluency end of the year benchmark scores for elementary students.

Table 3

STEEP Math Computational Fluency End of the Year Digits Correct Benchmarks for Elementary School

	Grade Level	
Proficiency Level	First-Third	Fourth-Sixth
Frustrational	0-19	0-39
Instructional	20-39	40-79
Mastery	40+	80+

Stanford Test of Achievement, Tenth Edition (Stanford-10). The Stanford-10 (Harcourt Educational Measurement, 2003) is a research-based, nationally norm-referenced, multiple choice achievement test first published in 1926. It provides information on student performance in core academic areas such as reading, language, and mathematics. This test (Harcourt Educational Measurement, 2003) yields several different types of scores, including raw scores, percentiles, scaled stanine scores, and

grade equivalent scores. As with any standardized assessment, the scaled scores and percentile rankings are of most interest because these allow educators and researchers to compare student performance against the set of same aged peers who took the test at the same time. Stanford-10 math items only yield a single standard score and are not divided into subskill areas as with the AIMS.

The individual questions comprising the Stanford-10 (Harcourt Educational Measurement, 2003) were drafted based on national and state instructional standards, as well as content-specific classroom curricula. Experts then generated the assessment blueprints from which testing professionals and practicing teachers would create the complete test items. A test blueprint assigns the percentage of questions that should measure each test concept. Finally, measurement specialists, content experts, and testing editors screened and finalized the test items (Harcourt Educational Measurement).

The Stanford-10 has strong reliability evidence supporting it. For instance, the Reading section of the Stanford-10 received an alpha reliability rating of .87, the Math section .80-.87, and the Language section .78-.84. Past research has also reported reliability coefficients (K-R20 and alternate form) in the .80-.90 range for the total math cluster and math multiple-choice portion of the test (Burns, VanDerHeyden, & Jiban, 2006). Although the Stanford-10 was designed for students from kindergarten through the twelfth grade, the state of Arizona only utilizes stand-alone administrations of the test for grades two and nine. However, the items from the Stanford-10 are embedded within the third, fourth, and eighth-grade administrations of AIMS tests to provide a dual purpose assessment featuring both norm-referenced and criterion-referenced components (Arizona Department of Education, 2013). The math portion of the Stanford-10 is

comprised of 25 items; however, while the results are reported as separate standard scores, 15 out of the 25 Stanford-10 items map on to the five main content strand areas of the AIMS test and contribute to both sets of scores reported by the Arizona Department of Education. Stanford-10 norm-referenced scores from the current study were obtained using the 2007 spring norms (Arizona Department of Education, 2013).

Arizona's Instrument to Measure Standards (AIMS). In the state of Arizona, the data obtained from the Stanford-10 items is used in conjunction with the AIMS test as a Dual Purpose Assessment (DPA; Arizona Department of Education, 2014) to more thoroughly measure levels of pupil achievement for statewide accountability purposes. The AIMS is an assessment designed to measure student proficiency in the areas of reading, writing, math, and science that is required by state and federal law (Arizona Department of Education, 2014). It is administered in grades 3 through 8, as well as high school, and is designed to measure performance on content standards, which were adopted in March 2003 for reading, June 2008 for mathematics, June 2004 for writing, and March 2005 for science (Arizona Department of Education, 2013). Reading and math are assessed on all AIMS administrations through the 12th grade. All skill areas, save written expression, are administered in a multiple-choice format. Student performance is then scored at one of four different proficiency levels for each content area, "Falls Far Below Standards", "Approaches Standards", "Meets Standards", or "Exceeds Standards". In addition, students obtain standardized scores as well as percentile rankings in relation to national norms. High school students must pass the AIMS as a requirement for graduation (Arizona Department of Education, 2012).

The most recent technical report for the AIMS test (Arizona Department of Education, 2013) indicates that the AIMS Reading/Language and Mathematics tests for the third, fourth, and eighth grades are used as dual-purpose assessments, meaning that a combination of both criterion-referenced scores based on state standards *and* nationally norm-referenced scores are generated based on student performance. Arizona teachers, curriculum specialists, and administrators made contributions to both test item development and the interpretation of results. Overall internal consistency reliability estimates for the spring 2013 AIMS assessment were calculated to be within the .82 to .93 range for the criterion-referenced items and within the .59 to .85 range for the norm-referenced items (Arizona Department of Education, 2013).

Including the Stanford-10 items, the third-grade AIMS math portion was made up of 76 operational items, of which 66 were divided among the five main content strands comprising Arizona's academic standards for mathematics, solely contributing to the AIMS standard score. The remaining 10 items contributed solely to the Stanford-10 standard score, thus allowing for a dual purpose assessment. According to the state item map of specifications, the Number and Operations strand (28 items) features the highest number of loaded test items in the earlier grades, while Data Analysis, Probability, and Discrete Mathematics (8 items), Patterns, Algebra, and Functions (11items), Geometry and Measurement (12 items), and Structure and Logic (7 items) feature fewer loaded items in the earlier grades that subsequently increase in the upper grades (Arizona Department of Education, 2013; D'Agostino, 2010).

Procedure

Data was gathered on the math computational fluency scores for study participants' first, second, and third grade school years in order to assess the strength of relationship between CBM-M scores and the selected assessment results. The Arizona State University Institutional Review Board (IRB; see Appendix B) approved this study as exempt. The participating school district then agreed to allow access to its archival data, which included STEEP-CBM-M universal screening, Stanford-10, and AIMS scores from 2010 to 2013. This data was collected from two separate online, internal district databases; one is known as *Arizona RTI* (AZRTI), and it supplied the math computational fluency screening scores from each school. The other database is called *Datacentral*. This database included the demographic information, Arizona's Instrument to Measure Standards Dual Purpose Assessment (AIMS DPA) scores, as well as Stanford-10 mathematics scores for the sample set. Information was gathered on the mathematics portion of the Stanford-10 and AIMS DPA performance at the end of the students' third grade year, along with relevant demographic variables. The AIMS DPA was administered during the month of April in the participants' third grade year.

Ideally, participants would have received CBMs of math computational fluency in the fall, winter, and spring. However, review of the data sources revealed that the participating schools had different math screening/collection procedures, which unavoidably led to some missing data points. For instance, one school did not administer a second grade winter screening. Fall screenings would have been completed approximately two to four weeks after the start of the school year, while the winter screenings were completed in December/January. The spring screenings were completed

in April/May approximately two to four weeks before the end of the school year. Math fluency screening probes were given and scored by school personnel trained in their administration.

To be included in the analysis, participants were required to have at least one first, one second, and one third-grade universal screening probe, as well as mathematics component scores from both the SAT-10 and AIMS DPA assessments administered in the third grade year. Missing data points on the math computational fluency probes were corrected using the regression formula, linear trend at point, which replaces missing values with the predicted value for that point. In total, there were 3,222 data points, out of which 28 percent were missing. Participants who had missing data were retained to ensure an accurate data sample, as the removal of non-random participants can cause distribution skewness (Tabachnick & Fidell, 1996; Devena, 2013).

Chapter 3

RESULTS

Power Analysis

In order to determine the achieved power for this study based on a set sample size, the G*Power software package (Faul, Erdfelder, Buckner, & Lang, 2009) was employed to perform post hoc power analyses for each type of test being utilized following from Cohen (1992), who outlined four essential parameters for statistical analysis including statistical power ($1 - \beta$), significance criterion (α), sample size (n), and effect size (q). Typically, three of these are known and subsequently used to derive the fourth (Cohen, 1988, 1992). Each statistical test in the present study retained an individual alpha level of .05 because each research question was generated based on meaningful interpretations of the data. The power analyses were performed post hoc due to unexpected limitations in the data sources, which ended data collection at the current sample set size before a priori analyses were completed.

The power analyses illustrated in Figure 1 were conducted across a range of possible effect sizes for research questions numbered one through four. Results of an exact model, two-tailed test for a given sample size (N = 358) where the null correlation was set at zero showed adequate power achieved from the lower critical r value of .10 for research questions one and two such that, when employing effect size conventions found in the literature (Cohen, 1969; Faul et al., 2009), to detect a relatively small effect size (q = .10) the given sample revealed a statistical power of .47, and a power of 1.0 for detecting both moderate (q = .30) and large (q = .50) effect sizes. Research questions

three and four required a *z*-test model, two-tailed test for two independent Pearson *r*'s from the same sample set split by gender (Female $N = 169$; Male $N = 189$) and then ethnicity (White Students $N = 251$; Combined Non-White Students $N = 107$) that likewise showed adequate power achieved from critical *z* value 1.96. Thus, the given sample disaggregated by gender revealed a statistical power of .15 for detecting a relatively small effect ($q = .10$), a power of .80 for detecting a moderate effect size ($q = .30$), while power exceeded .99 for the detection of a large effect size ($q = .50$); in much the same way, when the sample was disaggregated by ethnicity the analysis revealed statistical powers of .14, .73, and .99, respectively. Subsequently, the sample size of 358 was found to provide sufficient power for the current study because a majority of the effect sizes found in the CBM literature range from moderate to large.

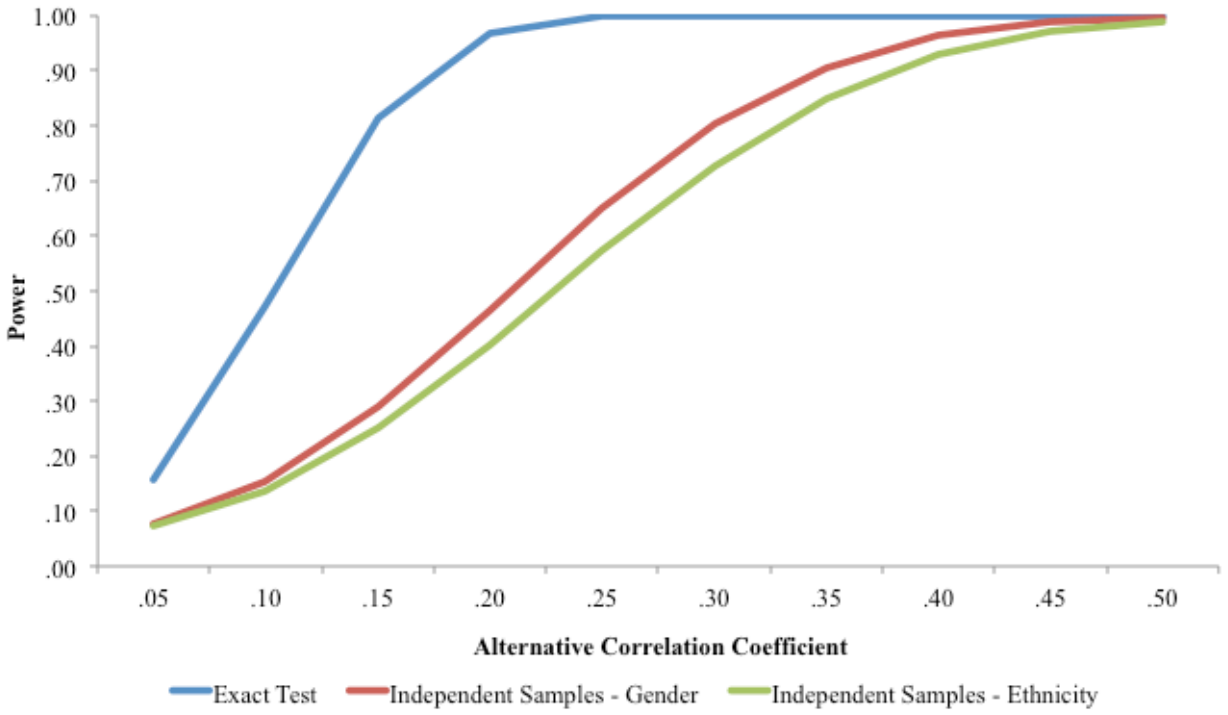


Figure 1. Post Hoc Power Analyses for Given Sample Size ($N = 358$) and Sample Size Split by Gender and Ethnicity: Research Questions 1-4.

To address research question number five, a *z*-model, post hoc, two-tailed test for the correlation between two dependent Pearson *r*'s with a common index was conducted. When the observed correlations were entered into the analysis, achieved power for the given sample size of 358 was .36. This relatively lower power measure was likely due to a negligible difference between the observed correlations and the fact that the measures being compared were highly correlated with one another.

Sample Characteristics

Means and standard deviations were calculated for select demographic variables relevant to this study. Table 4 shows the means and standard deviations for grade 1, grade 2, and grade 3 math computational fluency scores. On average, scores were similar for females and males across the three grade levels. Scores were more variable across ethnicity, with Asian and White scores being the highest across the first two years, while Asian and African American scores were the highest for the third grade. Hispanic and African American student scores were lowest in grade 1; students with two or more races scored the lowest in grade 2, and American Indian/Alaskan Native, Hispanic, and students with two or more races had the lowest scores for grade 3.

Table 4

Means and Standard Deviations of Math Computational Fluency Scores

	Grade 1		Grade 2		Grade 3	
Demographic Variable	*M*	SD	*M*	SD	*M*	SD
Gender						
Female	26.19	7.27	42.05	12.29	22.10	12.13
Male	26.50	7.41	46.09	16.04	22.94	12.65
Ethnicity						
White	27.17	7.01	46.25	14.23	23.70	11.95
Combined Non-White Ethnicity	24.44	7.76	39.33	14.07	19.84	13.05
Hispanic	22.73	5.41	37.23	12.86	16.99	10.06

Black/African American	22.05	5.52	39.21	14.15	25.89	17.64
American Indian/Alaskan Native	25.72	6.36	40.86	10.97	12.19	10.38
Asian	34.11	12.74	52.79	17.59	34.32	16.54
Native Hawaiian/Other Pacific Islander	24.34	0.94	41.67	5.18	16.91	6.00
Two or more races	26.05	8.48	34.39	10.37	17.04	7.61
Total	26.35	7.34	44.18	14.52	22.54	12.40

N = 358; Highest Possible Probe Scores: 1st Grade: 56; 2nd Grade: 91; 3rd Grade: 86

Table 5 shows the means and standard deviations of Stanford-10 mathematics component scores and AIMS DPA mathematics scaled score by select demographic variables. Across demographic variables, student performance on the SAT-10 math component scores and AIMS DPA math scaled scores displayed a pattern that was generally consistent with the math computational fluency scores presented above. Females and males achieved similar scores across the two measures; Asian and White students obtained the highest scores on both measures. Hispanic students had the lowest average score on the SAT-10 while American Indian/Alaskan Native students had the lowest average score in the AIMS DPA math portion. The combined Non-White ethnicity participants had to be collapsed into one category since there were too few students of each group with which to complete analyses of the individual ethnic categories. While the Asian group did show higher math scores, on average, than the other groups, the total number of Asian participants (*n* =13) made it necessary to combine them with the other Non-White participants. However, when collapsed into one category, the combined Non-White ethnicity group still averaged lower scores than White students, overall.

Table 5

Means and Standard Deviations of High Stakes Test Scores According to Select Demographic Variables

	Stanford-10 Math SS		AIMS-DPA Math SS	
Demographic Variable	*M*	SD	*M*	SD
Gender				
Female	647.59	42.11	394.59	47.78
Male	651.66	40.65	401.16	48.82
Ethnicity				
White	658.71	37.45	407.12	46.10
Combined Non-White Ethnicity	628.70	42.55	376.79	47.09
Hispanic	622.49	39.73	369.45	43.40
Black/African American	627.55	39.33	375.36	35.47
American Indian/Alaskan Native	628.25	49.07	338.50	30.90
Asian	656.92	50.81	420.08	58.73
Native Hawaiian/Other Pacific Islander	625.50	33.23	386.50	40.31

Note. AIMS DPA = Arizona Instrument to Measure Standards Dual Purpose Assessment; SS = Scale Score; SAT-10 = Stanford Achievement Test-10th Edition DPA Items. $N = 358$

Data Analysis

Pearson product-moment correlations were calculated to determine the degree to which select predictive variables and the corresponding criterion variables were linearly related in the study sample (Green & Salkind, 2011). Often in the research literature, the statistical significance of a coefficient is also interpreted as an effect size; that is, significant outcomes being construed as big effects and non-significant outcomes as being small, unimportant effects. This approach is problematic, however, because sometimes effects that are the same size can come out as highly significant, and at other times non-significant. Meanwhile, trivial effects can sometimes come out as highly significant, while important effects can register as non-significant. This occurs due to the fact that tests of statistical significance tend to confound the magnitude of impact and the size of the sample, which are two independent pieces of information. Thus, statistical significance, in and of itself, provides very little information as to the practical

significance or the relative impact of an effect size and so should never be used as the only measure of how much a relationship "matters" (Valentine & Cooper, 2003). As a result, R^2 coefficients were also calculated and interpreted in order to determine the overall effect size, or strength of relationship between math computational fluency screening scores and standardized state achievement math scores, as well as selected individual composite strands (Valentine & Cooper, 2003).

It was necessary to consider the two assumptions that underlie the Pearson correlation significance test; (a) that the variables are bivariately normally distributed, and (b) that the cases are a random sampling from the population and scores for one case are independent of the scores on the other cases (Green & Salkind, 2011). Scatterplots for the variables of comparison were generated in order to rule out non-linear relationships that might attenuate *r* (Cohen, Cohen, West, & Aiken, 2003). A majority of the scatterplots showed a generally positive linear relationship between variables, save for research question two (See Appendix A for several examples). The statistical program used for this study calculated the strand criterion scores for research question two as discrete rather than continuous quantitative variables due to limited item sets within the individual strands as compared to the CBM-Ms, indicating a violation of the normal distribution assumption. To address this in the analysis, correlations for research question one were reported with Spearman's rank coefficients in place of Pearson's *r*'s in order to necessarily treat the strand scores as ordinal-level rather than interval-level data.

Cohen (1988) suggested benchmarks for interpreting the R^2 effect size in the behavioral sciences (see also Valentine & Cooper, 2003) whereby an R^2 of .01 is

considered to represent a small effect, an R^2 of .09 is considered to represent a medium-sized effect, while large effects are generally considered equal or larger to an R^2 of .25.

When the comparisons were disaggregated by ethnicity and gender, respectively, Fisher's *r* to *z* transformation (Fisher, 1921) was applied in order to allow for comparison of the actual correlation values; this was due to the fact that *r*-values are not normally distributed (Howell, 2013). Steiger's *z*-test (Steiger, 1980) was used following Fisher's *r* to *z* transformation to determine the difference between the independent correlations (Weaver & Wuensch, 2013). Fisher's *r* to *z* transformation was also used when comparing the strength of relationship between the screening times of year for the third grade and standardized test results. Following this, Hotelling's *t*-test (Hotelling, 1931) was used to assess the difference between dependent correlations with a common index measure (Weaver & Wuensch, 2013).

First Research Question

The first research question addressed the relationship between general CBM-M computational math fluency screening scores and total performance on standardized Arizona state mathematics tests.

Pearson product-moment correlations were calculated to investigate the relationship between CBM-M math computational fluency scores and AIMS and Stanford-10 math composite scores calculated from the third grade DPA administration, as well as the intercorrelation between grade-level fluency performances in order to determine whether observed differences were due to chance. As a measure of effect size, R^2 coefficients were then calculated to assess the strength of relationship between the AIMS and Stanford-10 math portions and math computational fluency scores across

screenings and grade level. The results of the correlation analysis are provided in Table 6. As expected, all grade-level correlations were significant after controlling for Type 1 Error at .05.

Table 6

Means, Standard Deviations, and Correlations between Arizona Instrument to Measure Standards Dual Purpose Assessment Math Scores with Math Computational Fluency Measures

Variable	M	SD	1	2	3	4	5	6	7	8	9
AIMS Math	398.06	48.37	.35*	.35*	.23*	.56*	.47*	.47*	.49*	.43*	.55*
1. F Grade 1	13.96	9.17		.71*	.49*	.42*	.27*	.23*	.45*	.40*	.35*
2. W Grade 1	28.36	9.04			.59*	.49*	.36*	.38*	.47*	.46*	.40*
3. S Grade 1	36.74	7.44				.42*	.17*	.32*	.34*	.44*	.31*
4. F Grade 2	32.67	15.46					.65*	.63*	.64*	.49*	.60*
5. W Grade 2	47.48	16.12						.69*	.51*	.35*	.56*
6. S Grade 2	52.40	18.03							.55*	.48*	.61*
7. F Grade 3	16.54	13.08								.66*	.72*
8. W Grade 3	22.33	12.85									.68*
9. S Grade 3	28.75	15.78									

Note. AIMS = Arizona Instrument to Measure Standards Dual Purpose Assessment; F = Fall: W = Winter; S = Spring. * $p < .05$. $N = 358$

Table 7 shows the resulting R^2 coefficients derived from the correlations listed above. In general, as students moved through each grade level screening the coefficient and effect size magnitudes increased, with similar results occurring among the second and third grade CBM-M administrations. The first grade fall and winter both had the same moderate effect size, $r(356) = .35$, 90% CI [.27, .42], $p < .05$, with 12% of the variance in AIMS math being accounted for by its linear relationship with each corresponding CBM-M performance. The strongest overall association was observed with the second grade fall CBM-M administration, $r(356) = .56$, 90% CI [.50, .62], $p < .05$, with 31% of the variance in AIMS math being accounted for by its linear relationship with math computational fluency. The third grade spring CBM-M administration also

saw a strong association, $r(356) = .55$, 90% CI [.49, .61], $p < .05$, with 30% of the variance in AIMS math being accounted for by computational fluency. The weakest association occurred in the first grade spring, where 5% of the variance in the AIMS math was accounted for by CBM-M performance, $r(356) = .23$, 90% CI [.15, .31], $p < .05$.

Table 7

Means, Standard Deviations, and R^2 between Arizona Instrument to Measure Standards Dual Purpose Assessment Math Scores with Math Computational Fluency Measures

Variable	M	SD	1	2	3	4	5	6	7	8	9
AIMS Math	398.06	48.37	.12	.12	.05	.31	.23	.22	.24	.18	.30
1. F Grade 1	13.96	9.17		.50	.24	.18	.07	.05	.20	.16	.12
2. W Grade 1	28.36	9.04			.35	.24	.13	.14	.22	.21	.16
3. S Grade 1	36.74	7.44				.18	.03	.10	.12	.19	.10
4. F Grade 2	32.67	15.46					.42	.40	.41	.24	.36
5. W Grade 2	47.48	16.12						.48	.26	.12	.31
6. S Grade 2	52.40	18.03							.30	.23	.37
7. F Grade 3	16.54	13.08								.44	.52
8. W Grade 3	22.33	12.85									.46
9. S Grade 3	28.75	15.78									

N = 358

Table 8 shows the correlations between the Stanford-10 math scores and math computational fluency scores across all three grade level screenings. Similar to the AIMS results, all grade-level correlations were significant after controlling for Type 1 Error at .05.

Table 8

Means, Standard Deviations, and Correlations between Stanford Achievement Test-10th Edition DPA Math Component Scores with Math Computational Fluency Measures

Variable	M	SD	1	2	3	4	5	6	7	8	9
SAT-10 Math	649.74	41.34	.30*	.28*	.19*	.51*	.43*	.45*	.43*	.35*	.48*
1. F Grade 1	13.96	9.17		.71*	.49*	.42*	.27*	.23*	.45*	.40*	.35*
2. W Grade 1	28.36	9.04			.59*	.49*	.36*	.38*	.47*	.46*	.40*
3. S Grade 1	36.74	7.44				.42*	.17*	.32*	.34*	.44*	.31*
4. F Grade 2	32.67	15.46					.65*	.63*	.64*	.49*	.60*
5. W Grade 2	47.48	16.12						.69*	.51*	.35*	.56*

6. S Grade 2	52.40	18.03	.55*	.48*	.61*
7. F Grade 3	16.54	13.08		.66*	.72*
8. W Grade 3	22.33	12.85			.68*
9. S Grade 3	28.75	15.78			

Note. Stanford Achievement Test-10th Edition DPA Standard Scores; F = Fall: W = Winter; S = Spring.
* $p < .05$. $N = 358$

Table 9 shows the R^2 coefficient for each of the SAT-10 correlations. Moderate to strong relationships can be noted across all grade-level CBM-M administrations. Consistent with AIMS results, the strongest associations occurred in the second grade fall screening, $r(356) = .51$, 90% CI [.44, .57], $p < .05$, with 26% of the variance in the SAT-10 math score being accounted for by CBMs of computational fluency, and third grade spring, $r(356) = .48$, 90% CI [.41, .54], $p < .05$, with 23% of variance accounted for. The weakest relationship was once again found to be the first grade spring administration, $r(356) = .19$, 90% CI [.11, .27], $p < .05$, with 4% of the variance in the SAT-10 math score being accounted for by the computational fluency CBM.

Table 9

Means, Standard Deviations, and R^2 between Stanford Achievement Test-10th Edition Math DPA Component Scores with Math Computational Fluency Measures

Variable	M	SD	1	2	3	4	5	6	7	8	9
SAT-10 Math	649.74	41.34	.09	.08	.04	.26	.18	.20	.18	.12	.23
1. F Grade 1	13.96	9.17		.50	.24	.18	.07	.05	.20	.16	.12
2. W Grade 1	28.36	9.04			.35	.24	.13	.14	.22	.21	.16
3. S Grade 1	36.74	7.44				.18	.03	.10	.12	.19	.10
4. F Grade 2	32.67	15.46					.42	.40	.41	.24	.36
5. W Grade 2	47.48	16.12						.48	.26	.12	.31
6. S Grade 2	52.40	18.03							.30	.23	.37
7. F Grade 3	16.54	13.08								.44	.52
8. W Grade 3	22.33	12.85									.46
9. S Grade 3	28.75	15.78									

$N = 358$

Second Research Question

The second research question addressed the relationship between CBM-M computational math fluency screening scores and composite skill areas assessed by Arizona's state math measures including: (a) Number and Operations; (b) Data Analysis, Probability, and Discrete Mathematics; (c) Patterns, Algebra, and Functions; (d) Geometry and Measurement; and (e) Structure and Logic. This analysis was specific to the AIMS because the Stanford-10 assessment only yields a single standard score and is not divided into math subskills areas.

Spearman's rank coefficient was calculated in place of Pearson's *r* for question one to assess the relationships between math computational fluency by grade level and standardized scores on the specific math skills measured by the AIMS DPA. The computational fluency scores across the three years were correlated with the individual AIMS DPA math score components. As indicated above, this was done because scatterplot analysis indicated a violation of the assumption of normalcy, which was primarily due to the fact that there are substantially fewer items making up each strand area relative to the item count on the math fluency measures. In other words, as a function of the test blueprint, the measurement scale underlying these variables was ordinal rather than interval due to the discrete nature of the strand scores and so an alternative analysis was necessary to make the correction (Green & Salkind, 2011). The probability of generating statistically significant test results increases as the number of tests increases (Type I Error), so Holm's Sequential Bonferroni Procedure was used to control for Type I Error across the multiple correlations (Abdi, 2010). The resulting

correlation coefficients and R^2 coefficients were assessed in order to determine the strength of relationship across comparisons.

A mean value was calculated for first, second, and third grade from the CBM-M values collected across the three years. The grade means were correlated with the AIMS DPA math score components in order to determine whether observed differences were due to chance. As a measure of effect size, R^2 coefficients were then calculated to assess the strength of relationships between math computational fluency by grade level and standardized scores on the specific math skills measured by the AIMS DPA. The results of the correlation analysis are provided in Table 10. Mean score analysis shows that participants' scores were highest for Number and Operations and lowest for Structure and Logic. As expected, all grade-level mean correlations were significant after controlling for Type 1 Error at .05. This accurately reflected the results from all of the individual correlations, which also showed a majority as being significant at the $P < .01$ level.

Table 10

Means, Standard Deviations, and Spearman's Correlations between Arizona Instrument to Measure Standards Dual Purpose Assessment Math Strand Scores and Mean Scores of Math Computational Fluency by Grade Level

			CBM-M Year		
AIMS DPA Math Score Component	M	SD	Grade 1	Grade 2	Grade 3
Number and Operations	21.01	4.98	.35*	.54*	.50*
Data Analysis, Probability, and Discrete Mathematics	5.97	1.72	.20*	.40*	.38*
Patterns, Algebra, and Functions	8.64	2.11	.27*	.47*	.48*
Geometry and Measurement	9.38	1.89	.25*	.44*	.41*
Structure and Logic	4.90	1.71	.26*	.43*	.42*

Note. AIMS DPA = Arizona Instrument to Measure Standards Dual Purpose Assessment; CBM-M = Curriculum Based Measurement in Math Fluency. * $p < .01$. $N = 358$

Table 11 shows the resulting R^2 coefficients derived from the listed correlations. In general, as students moved from first to second grade the coefficient and effect size magnitudes increased, while results between the second and third grade were found to be similar. The strongest overall association occurred in the second grade, $r(356) = .54$, 90% CI [.48, .60], $p < .01$, with 29% of the variance in AIMS Number and Operations being accounted for by its linear relationship with the corresponding CBM-M performance. The third grade also saw a strong association with the Number and Operations strand, $r(356) = .50$, 90% CI [.43, .56], $p < .01$, with 25% of variance accounted for by a linear relationship with computational fluency. The weakest associations consistently occurred in the first grade. For instance, only 4% of the variance in the Data Analysis, Probability, and Discrete Mathematics strand was accounted for by its relationship to CBM-M performance, $r(356) = .20$, 90% CI [.12, .28], $p < .01$.

Table 11

Means, Standard Deviations, and derived R^2 coefficients between Arizona Instrument to Measure Standards Dual Purpose Assessment Math Strand Scores and Mean Scores of Math Computational Fluency by Grade Level

			CBM-M Year		
AIMS DPA Math Score Component	M	SD	Grade 1	Grade 2	Grade 3
Number and Operations	21.01	4.98	.12	.29	.25
Data Analysis, Probability, and Discrete Mathematics	5.97	1.72	.04	.16	.14
Patterns, Algebra, and Functions	8.64	2.11	.07	.22	.23
Geometry and Measurement	9.38	1.89	.06	.19	.17
Structure and Logic	4.90	1.71	.07	.18	.18

$N = 358$

Third Research Question

The third research question addressed the relationship between general CBM-M computational math fluency screening scores and total performance on standardized Arizona state mathematics tests when students were disaggregated by gender.

Pearson product-moment correlations were calculated for each sample to assess the relationships between math computational fluency scores from the first through third-grade screenings and the standardized scores for AIMS DPA and Stanford-10 math tests. The alpha level for each test was set in an effort to maintain the error rate at .05. Fisher's *r* to *z* transformation was used to convert the correlation coefficients to *z* scores with a mean of zero and a standard deviation of one in order to allow for comparison of the actual correlation values. Steiger's *z*-test was then used to determine the difference between the two independent correlations. Table 12 shows the correlations between CBM-M computational fluency scores and the math scores from both standardized tests for both genders. The results show that when the sample was analyzed after being disaggregated by gender, most of the correlations between math screening performances and both sets of standardized test scores were found to be significant, which is consistent with the previous research.

Table 12

Correlations between Math Computational Fluency and High Stakes Test Scores by Gender

CBM-M Administration	M	SD	AIMS Math	SAT-10 Math
Female				
First Grade Fall	13.43	8.84	.31*	.27*
First Grade Winter	28.30	9.59	.32*	.25*
First Grade Spring	36.83	7.24	.20*	.21*
Second Grade Fall	29.62	12.58	.53*	.39*
Second Grade Winter	46.11	14.70	.46*	.38*

Second Grade Spring	50.43	16.80	.39*	.40*
Third Grade Fall	15.54	11.89	.48*	.40*
Third Grade Winter	22.19	13.18	.39*	.35*
Third Grade Spring	28.55	15.11	.49*	.47*
Male				
First Grade Fall	14.43	9.46	.38*	.33*
First Grade Winter	28.42	8.53	.38*	.31*
First Grade Spring	36.65	7.64	.26*	.17
Second Grade Fall	35.40	17.22	.59*	.52*
Second Grade Winter	48.70	17.23	.48*	.47*
Second Grade Spring	54.16	18.93	.54*	.48*
Third Grade Fall	17.44	14.04	.49*	.46*
Third Grade Winter	22.44	12.58	.47*	.35*
Third Grade Spring	28.94	16.39	.59*	.48*

Note. CBM-M = Curriculum Based Measurement in Math Fluency; AIMS DPA = Arizona Instrument to Measure Standards Dual Purpose Assessment; SAT-10 = Stanford Achievement Test-10th Edition DPA Items. * $p < .01$. Females $n = 169$, Males $n = 189$

The Fisher's *r* to *z* transformations of math fluency and standardized math test correlation scores by gender are given in Table 13. Steiger's *z*-test was then employed as an inferential measure to determine whether the value of the difference between the gender correlation coefficients was statistically significant. However, there were no significant differences between respective gender correlations across all grade level math screenings. Only the second grade spring Steiger's *z*-test approached significance ($p = .07$) for gender correlational differences.

Table 13

Results of Steiger's z-Test on Differences in Pearson r Coefficients for Gender.

CBM-M Administration	AIMS		SAT-10	
	Stieger's *Z*	*P* Value	Stieger's *Z*	*P* Value
First Grade Fall	0.74	.46	0.62	.54
First Grade Winter	0.64	.52	0.61	.54
First Grade Spring	0.59	.56	0.39	.70
Second Grade Fall	0.82	.41	1.54	.12
Second Grade Winter	0.24	.81	1.03	.30
Second Grade Spring	1.80	.07	0.93	.35
Third Grade Fall	0.12	.90	0.69	.49
Third Grade Winter	0.92	.36	0.00	1.00
Third Grade Spring	1.33	.18	0.12	.90

* Significant Difference: Two-Tailed $p < .025$

Fourth Research Question

The fourth research question addressed the relationship between general CBM-M computational math fluency screening scores and general performance on standardized Arizona state mathematics tests when students were disaggregated by ethnicity.

Due to limits in the available data, multiple comparisons across each of the ethnicity categories was not feasible due to an unbalanced allocation ratio in the power analysis. This would have required sample sizes for each sample beyond what the present data set could adequately satisfy. As a result, the data set was split in such way that all of the ethnic minority students were collapsed together into a single category, distinct from White students. After disaggregating the data set by the two ethnic categories, Pearson product-moment correlations were calculated for each sample to assess the relationships between math computational fluency scores from the first through the third-grade screening scores and the standardized scores for AIMS DPA and Stanford-10 math domains. The alpha level for the test was set in an effort to maintain the error rate at .05. Fisher's *r* to *z* transformation was used to convert the correlation coefficients to *z* scores with a mean of zero and a standard deviation of one in order to allow for comparison of the actual correlation values. Steiger's *z*-test was then used to determine the difference between each pair of independent correlations.

Table 14 shows the correlations for both ethnicity categories between math computational fluency and the standardized math scores. All correlations were significant at the .01 level for non-White ethnic groups, but for White students, the first grade spring administration was non-significant when correlated with AIMS math and none of the first

grade CBM-M administrations for White students were significant when correlated with the Stanford-10 selected items.

Table 14

Correlations between Math Computational Fluency and High Stakes Test Scores by Ethnicity

CBM-M Administration	M	SD	AIMS Math	SAT-10 Math
White				
First Grade Fall	15.25	8.66	.17*	.12
First Grade Winter	29.32	8.52	.21*	.12
First Grade Spring	36.94	7.90	.16	.12
Second Grade Fall	35.08	15.66	.55*	.48*
Second Grade Winter	50.01	15.54	.43*	.36*
Second Grade Spring	53.67	17.46	.48*	.44*
Third Grade Fall	18.05	12.72	.45*	.39*
Third Grade Winter	23.30	12.44	.42*	.33*
Third Grade Spring	29.74	15.60	.54*	.48*
Combined Non-White Ethnicity				
First Grade Fall	10.93	9.67	.61*	.50*
First Grade Winter	26.11	9.83	.54*	.47*
First Grade Spring	36.27	6.28	.44*	.37*
Second Grade Fall	27.02	13.44	.50*	.45*
Second Grade Winter	41.56	15.97	.45*	.43*
Second Grade Spring	49.42	19.06	.44*	.43*
Third Grade Fall	13.02	13.31	.50*	.43*
Third Grade Winter	20.03	13.54	.42*	.33*
Third Grade Spring	26.46	16.04	.55*	.47*

Note. CBM-M = Curriculum Based Measurement in Math Fluency; AIMS DPA = Arizona Instrument to Measure Standards Dual Purpose Assessment; SAT-10 = Stanford Achievement Test-10th Edition DPA Items.* $p < .01$. White Students $n = 251$, Combined Non-White Ethnicity $n = 107$

Table 15 lists the Fisher's *r* to *z* transformations of math fluency and standardized math test correlation scores by ethnicity. As in the gender analysis, Steiger's *z*-test was employed as an inferential measure to determine whether the value of the difference between the ethnicity correlation coefficients was statistically significant or due to chance factors. The ethnicity differences for the first grade CBM-M administrations, as correlated with both state standardized scores, were found to be significantly different at the two-tailed .025 level; however, no other significant differences were observed.

Table 15

Results of Steiger's z-Test on Differences in Pearson r Coefficients for Ethnicity.

	AIMS		SAT-10	
CBM-M Administration	Stieger's *Z*	*P* Value	Stieger's *Z*	*P* Value
First Grade Fall	4.60	.00*	3.67	.00*
First Grade Winter	3.35	.00*	3.33	.00*
First Grade Spring	2.66	.01*	2.29	.02*
Second Grade Fall	0.59	.56	0.33	.74
Second Grade Winter	0.21	.83	0.71	.48
Second Grade Spring	0.43	.67	0.11	.91
Third Grade Fall	0.55	.58	0.41	.68
Third Grade Winter	0.00	1.00	0.00	1.00
Third Grade Spring	0.12	.90	0.11	.91

* Significant Difference: Two-Tailed $p < .025$

Fifth Research Question

The fifth research question addressed how the universal CBM-M screening time of year related to student success on high stakes tests.

The means, standard deviations, and Pearson product-moment correlations were calculated to assess the relationships between the first through third-grade math CBM fluency screening scores and the standardized scores for AIMS DPA and Stanford-10 math tests. Mean score analysis was conducted for all screening times across all grade levels. A mean value was also calculated for fall, winter, and spring from the values collected across each given grade level and Fisher's *r* to *z* transformation was used to convert the third grade mean correlation coefficients to *z* scores with a mean of zero and a standard deviation of one in order to allow for comparison across all of the actual correlation values. The Hotelling's *t*-test was then used to assess of the difference between dependent correlations with one common measure. The third grade was selected for this portion of the analysis by virtue of the chronological proximity of both the

screening and standardized test administrations. Holm's Sequential Bonferroni Procedure was used to control for Type I Error across the multiple correlations (Abdi, 2010).

Table 16 shows the means, standard deviations, and correlation coefficients calculated between math fluency probe administration time and results on the AIMS DPA and SAT-10 math portions. Mean score analysis shows an overall increase in math computational fluency scores for all grade levels across screenings. However, the largest score decrease occurred over the summer between the second grade spring and third grade fall administrations. On average, the participants scored within the STEEP end of year Instructional range for math calculation fluency by the winter and spring screenings, with only the second grade winter and spring administrations meeting or exceeding the Mastery range of 40 or greater place-value digits correct threshold for the end of year. Consistent with previous findings, moderate to strong significant correlations were observed between all CBM-M screening times of year and both state standardized math scores.

Table 16

Correlations between Math Computational Fluency Probe Administration Time and High Stakes Test Scores

CBM-M Administration	M	SD	AIMS Math	SAT-10 Math
First Grade Fall	13.96	9.17	.35*	.31*
First Grade Winter	28.36	9.04	.35*	.28*
First Grade Spring	36.74	7.44	.23*	.19*
Second Grade Fall	32.67	15.46	.56*	.51*
Second Grade Winter	47.48	16.12	.47*	.43*
Second Grade Spring	52.40	18.03	.47*	.45*
Third Grade Fall	16.54	13.08	.49*	.43*
Third Grade Winter	22.33	12.85	.43*	.35*
Third Grade Spring	28.75	15.78	.55*	.48*
All Grades Fall Mean	21.06	10.43	.59*	.52*
All Grades Winter Mean	32.72	9.80	.56*	.48*
All Grades Spring Mean	39.30	11.22	.56*	.50*

Note. CBM-M = Curriculum Based Measurement in Math Fluency; AIMS DPA = Arizona Instrument to Measure Standards Dual Purpose Assessment; SS = Scale Score; SAT-10 = Stanford Achievement Test-10th Edition DPA Items. * $p < .01$. $N = 358$

The R^2 coefficients for each of the listed correlations are shown in Table 17. Moderate to strong relationships can be noted across all grade-level CBM-M administrations with the strongest effect sizes being observed in the second and third grade. The listed effect sizes are identical to research question number two.

Table 17

R^2 Coefficients between Math Computational Fluency Probe Administration Time and High Stakes Test Scores

CBM-M Administration	M	SD	AIMS Math	SAT-10 Math
First Grade Fall	13.96	9.17	.12	.10
First Grade Winter	28.36	9.04	.12	.08
First Grade Spring	36.74	7.44	.05	.04
Second Grade Fall	32.67	15.46	.31	.26
Second Grade Winter	47.48	16.12	.22	.18
Second Grade Spring	52.40	18.03	.22	.20
Third Grade Fall	16.54	13.08	.24	.18
Third Grade Winter	22.33	12.85	.18	.12
Third Grade Spring	28.75	15.78	.30	.23
All Grades Fall Mean	21.06	10.43	.35	.27
All Grades Winter Mean	32.72	9.80	.31	.23
All Grades Spring Mean	39.30	11.22	.31	.25

$N = 358$

After applying the Fisher *r* to *z* transformation, Hotelling's *t*-test was employed as an inferential measure to determine whether the value of the differences between dependent correlations with a common index measure for the third grade screening times of year were statistically significant. The third grade screening times of year were chosen for analysis because they displayed consistently strong relationships to state standardized outcomes and also because, when considering the present research question, it would not be possible to distinguish between the predictive validity of the CBM-M screening administration times from the earlier years and the natural maturational process that

students go through as they gain more academic experience. The results are displayed in Table 18.

Table 18

Results of Hotelling's t-Test on Third Grade Math Computational Fluency Administrations

	AIMS		SAT-10	
Comparison	Hotelling's *T*	*P* Value	Hotelling's *T*	*P* Value
Fall—Winter	1.59	1.58	2.03	2.08*
Fall—Spring	1.83	1.82	1.45	1.44
Winter—Spring	3.40	3.33*	3.49	3.43*

*Two-tailed critical is 1.96 for $p < .05$ and 2.58 for $p < .01$. $N = 358$

The correlation differences between the winter and spring computational fluency screenings with the AIMS math score were significant, $Z = 3.33$, $p < .01$. In addition, the correlation differences between the fall and winter screenings with the SAT-10 score were significant, $Z = 2.08$, $p < .05$, as were the correlation differences between the winter and spring screenings with the SAT-10, $Z = 3.43$, $p < .01$.

Chapter 4

DISCUSSION

Research Summary

The RTI model is intended to enable educators to identify students who have needs in fundamental academic skill areas found to be associated with overall classroom success, and to provide them with the necessary supports in order to close any potential gaps between deficiency and proficiency (Clarke et al., 2011). Response to Intervention also plays a key role in special education identification since the reauthorization of IDEA in 2004 (Lembke, Hampton, & Beyers, 2012). Additionally, in recent decades educational policy in the United States has become increasingly focused on results and accountability as measured primarily by standardized, high-stakes test scores. (Minskoff & Allsopp, 2003).

As a systematic means of quantifying the growth of students' basic academic skill sets, CBMs are essential to any RTI service delivery because they provide an efficient and cost-effective method of collecting a large number of meaningful data points. Math computational fluency CBMs are used within the RTI system to determine which students may benefit from math interventions because students struggling with basic mathematical concepts may experience a much more difficult time in passing math courses and performing well on standardized tests, all of which could lead to academic failure (Clarke et al., 2011; Minskoff & Allsopp, 2003). Most statewide achievement tests are first administered at the end of the third grade (Keller-Margulis, Shapiro, & Hintze, 2008), so it is important to understand the correlation between CBM scores from earlier grades to third grade test performance in order to better identify academically at-risk

students as early as possible (Keller-Margulis et al., 2008). In addition, third grade standardized test results are becoming increasingly more critical, with some states, such as Arizona for example, mandating that failure in the third grade will result in grade retention. Although this law is currently only in effect for reading (ARS, 15-101), it could be extended to other academic areas in the future. Being able to look at CBM-M scores as one valid predictor of how students may perform on state standardized math tests would provide further evidence that there is an adequate correlation between what these brief probes measure and what is being taught in the classroom, which in turn would help to validate their use as an evidence-based practice (Keller-Margulis et al., 2008). Previous research has identified moderate to strong relationships between CBM-M data and state standardized test performance in Pennsylvania (Shapiro et al., 2006) and Minnesota (Jiban & Deno, 2007). While studies such as these have provided promising results, there still remains a limited amount of applied research focused on math CBMs, especially when compared to the available empirical support for analogous measures of oral reading fluency (Jiban & Deno, 2007; Keller-Margulis et al., 2008).

The present study was conducted to further explore the relationship between CBMs of math computational fluency and high stakes standardized testing performance by analyzing the predictive validity of CBM-M universal screening probe administrations on the math portions of Arizona's state standardized tests, the AIMS DPA and Stanford-10. In addition, this study assessed the relationship between math computational fluency and specific areas of mathematics, such as number sense, operations, geometry, and algebraic knowledge. Finally, the study also investigated whether any particular CBM-M screening time of year had a stronger relationship to high-stakes testing than the others.

The participants in this study were 358 students selected from six elementary schools in a large suburban Arizona school district. The measures used were the first through third grade math computational fluency universal screening probes from the System to Enhance Educational Performance (STEEP), and the mathematics portions of the AIMS DPA and Stanford-10, including the five math blueprint strands from the AIMS test: Number and Operations; Data Analysis, Probability, and Discrete Mathematics; Patterns, Algebra, and Functions; Geometry and Measurement; and Structure and Logic.

Standardized Math Testing

Conclusions. Scaled scores from the state and national standardized math tests administered in Arizona were correlated with participants' first through third grade math CBM computational fluency screening scores in order to determine the presence of a significant relationship that is not due to chance. Following this, R^2 coefficients were calculated as a measure of effect size for those relationships. It was hypothesized that that there would be significant correlations of moderate to strong effect size (R^2's ranging from .09 to .25) between general computational math fluency screening scores and performance on the AIMS and Stanford-10 scores.

As hypothesized, there were significant correlations observed between math computational fluency across all three grades and results on the AIMS test. Further, moderate to strong effect sizes were also observed with the CBM-M screening scores, accounting for 12 to 30 percent of the variance in AIMS performance. The one exception to this was the first grade spring screening. Similarly, there were also significant correlations found between all three years of math computational fluency screenings and

Stanford-10 performance. Again, moderate to strong effect sizes of 9 to 26 percent of variance accounted for were detected in most of the associations, with the exception of weaker relationships demonstrated in the first grade winter and spring screenings.

Implications. These findings support prior research suggesting that computational fluency CBM has overall good predictive capability for standardized state test performance (Foegen, Jiban, & Deno, 2007; Shapiro et al., 2006). Similar to Shapiro and colleagues, while the observed correlations may not have been as strong as is typically seen for reading (Devena, 2013) they still tended to be statistically significant with effect sizes in the moderate to strong range. Based upon these results, student performance on early primary math computational fluency measures could be effectively interpreted in conjunction with other data to identify individuals who may need support in mathematics or to forecast a student's potential performance on high-stakes state testing and remediate as necessary. This, in turn, could lead to fewer referrals for special education evaluation and play an integral part in getting students the intervention support that they need in a shorter amount of time (VanDerHeyden & Witt, 2008).

Specific Areas of Math

Conclusions. The key math strands from the AIMS test were correlated with participants' first through third grade math computational fluency screening scores in order to determine the presence of a significant relationship that is not due to chance. Following this, R^2 coefficients were calculated as a measure of effect size for those relationships. It was hypothesized that there would be a moderate relationship (i.e., R^2's ~ .09) between computational math fluency screening scores and the composite skill areas, that the third grade screening scores would have the strongest predictive validity among

the screenings, and that compared to the other strands, the strongest strength of relationship would be observed with Number and Operations.

As hypothesized, all grade-level mean correlations were statistically significant with coefficients between .20 and .54. However, the third grade screenings did not demonstrate the strongest predictive validity evidence, as both the second *and* third grade screenings were quite comparable in terms of effect size across the strands. This could be due to the similarity between the relatively higher level of operation complexity inherent in the second and third grade fluency probes as compared to the simple facts found on the first grade probe. The first grade screenings exhibited the weakest predictive validity overall. As expected though, among the other strands the strongest relationship across all three grade levels was shown with Number and Operations. This is likely due to the fact that the math computational fluency screening probes for first, second, and third grade tend to be developed via the *robust indicators* method (Foegen, Jiban, & Deno, 2007), meaning they are constructed out of the grade-appropriate areas of arithmetic proficiency. These would be simple numeration and operations concepts, and not components from areas such as geometry, probability or algebra (Lembke, Hampton, & Beyers, 2012).

Implications. These results support prior findings suggesting that there is moderate predictive capability demonstrated with CBM-M (Foegen, Jiban, & Deno, 2007; Shapiro et al., 2006), while adding additional support to an area with a historically limited body of research (Jiban & Deno, 2007; Keller-Margulis et al., 2008). When considering the use of early computational fluency measures as a means to identify specific math areas where struggling students need support, it is important to understand the limits of their technical adequacy to that end. Based on results of the current study,

while first grade data is moderately related to the Number and Operations area, it is weakly related to the other areas. However, correlated scores from the second and third grade measures demonstrate moderate to strong effect sizes with all five of the math areas assessed on the AIMS test, which may be useful to consider when gathering data for a remediation plan. Further, with second and third grade scores being so comparably related, the opportunity is there to put interventions into place prior to a struggling student's first high-stakes test administration.

Standardized Math Testing by Gender

Conclusions. After disaggregating the data set by gender, the scaled scores from Arizona's state standardized math tests were correlated with participants' first through third grade math computational fluency screening scores in order to determine the presence of a significant relationship. Fisher's *r* to *z* transformation was used to convert the correlation coefficients to *z* scores in order to allow for comparison of the actual correlation values. Steiger's *z*-test was then used as an inferential measure to determine whether any differences between independent gender correlations were statistically significant. It was hypothesized that there would not be significant differences between the male and female correlations.

The means and standard deviations for males and females on both the AIMS and Stanford-10 were approximate in this sample and did not suggest a disparity in performance between the genders. Consistent with previous results from this study, all correlations across grade level and genders were significant except for the male first grade spring screening. These results support the present hypothesis, as the Steiger's *z*-test did not reveal any significant differences between respective gender correlations for

math computational fluency with both the Stanford-10 and AIMS scores across grade the levels.

Implications. These results support the current trend in gender achievement research literature indicating that boys and girls are more evenly matched in the area of math during the elementary grades than has been previously reported (Cole, 1997; Leahey & Guo, 2001; Tsui, 2007), while running contrary to the notion that girls score significantly lower than boys on standardized testing (Arroyo et al., 2013; Hyde et al., 2008). As there is very little literature specifically addressing gender differences with regard to CBM-M and high stakes state testing, these results provide a beneficial contribution by showing that math computational fluency scores for males and females may have approximately equivalent predictive validity for state test performance. It also suggests that math computational fluency data from the early elementary school years can be used with equal confidence for both males and females in helping to identify support needs in the early elementary years.

Standardized Math Testing by Ethnicity

Conclusions. After the data set was disaggregated by ethnicity, the scaled scores from Arizona's state standardized math tests were correlated with participants' first through third grade math computational fluency screening scores in order to determine the presence of a significant relationship. Fisher's *r* to *z* transformation was used to convert the correlation coefficients to *z* scores in order to allow for comparison of the actual correlation values. Steiger's *z*-test was then used as an inferential measure to determine whether any differences between independent ethnicity correlations were statistically significant, indicating real differences between the groups. It was

hypothesized that there would not be significant differences in the observed predictive validity of computational math fluency screening scores and performance on standardized Arizona statewide mathematics tests when disaggregating by ethnicity.

A review of the means and standard deviations for White and combined ethnicity students on both the AIMS and Stanford-10 showed differences in test scores, with White students scoring higher on the AIMS and SAT-10, on average, than did the other combined ethnic groups. Consistent with previous results from this study, most coefficients across grade levels and ethnic categories were significant with the exception of the correlations between White students' AIMS and the first grade spring screening, as well as their SAT-10 correlations to all three first grade screenings.

The Steiger's *z*-test for this analysis did show significant group differences reflected between the respective ethnicity correlations for all three first grade screening administrations and both the Stanford-10 and the AIMS math test scores. This suggests that, to a significant degree, combined Non-White students' scores on math computational fluency from the first grade screenings were more strongly related to their AIMS and SAT-10 performance than were the White students' scores. As such, the present hypothesis is only partially supported. However, this portion of the result could be construed as trivial since previous findings have indicated that the first grade screenings tend to exhibit weak to moderate predictive validity for both tests, overall. In addition, these observed differences are not surprising considering that they occurred when all of the first grade correlations for the combined ethnicity group were statistically significant while most correlations for White students were non-significant. Moreover,

there were no observed group differences between analogous test correlations for any of the second and third grade screenings.

Implications. These results partially support the research literature on ethnic differences in achievement by showing that despite lower, on average, standardized test performance (Bell, Lentz, & Graden, 1992), CBM-M may be able to generate data that is more in step with students' acquisition of curricula regardless of gender and ethnicity (VanDerHeyden et al., 2007). Similar to gender, there is relatively little literature addressing ethnic differences with regard to CBM-M and high stakes state testing; these obtained results suggest that second and third grade math computational fluency scores for White students and other ethnic groups may have approximately equivalent predictive validity for state test performance, which in turn demonstrates the potential usefulness of early math computational fluency data to locate at-risk students regardless of reported ethnic identity. This is also promising because if used to its full potential in a culturally responsive RTI system, data on CBM-M performance could play a part in reducing the disproportionality of ethnically diverse students being placed in special education. These results should be interpreted with some caution, however, because due to inadequate numbers, students identified as Asian represented only 3.6 percent of the total sample and 12 percent of the ethnically diverse sample. As such, they were included in the combined ethnically diverse category with registered mean scores higher than other subgroups on both CBM-M and the standardized tests, which may have impacted the results.

Time of Year

Conclusions. In order to investigate which screening times of year demonstrate the strongest relationship to standardized testing outcomes, the scaled scores of the

standardized math tests (Stanford-10 and AIMS) were correlated with participants' first through third grade math computational fluency screening scores. Following this, R^2 coefficients were calculated as a measure of effect size for those relationships. A mean score analysis was conducted for all screening times across each grade level, showing an overall increase in math computational fluency scores for all grade levels across screenings. Fisher's *r* to *z* transformation was used to convert coefficients from the third grade screenings to *z* scores. Hotelling's *t*-test was then used to assess of the difference between dependent correlations with a common measure. It was hypothesized that there would be a stronger correlation between math fluency scores and test performance when analyzing the winter and spring screening data per grade level and a weaker correlation with the fall screenings. It was further expected that there would be a significantly stronger correlation with the third grade spring math fluency scores and Arizona statewide mathematics test performance than with either the fall or winter of the third grade scores.

Results of the correlation analysis between math screening times and the math scaled scores were all statistically significant with moderate to strong effect sizes, as reported in research question number two. Across all three of the grade levels and for both state tests the results were mixed. For example, several of the fall correlations demonstrated effect sizes that were comparable to or bigger than winter and spring correlations. Winter computational fluency scores generally had the smallest effect sizes to the AIMS DPA and SAT-10 math component scale scores, which did not support the first part of the current hypothesis.

The remaining portion of the hypothesis was partially supported. When examining the relationship between the third grade CBM-Ms and the standardized test scores on both the AIMS and the SAT-10, the spring correlations displayed the strongest relationship to testing outcomes in the third grade, the fall displayed the second strongest, and the winter the weakest. On the AIMS test, only the difference between the winter and spring correlations was significant, indicating that the spring screening is significantly more predictive than the winter screening, but not significantly more predictive than the fall. On the SAT-10, both the difference between the fall and winter screenings, as well as the winter and spring screenings were statistically significant, while the difference between the fall and spring screenings was non-significant. This further indicates that both the spring and fall screenings have significantly more predictive capability than the winter screening, but that the spring does not demonstrate any more predictive capability than the fall.

Implications. These results are consistent with previous research that has reported somewhat inconsistent findings among CBM screening times of year for reading (Adkins, 2013; Devena, 2013) and overall growth rates with regards to CBM-M screening performance, while still suggesting that the strongest growth, in general, typically occurs between the winter and the spring screenings (Graney et al., 2009). It should be noted, however, that best practice for RTI systems still calls for three universal screening administrations per year (Reschly & Bergstrom, 2009) and that the clearest picture of performance and need only really come into focus when many data points are taken in aggregate, with no single score holding more weight than the others. Still, when it comes to more detailed interpretations of collected data over the course of a school

year, it could be beneficial for RTI teams to take fall and spring screening scores into greater consideration when identifying needs and planning remediation strategies aimed at improving test performance.

Limitations and Future Research Direction

Several limitations and directions for future research should be noted for the current study. One limitation was the use of a single standardized dual purpose assessment with which to assess the predictive validity of math computational fluency on the individual test strands. Standardized tests vary considerably across the remaining states (Shapiro et al., 2006), and these findings may not generalize to other versions. It is also important to note that, including the 15 featured Stanford-10 questions, there are 66 items making up the math portion of the AIMS with each individual strand being only briefly represented item-wise, as was demonstrated in the initial scatterplot analysis (see Figure 2 in Appendix A). Consequently, this limited item content compromises the internal reliability of the individual strand areas themselves.

Another limitation is that there was an unavoidable overlap between the AIMS DPA and Stanford-10 scores in that 15 of the Stanford-10 items were used to calculate both standardized test results. Consequently, it was not possible to draw specific conclusions based on norm-referenced versus criterion-referenced testing outcomes. Further, only the Stanford-10 math portion as paired with the AIMS DPA was investigated. Therefore, the predictive validity of using CBM-M on the full Stanford-10 battery may be different, as there would be a substantially higher number of items. Future research could look at CBM-M as it relates to a full, stand-alone administration of the Stanford-10 in order to generate more concrete results.

In addition, Arizona recently discontinued the AIMS test as its measure of state educational standards. As of this writing, the Arizona Department of Education has chosen Arizona's Measurement of Educational Readiness to Inform Teaching (AzMerit) as the replacement measure, and it will be administered for the first time during the month of April 2015. In future research endeavors, it would be useful to conduct a similar analysis with this updated Arizona assessment to determine if current findings are consistent across different measures. Moreover, it would also be more beneficial to address the relationship between math computational fluency and the different skill areas of mathematics using broader measures of each particular subcomponent.

Another limitation of this study had to do with the ability to generalize findings to a wider population of students, as the analysis featured a general education sample from six elementary schools in just one school district. Consequently, the sample was demographically limited in its regional and national representation. Further, as the number of ethnically diverse students in the sample was small compared to the number of White students, the non-White students were collapsed into one category. This combined ethnicity category included Asian students, who had higher score outcomes, on average, than the students from other ethnic backgrounds, which could have had an effect on the results. Future research along these lines should take steps to employ a more representative sample with an adequate number of diverse students in order to ensure broader generalizability and strength of findings. In addition to this, it would be beneficial if future studies were able to assess longitudinal differences to see if trends reported in the research vary in upper grade levels.

The data set for this study was built from archival data and so it was not possible to control for extraneous variables such as the actual time of administration across schools, which could potentially impact the relationship between math computational fluency and high stakes test performance. To avoid this kind of limitation, future research should exact more control on time of administration in order to ensure higher confidence in the observed results.

Another limitation of this study was the fact that the only CBM-M information available was screening data. This is a direct reflection of how little attention CBMs of math receive compared to oral reading fluency. By itself, a single CBM fluency score is insufficient for the purpose of making educational decisions for students. The bigger picture on student performance and potential need only truly comes into focus when an appropriate amount of data is aggregated and analyzed for problem solving. Future research would greatly benefit from having an abundance of progress monitoring math computational fluency data points in addition to screening data.

Finally, there were several missing scores in the available data used to compile the sample set for this study, necessitating the replacement of those missing values with calculated predicted values. This could have had an impact on the overall results as compared to an intact data set with no missing values. Future research should take care to minimize the need for statistically generated scores by ensuring that there are minimal missing data pieces. Taken altogether, these limitations require a cautious interpretation of the research results.

Conclusion

At its core, like all psychological research, this study was about behavior. More specifically, it was about assessing simple response capabilities as potential determinants of more complex response capabilities. Thought of in this sense, we can reduce our fundamental understanding of CBM, high stakes standardized testing or any type of ability assessment as an evaluation of the current status of an individual's behavioral repertoire within a selected domain. The depth and breadth of a desired evaluation is directly proportionate to practical factors such as item count, item type, adequacy of test design, time availability, etc., and exists on a continuum with instruments as circumscribed as CBM on one end and meticulously constructed standardized instruments on the other. For the purpose of identifying which students in a given school population may need intervention support in order to prevent them from falling behind in the curriculum, CBM is well supported in the research as simple and effective. As such, if it has been consistently demonstrated that CBM targets the appropriate fundamental skill sets by which to accurately forecast future challenges, it should follow that CBM can also likewise provide some information about the likelihood of certain testing behaviors, which could prove to be very useful to educators and students everywhere.

The challenge addressed in these pages has been the relative dearth of research on math computational fluency CBM and whether it can consistently demonstrate the same kind of utility as its oral reading fluency counterpart (Shapiro et al., 2006), which is important because of the valuable resource allocations that hinge on the interpretation of this and other such data. Limitations notwithstanding, the obtained results were generally congruent with previous research and provide further support for the usefulness of CBM-

M to identify students at risk of academic failure and potentially poor outcomes on high stakes test performance. The findings showed that 2-minute samples of math computation ability recorded in the fall and spring of the second and third grade years have a moderate to strong relationship to student performance on both normative and criterion referenced standardized math assessments given in Arizona regardless of gender and ethnicity.

This study contributes to the empirical research base supporting the use of CBM-M as a predictor of statewide assessments in mathematics. With the collection of a simple instance of behavior, a plan for remediation can be initiated that is aimed at building a more efficient and effective behavioral repertoire for the mathematics domain. In this way, simple behavior has far reaching possibilities and ceases to be so simple after all.

References

A New Era: Revitalizing Special Education for Children and Their Families. (2002). Report of the Presidents Commission on Excellence in Special Education. Washington, DC: U.S. Department of Education.

Abdi, H. (2010). Holm's sequential Bonferroni procedure. In N. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 1–8). Thousand Oaks, CA: Sage.

Adkins, J. (2013). *An examination of bias in oral reading fluency: Differential effects across race, gender, and socioeconomic status* (Doctoral dissertation) Retrieved from Dissertations and Theses database. (UMI No. 3604787)

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association. (2003). Guidelines on multicultural education, training, research, practice, and organizational change for psychologists. *American Psychologist, 58*, 377-402.

Anderson, D., Lai, C. F., Alonzo, J., & Tindal, G. (2011). Examining a grade-level math CBM designed for persistently low-performing students. *Educational Assessment, 16*, 15-34.

Arizona Department of Education. (2012). *Arizona's Instrument to Measure Standards: Spring guide to test interpretation AIMS and Stanford 10.* Arizona: Author.

Arizona Department of Education. (2013). *Arizona's Instrument to Measure Standards 2012 technical report.* Arizona: Author. Retrieved from the Arizona Department of Education http://www.azed.gov.

Arizona Department of Education. (2014). Assessment overview. Retrieved from http://www.azed.gov/standards-development-assessment/

Arizona State Legislature. (2013). Arizona Revised Statutes (ARS) §§ 15-741 - 15-744.

Arroyo, I., Burleson, W., Tai, M., Muldner, K., & Woolf, B. P. (2013). Gender differences in the use and benefit of advanced learning technologies for mathematics. *Journal of Educational Psychology, 105,* 957-969.

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

Barnes, A. C., & Harlacher, J. E. (2008). Clearing the confusion: Response-to-intervention as a set of principles. *Education and Treatment of Children, 31,* 417-431.

Batsche, G. (2007, Month). Problem solving and response to intervention. Paper presented at the Illinois IEA Professional Development Workshop: Response to intervention: Accelerating Achievement for ALL students, Illinois, Retrieved September 27, 2014, from www.ieanea.org/media/DrBatschePresentation.ppt.

Beal, C. R. (1999). Special issue on the math-fact retrieval hypothesis. *Contemporary Educational Psychology, 24,* 171-180.

Bell, P. F., Lentz, F. E., & Graden, J. L. (1992). Effects of curriculum-test overlap on standardized test scores: Identifying systematic confounds in educational decision making. *School Psychology Review, 21,* 644-655.

Berkeley, S., Bender, W. N., Peaster, L. G., Saunders, L. (2009). Implementation of response to intervention: A snapshot of progress. *Journal of Learning Disabilities, 42,* 85-95.

Bradley, R., Danielson, L., & Hallahan, D. P. (Eds.) (2002). *Identification of learning disabilities: Research to practice.* Mahwah, NJ: Lawrence Erlbaum.

Bums, M. K. (2004). Using curriculum-based assessment in consultation: A review of three levels of research. *Journal of Educational and Psychological Consultation, 15*, 63-78.

Burns, M. K., Appleton, J., & Stehouwer, J. (2005). Meta-analytic review of responsiveness-to-intervention research: Examining field based and research implemented models. *Journal of Psychoeducational Assessment, 23,* 381-394.

Burns, M. K., VanDerHeyden, A. M., & Jiban, C. L. (2006). Assessing the instructional level for mathematics: A comparison of methods. *School Psychology Review, 35,* 401-418.

Casey, M., Nuttall, R., Pezaris, E., & Benbow, C. (1995). The influence of spatial ability on gender differences in math college entrance test scores across diverse samples. *Developmental Psychology, 31*, 697–705. doi: 10.1037/0012-1649.31.4.697

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159. doi:10.1037/0033-2909.112.1.155

Cole, N. S. (1997). *The ETS gender study: How females and males perform in educational settings.* Princeton, NJ: Educational Testing Service. Reports-Research (ERIC Document Reproduction Service No. ED 424 337).

CTB/McGraw-Hill. (2002). *TerraNova, the second edition.* Monterey, CA: Author.

Clements, D. H., & Sarama, J. (2008). Focal points: Pre-K to kindergarten. *Teaching Children Mathematics, 14,* 361-365.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 122,* 155-159.

Clarke, B., Smolkowski, K., Baker, S. K., Fien, H., Doabler, C. T., & Chard, D. J. (2011). The impact of a comprehensive tier I core kindergarten program on the achievement of students at risk for mathematics. *The Elementary School Journal 111*, 561-584.

D'Agostino, J. V. (2010). *Arizona mathematics standard and assessments alignment*. Retrieved from the Arizona Department of Education http://www.azed.gov/assessment/files/2014/04/aims-2010-math-final-report.pdf

Delazer, M., Girelli, L., Grana, A., & Domahs, F. (2003). Number processing and calculation: Normative data from healthy adults. *The Clinical Neuropsychologist, 17,* 331-350.

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

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

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

Devena, S. (2013). *Relationship of oral reading fluency probes on students' reading achievement test scores* (Doctoral dissertation) Retrieved from Dissertations and Theses database. (UMI No. 3602853)

Donovan, M. S., & Cross, C. T. (Eds.). (2002). *Minority students in special and gifted education.* Washington, DC: National Academy Press.

Dozier, A. L., & Barnes, M. J. (1997). Ethnicity, drug user status and academic performance. *Adolescence*, *32*, 825-837.

Duncan, G. J., Claessens, A., & Engel, M. The contributions of hard skills and socio-emotional behavior to school readiness. Working papers, Institute for Policy Research at Northwestern University, 2004. www.northwestern.edu/ipr/publications /workingpapers/wpabstracts05/wp0501.html.

Elliott, S. N., & Fuchs, L. S. (1997). The utility of curriculum-based measurement and performance assessment as alternatives to traditional intelligence and achievement tests. *School Psychology Review, 26*, 224-234.

Elliott, S. N., Huai, N., & Roach, A. T. (2007). Universal and early screening for educational difficulties: Current and future approaches. *Journal of School Psychology, 45,* 137-161.

Espinosa, L. M. (2005). Curriculum and assessment considerations for young children from culturally, linguistically, and economically diverse backgrounds. *Psychology in the Schools, 42,* 837-853.

Erickson, R., Ysseldyke, J., Thurlow, M., & Elliot, J. (1998). Inclusive assessments and accountability systems. *Teaching Exceptional Children, 31,* 4-9.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,* 1149-1160.

Fisher, R.A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron, 1,* 3-32.

Fleischman, H. L., Hopstock, P. J., Pelczar, M. P., & Shelley, B. E. (2010). *Highlights from PISA 2009: Performance of 15-year-old students in science and mathematics literacy in an international context.* Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

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., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal or Special Education, 41,* 121-139.

Fore, C., Boone, R., Lawson, C., & Martin, C. (2007). Using curriculum-based measurement for formative instructional decision-making in basic mathematics skills. *Education, 128 (2),* 324-332.

Fuchs, D., & Fuchs, L. S. (2001). Responsiveness-to-intervention: A blueprint for practitioners, policymakers, and parents. *Teaching Exceptional Children, 38,* 57-61.

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

Fuchs, L., Fuchs, D., Yazdian, L., & Powell, S. (2002). Enhancing first grade children's mathematical development with peer-assisted learning strategies. *School Psychology Review, 31*, 569-583.

Gibbons, K. (2008). Evaluating RTI's effectiveness over the long term. *School Administrator, 65,* 13-14.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*, 117-135.

Good, R. H., III, & Salvia, J. (1988). Curriculum bias in published, norm-referenced reading tests: Demonstrable effects. *School Psychology Review, 17,* 51-60.

Graney, S. B., Missall, K., Martínez, M., & Bergstrom, M. (2009). A preliminary investigation of within-year growth patterns in reading and mathematics curriculum-based measures. *Journal of School Psychology, 47*, 121-142.

Green, S., B., & Salkind, N. J. (2011). *Using SPSS for windows and Macintosh: Analyzing and understanding data* (6th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Gresham, F. M. (2007). Evolution of the response-to-intervention concept: Empirical foundations and recent developments. In S. Jimerson, M. Burns, & A. VanDerHeyden (Eds.), *Handbook of response to intervention: The science and practice of assessment and intervention* (pp. 10–24). New York: Springer.

Harcourt Educational Measurement (2003). *Technical manual: Stanford Achievement Test* (10th ed.). San Antonio, TX: Harcourt Assessment.

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

Hernández-Finch, M. E. (2012). Special considerations with response to intervention and instruction for students with diverse backgrounds. *Psychology in the Schools, 49,* 285-296.

Hintze, J. M. (2009). Curriculum-based assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The Handbook of School Psychology, Fourth Edition* (pp. 397-409). Hoboken, NJ: John Wiley & Sons, Inc.

Hosp, J. L., Hosp, M. A., & Dole, J. K. (2011). Potential bias in predictive validity of universal screening measures across disaggregation subgroups. *School Psychology Review, 40,* 108-131.

Hosp, J. L., & Madyun, N. (2007). Addressing disproportionality with response to intervention. In S. Jimerson, M. Burns, & A. VanDerHeyden (Eds.), *Handbook of response to intervention: The science and practice of assessment and intervention* (pp. 10–24). New York: Springer.

Hotelling, H. 1931. The generalization of Student's ratio. *Annals of Mathematical Statistics 2,* 360–378.

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Cengage Wadsworth.

Hughes, S. J. (2008). Comprehensive assessment must play a role in RTI. In E. Fletcher-Janzen & C. R. Reynolds (Eds.), *Neuropsychological Perspectives on Learning Disabilities in the Era of RTI: Recommendations for Diagnosis and Intervention* (pp. 115-130). Hoboken, NJ: John Wiley & Sons, Inc.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008, July 25). Gender similarities characterize math performance. *Science, 321*, 494 – 495. doi:10.1126/science.1160364

Ikeda, M. J., Neessen, E., & Witt, J. C. (2008). Best practices in universal screening. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 103-114). Bethesda, MD: National Association of School Psychologists.

Individuals with Disabilities Education Act of 2004*,* Pub. L. No. 108-446, 20 U.S.C §§1400-1491 (2004).

Jacob, S., & Hartshorne, T. S. (2007). *Ethics and law for school psychologists* (5th ed.). Hoboken, NJ: John Wiley & Sons, Inc.

James, D. W., Jurich, S., & Estes, S. (2001). *Raising minority academic achievement: A compendium of education programs.* Washington, DC: American Youth Policy Forum.

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

Jordan, N., Kaplan, D., Locuniak, M., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice, 22,* 36-46.

Kavale, K. (1990). The effectiveness of special education. In T. B. Gutkin & C. R. Reynolds (Eds.), *The Handbook of School Psychology* (2nd ed., pp. 868-898). New York: Wiley.

Kavale, K. A. (2005). Effective intervention for students with specific learning disability: The nature of special education. *Learning Disabilities, 13,* 127-138.

Kavale, K. A. (2007). Quantitative research synthesis: Meta-analysis of research on meeting special education needs. In L. Florian (Ed.), *The Sage Handbook of Special Education* (pp. 207-201). London: Sage Publications.

Kavale, K. A., & Forness, S. R. (1999). Effectiveness of special education. In T. B. Gutkin & C. R. Reynolds (Eds.), *The Handbook of School Psychology* (3rd ed., pp. 984-1024). New York: Wiley.

Kavale, K. A., Kauffman, J. M., Bachmeier, R. J., & LeFever, G. B. (2008). Response-to-intervention: Separating the rhetoric of self-congratulation from the reality of specific learning disability. *Learning Disability Quarterly, 31,*135-150.

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

Klingner, J. K., & Edwards, P. A. (2006). Cultural considerations with response to intervention models. *Reading Research Quarterly, 41*, 108–117.

Kranzler, J. H., Flores, C. G., & Coady, M. (2010). Examination of the cross-battery approach for the cognitive assessment of children and youth from diverse linguistic and cultural backgrounds. *School Psychology Review, 39*(3), 431-446.

Leahey, E., & Guo, G. (2001) Gender differences in mathematical trajectories. *Social Forces, 80,* 713-732.

Linacre, J. M. (1994). *Many-facet rasch measurement* (2nd ed.). Chicago, IL: University of Chicago Social Research.

MacMillan, P. D. (2001). Simultaneous measurement of mathematics growth, gender, and relative-age effects: Many-faceted rasch applied to CBM scores. Paper presented at the annual conference of the Canadian Society for the Study of Education (CSSE), Laval, QU, Canada.

Manzo, K. K., & Galley, M. (2003). Math climbs, reading flat on '03 NAEP. *Education Week, 23*, 1-2.

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

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). New York, NY: Psychology Press.

Mazzocco, M., & Thompson, R. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research & Practice, 20,* 142-155.

Minskoff, E., & Allsopp, D. H. (2003). Academic success strategies for adolescents with Learning disabilities and ADHD. Baltimore: Brookes Publishing.

Morgan, S. L., & Mehta, J. D. (2004). Beyond the Laboratory: Evaluating the survey evidence for the disidentification explanation o f Black-White differences in achievement. *Sociology of Education*, *77*, 82-101.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study

Center, Boston College. Retrieved from http://timssandpirls.bc.edu/timss2011/international-results-mathematics.html

National Center for Education Statistics. (2005). *National Assessment of Educational Progress: Mathematics assessment*. Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (2013). *The Condition of Education 2013*. Washington, DC: U.S. Department of Education.

National Center for Education Statistics (2013). *The Nation's Report Card: Trends in Academic Progress 2012* (NCES 2013–456). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington, DC: U.S. Department of Education.

National Reading Panel (2000). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* Washington, DC: National Institute of Child Health and Human Development. www.nichd.nih.gov/publications/nrp/smallbook.pdf.

National Science Board. (2003, August 14). The science and engineering workforce: Realizing America's potential. Retrieved from http://www.nsf/gov/nsb/documents/2003/nsb0369/

Nintzel, J. (June 14, 2013). Hispanics leading minority growth in AZ. *Tucson Weekly.* Retrieved from http://www.tucsonweekly.com/TheRange/archives/2013/06/14/hispanics-leading-minority-growth-in-az

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Nyberg, K. L., McMillin, J. D., O'Neill-Rood, N., & Florence, J. M. (1997). Ethnic differences in academic retracking: A four-year longitudinal study. *Journal of Educational Research, 91*, 33-44.

OECD (2010), PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science (Volume I). http://dx.doi.org/10.1787/9789264091450-en

Ortiz, S. O. (2006). Multicultural issues in school psychology practice: A critical analysis. In Bonnie K. Nastasi, (Ed.). *Multicultural issues in school psychology* (pp. 151-165).

Paleologos, T., & Brabham, E. (2011). The effectiveness of DIBELS oral reading fluency for predicting reading comprehension of high- and low-income students. *Reading Psychology, 32,* 54-74.

Patton, J. R., Cronin, M. E., Bassett, D. S., & Koppel, A. E. (1997). A life skills approach to mathematics instruction: Preparing students with learning disabilities for the real-life demands of adulthood. *Journal of Learning Disabilities, 30*, 178-187.

Peterson, P. E., Woessmann, L., Hanushek, E. A., & Lastra-Anadón, C. X. (2011). *Globally challenged: Are U. S. students ready to compete? The latest on each state's international standing in math and reading.* Cambridge, MA: Harvard's Program on Education Policy and Governance & Education Next, Taubman Center for State and Local Government, Harvard Kennedy School.

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

Reschly, D. (August, 2005). *RTI Paradigm Shift and the Future of SLD Diagnosis and Treatment.* Paper presented to the Annual Institute for Psychology in the Schools of the American Psychological Association, Washington, DC.

Reschly, D. J. (2008). School psychology paradigm and beyond. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 3-15). Bethesda, MD: National Association of School Psychologists.

Reschly, D. J., & Bergstrom, M. K. (2009). Response to intervention. In T. B. Gutikin & C. R. Reynolds (Eds.), *The Handbook of School Psychology, Fourth Edition* (pp. 434-460). Hoboken, NJ: John Wiley & Sons, Inc.

Reschly, A., Busch, T., Betts, J., Deno, S., & Long, J., (2009). Curriculum-based measurement oral reading as an indicator of reading achievement; A meta-analysis of the correlational evidence, *Journal of School Psychology, 47*, 427-469.

Reynolds, C. R. (2008). RTI, neuroscience, and sense: Chaos in the diagnosis and treatment of learning disabilities. In E. Fletcher-Janzen & C. R. Reynolds (Eds.), *Neuropsychological Perspectives on Learning Disabilities in the Era of RTI: Recommendations for Diagnosis and Intervention* (pp. 14-27). Hoboken, NJ: John Wiley & Sons, Inc.

Reynolds, C. R., & Shaywitz, S. E. (2009). Response to intervention: Ready or not? or, from wait-to-fail to watch-them-fail. *School Psychologist Quarterly, 24,* 130-145. doi: 10.1037/a0016158

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology, 46,* 343-366.

Royer, J. M., & Walles, R. (2007). Influences of gender, motivation and socioeconomic status on mathematics performance. In D. B. Berch & M. Mazzocco (Eds.), *Why is math so hard for some children?* (pp. 349 –368). Baltimore, MD: Brookes.

Sadker, D. (1999). Gender equity: Still knocking at the classroom door. *Educational Leadership, 56,* 22-25.

Saffer, N. (1999). Math and your career. *Occupational Outlook Quarterly, 43*, 31-35.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2007). *Assessment in special and inclusive education* (10th ed.). Boston: Houghton Mifflin.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). San Diego: Jerome M. Sattler.

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,* 19-35.

Shinn, M. (August, 2005). *Who is LD? Theory, Research, and Practice.* Paper presented to the Annual Institute for Psychology in the Schools of the American Psychological Association, Washington, DC.

Shinn, M. R. (2008). Best practices in using curriculum-based measurement in a problem-solving model. In A. Thomas & J. Grimes (Eds.), *Best Practices in School Psychology V* (pp. 243-261). Bethesda, MD: National Association of School Psychologists.

Schacht, W. H. (2009). *Industrial competitiveness and technological advancement: Debate over government policy* (Order Code RL33528). CRS Issue Brief for Congress. Washington, DC: Congressional Research Service.

Statistical Solutions: Intelligence in Data. (2012). *Stanford Achievement Test-10.* Retrieved from http://www.statisticssolutions.com/academic-solutions/resources/directory-of-survey-instruments/standford-achievement-test-10-sat-10/#sthash.FGEJJ77w.KiQf8fYF.dpuf

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*, 795-819.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245-251.

Sullivan, A. L., & Kucera, M. (2011, February). *A framework for culturally responsive assessment.* Poster presented at the meeting of the National Association of School Psychologists, San Francisco, CA.

Swanson, H. L. (2008). Neuroscience and RTI: A complementary role. In E. Fletcher-Janzen & C. R. Reynolds (Eds.), *Neuropsychological Perspectives on Learning Disabilities in the Era of RTI: Recommendations for Diagnosis and Intervention* (pp. 28-53). Hoboken, NJ: John Wiley & Sons, Inc.

Tabachnick, B., & Fidell, L. (1996). *Using multivariate statistics* (3th ed.). New York: Herper Collins College Publishers.

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*, 493-513.

Tilly, W. D. (2008). The evolution of school psychology to science-based practice: Problem solving and the three-tiered model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 17-36). Bethesda, MD: National Association of School Psychologists.

Tindall, G., Marston, D., & Deno, S. L. (1983). *The reliability of direct and repeated measurement (Research Rep. No. 109)*. Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Tsui, M. (2007). Gender and mathematics achievement in China and the United States. *Gender Issues, 24,* 1-11.

United States Department of Education. (2010). *Education secretary Arne Duncan issues statement on the results of the program for international student assessment.* Retrieved from http://www.ed.gov/news/press-releases/education-secretary-arne-duncan-issues-statement-results-program-international-s

United States Department of Education Office of Special Education and Rehabilitative Services. (2002). *A new era: Revitalizing special education for children and their families.* Washington, DC: Author.

Valentine, J. & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes.* Washington, DC: What Works Clearinghouse. Retrieved from http://www.wmich.edu.ezproxy1.lib.asu.edu/evalphd/wp-content/uploads/2010/05/Effect_Size_Substantive_Interpretation_Guidelines.pdf

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

VanDerHeyden, A. M. & Witt, J. C. (2008). Best practices in can't do/won't do assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 131-139). Bethesda, MD: National Association of School Psychologists.

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,* 225-256.

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

Wanzek, J., Roberts, G., Linan-Thompson, S., Vaughn, S., Woodruff, A. L., & Murray, C. S. (2010). Differences in relationship of oral reading fluency and high-stakes measures of reading comprehension. *Assessment for Effective Intervention, 35,* 67-77.

Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41,* 85-120.

Weaver, B., & Wuensch, K. L. (2013). SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients. *Behavior Research Methods, 45,* 880-895.

Weiderholt, L. (1974). Historical perspective on the education of the learning disabled. In L. Mann & D. Sabatino (Eds.), *The second review of special education* (pp. 103-152). Austin: Pro-Ed.

Whitebook, M. (2003). Early education quality: Higher teacher qualifications for better learning environments. Center for the Study of Child Care Employment. Retrieved July 2014 from http://iir.berkeley.edu/cscce

Witt, J. (2002). *Screening to Enhance Educational Performance (STEEP).* Retrieved on May 4, 2014 from, http://www.joewitt.org.

Wilson, J. (2005). *The relationship of Dynamic Indicators of Basic Early Literacy Skills (DIBELS) oral reading fluency to performance on Arizona Instrument to Measure Standards (AIMS)* (Technical Report). Tempe, AZ: Tempe School District No.3.

Wodrich, D. L., & Schmitt, A. J. (2006). *Patterns of learning disorders: Working systematically from assessment to intervention.* New York, NY: The Guilford Press.

Wodrich, D. L., Spencer, M. L. S., & Daley, K. B. (2006). Combining use of RTI and psychoeducational testing: What we must assume to do otherwise. *Psychology in the Schools, 43*, 798-806.

Ysseldyke, J., Burns, M., Dawson, P., Kelley, B., Morrison, D., Ortiz, S., Rosenfeld, S., & Telzrow, C. (2006). *School psychology: A blueprint for training and practice III.* Bethesda, MD: National Association of School Psychologists.

Ysseldyke, J., Burns, M., Dawson, P., Kelley, B., Morrison, D., Ortiz, S., Rosenfeld, S., & Telzrow, C. (2008). The blueprint for training and practice as the basis for best practices. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 37-70). Bethesda, MD: National Association of School Psychologists.

Zirkel, P. A. (2013). The Legal Dimension of RTI: Part II. State Laws and Guidelines. Retrieved from http://www.rtinetwork.org

APPENDIX A

SCATTERPLOT ANALYSIS

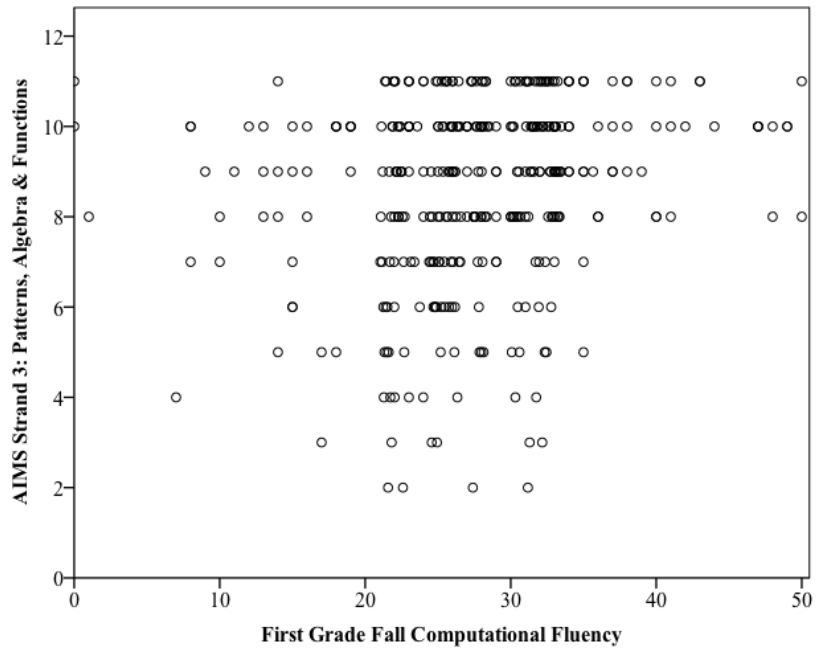


Figure 2. Scatterplot depicting the relationship between first grade fall math computational fluency scores and the AIMS Math Strand #3: Patterns, Algebra, and Functions scores as an example of the violation of normality assumption via discrete interval-level data in research question #1.

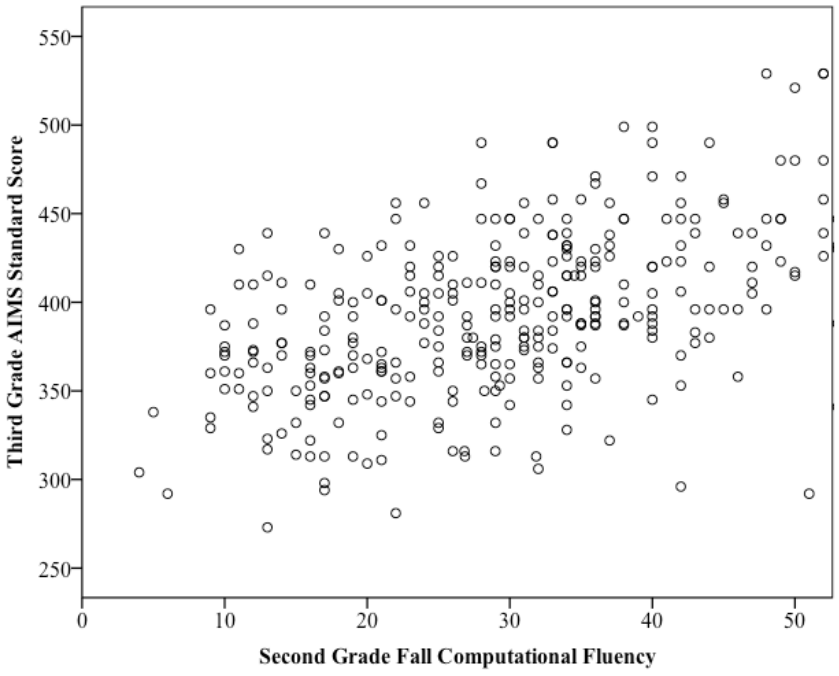


Figure 3. Scatterplot depicting a general linear relationship between second grade fall math computational fluency scores and AIMS Scores for research questions 2 & 5.

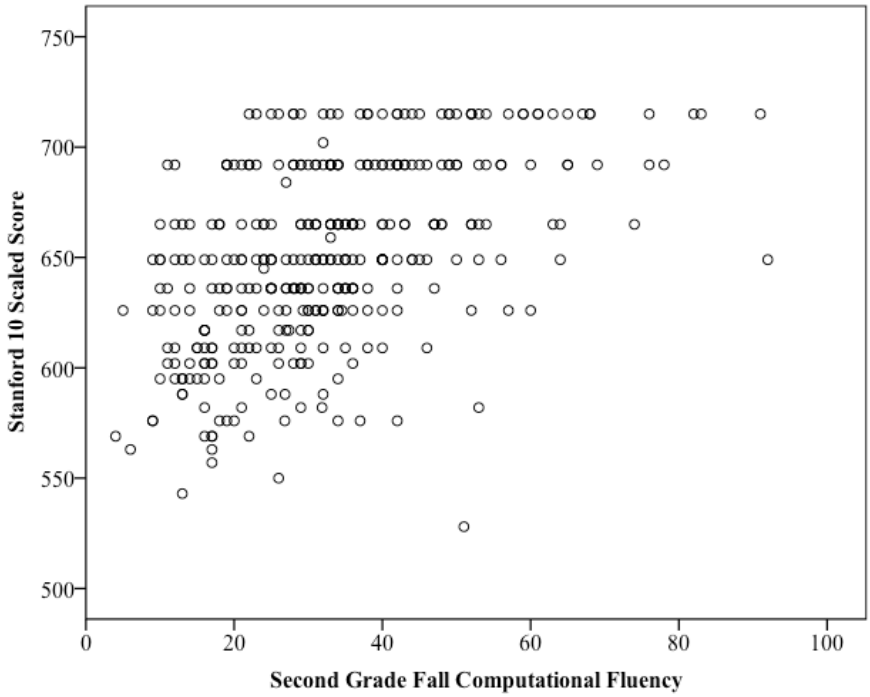


Figure 4. Scatterplot depicting a general linear relationship between second grade fall math computational fluency scores and Stanford-10 Scores for research questions 2 & 5.

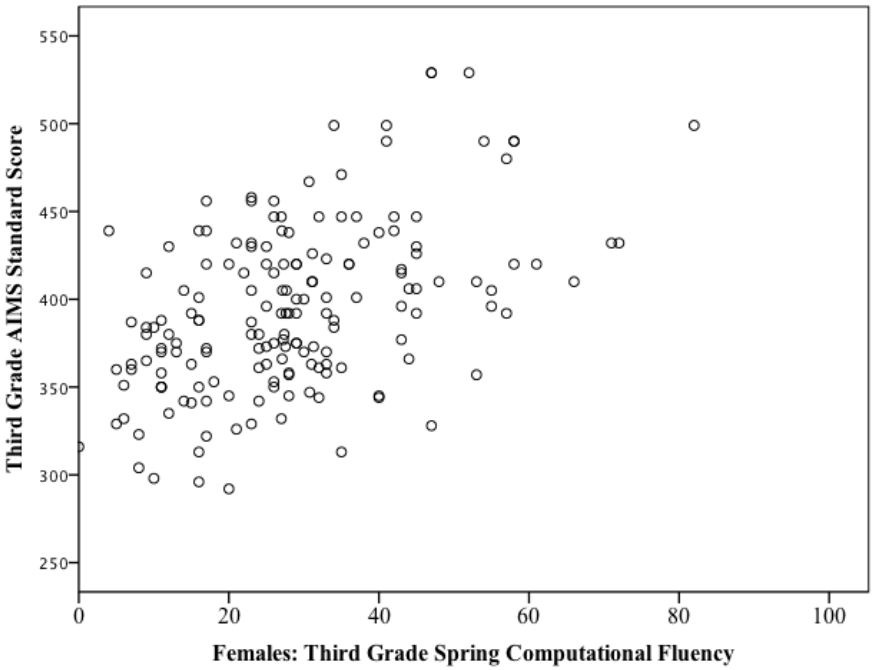


Figure 5. Scatterplot depicting a general linear relationship between female third grade spring math computational fluency scores and AIMS Scores for research question # 3.

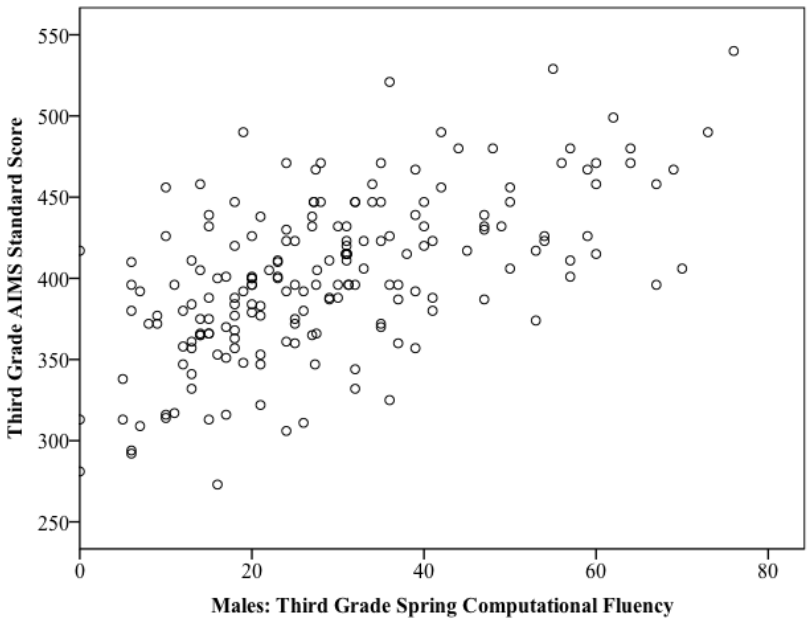


Figure 6. Scatterplot depicting a general linear relationship between male third grade spring math computational fluency scores and AIMS Scores for research question # 3.

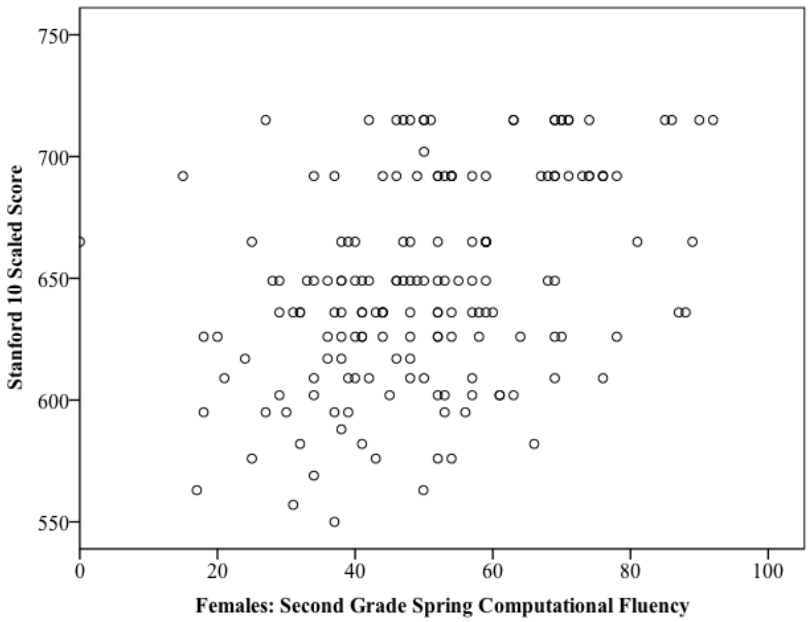


Figure 7. Scatterplot depicting a general linear relationship between female second grade spring math computational fluency scores and Stanford 10 Scores for research question # 3.

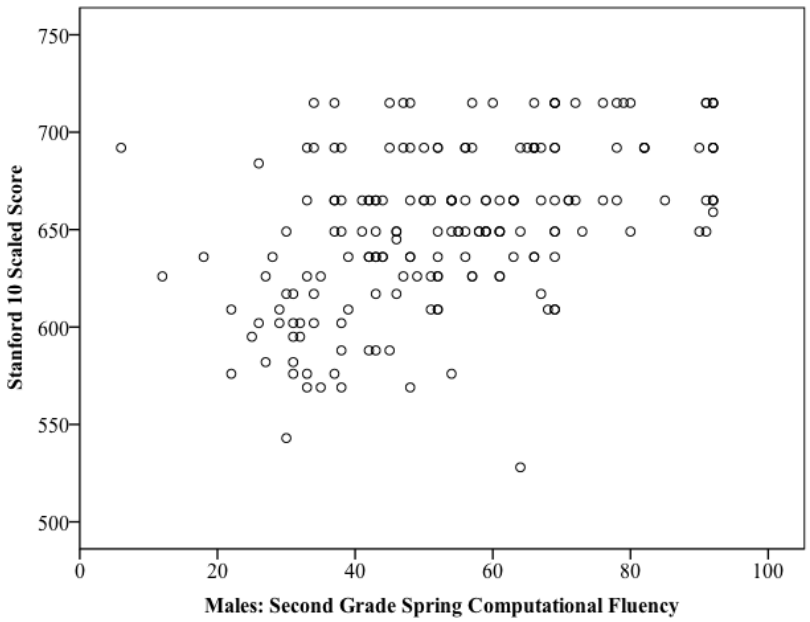


Figure 8. Scatterplot depicting a general linear relationship between male second grade spring math computational fluency scores and Stanford 10 Scores for research question # 3.

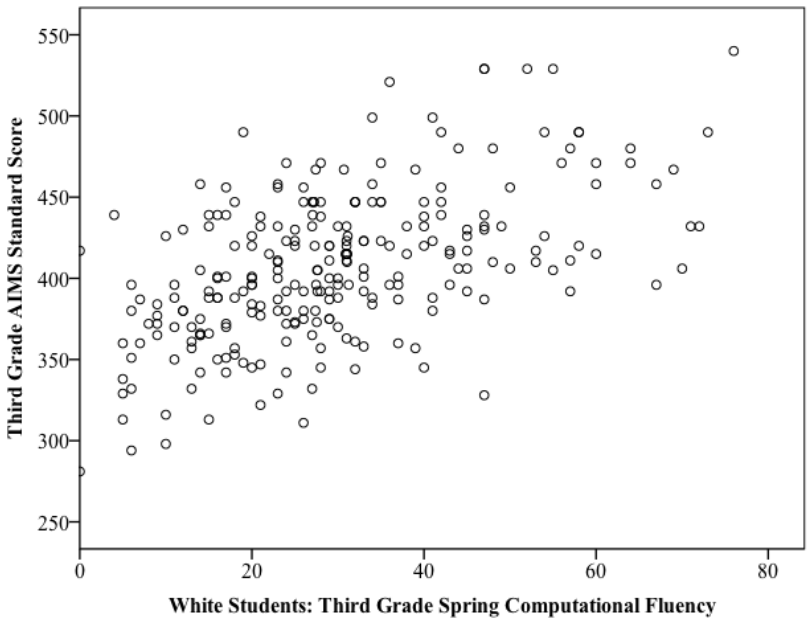


Figure 9. Scatterplot depicting a general linear relationship between White student third grade spring math computational fluency scores and AIMS Scores for research question # 4.

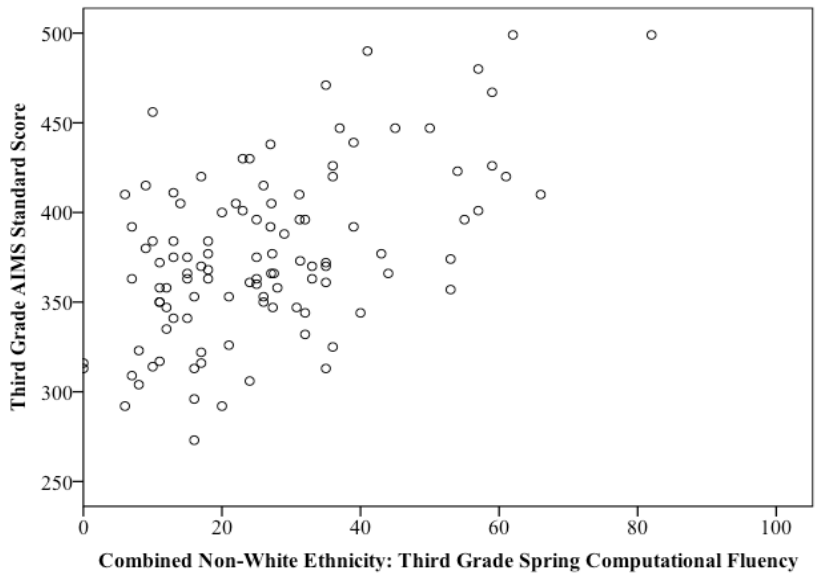


Figure 10. Scatterplot depicting a general linear relationship between Non-White student third grade spring math computational fluency scores and AIMS Scores for research question # 4.

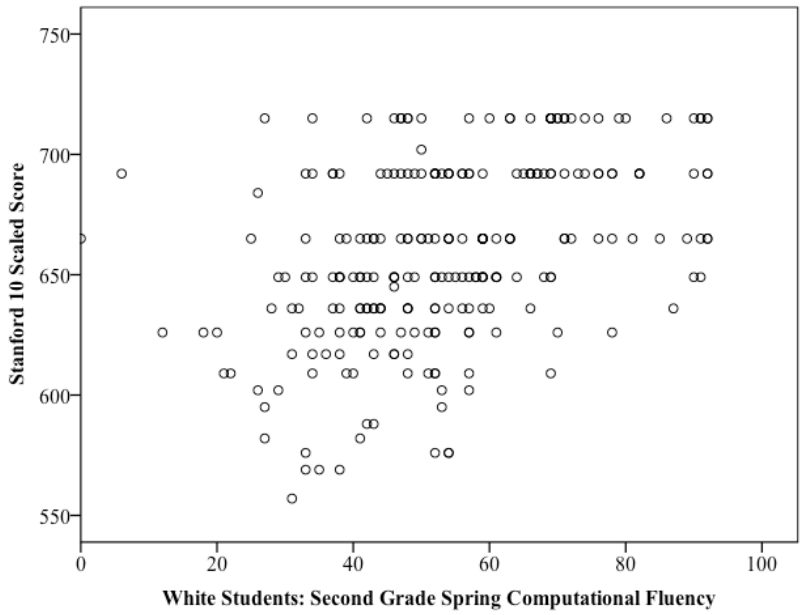


Figure 11. Scatterplot depicting a general linear relationship between White student second grade spring math computational fluency scores and Stanford 10 Scores for research question # 4.

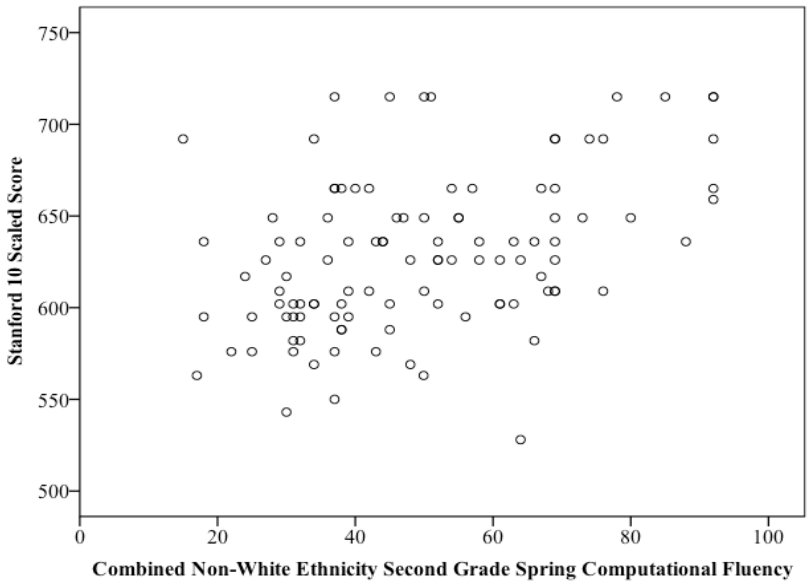
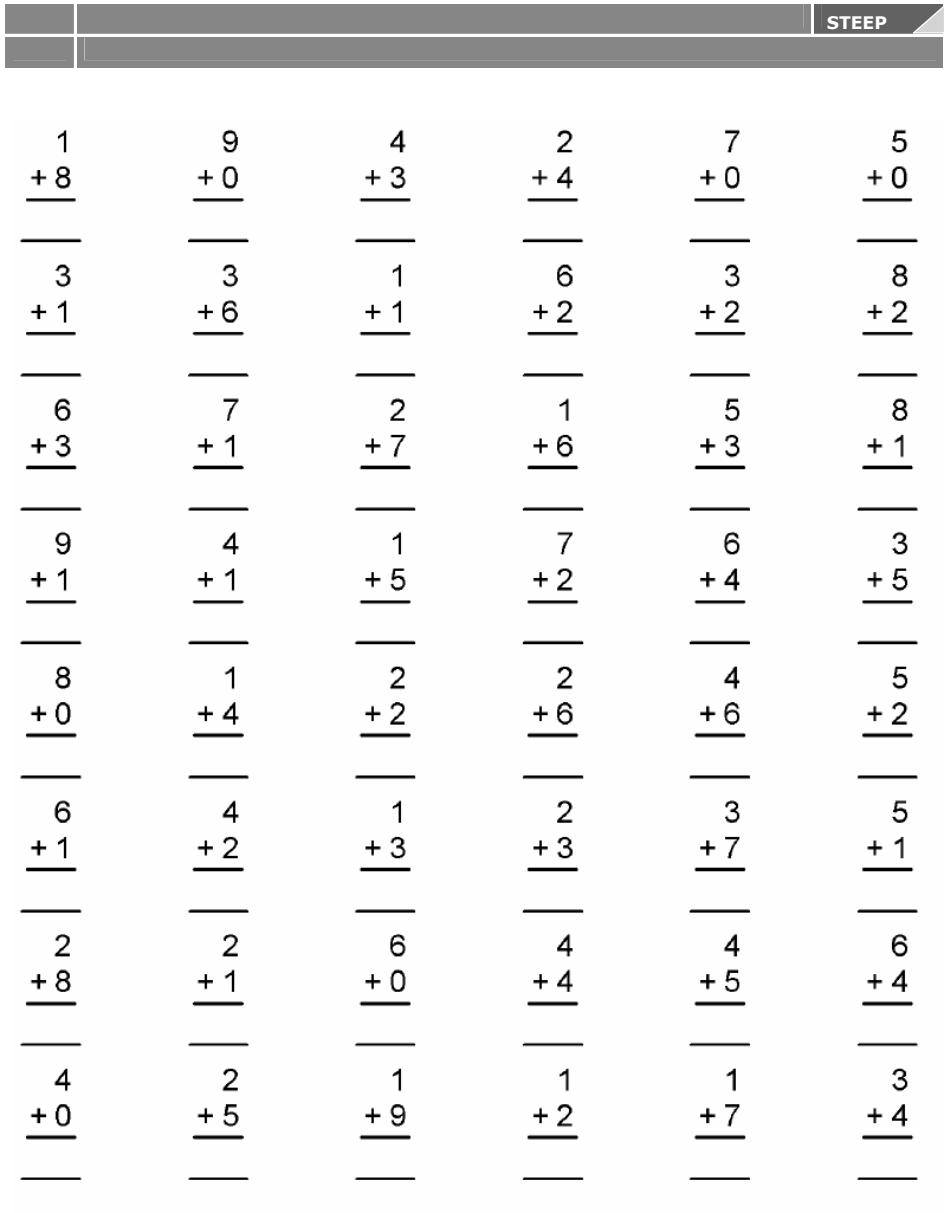


Figure 12. Scatterplot depicting a general linear relationship between Non-White student second grade spring math computational fluency scores and Stanford 10 Scores for research question # 4.

APPENDIX B

STEEP MATH COMPUTATIONAL FLUENCY CBM SAMPLES

STEEP Math Computational Fluency CBM Probe, Grade 1

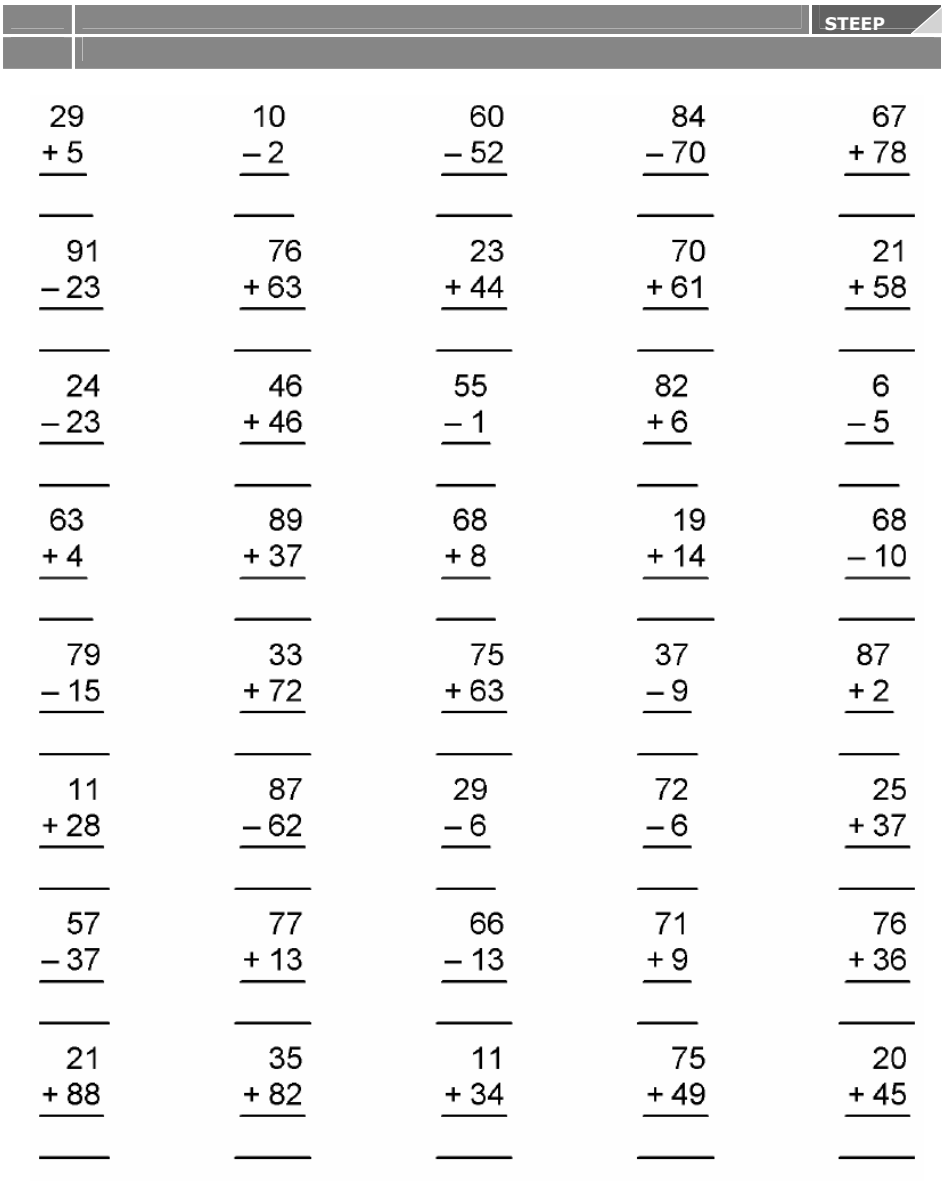


STEEP Math Computational Fluency CBM Probe, Grade 2

STEEP

5 + 1 =	4 + 4 =	10 + 2 =	20 + 3 =	4 + 9 =	6 + 5 =	8 + 5 =
32 + 5 =	34 + 3 =	2 + 4 =	21 + 7 =	32 + 1 =	41 + 4 =	9 + 1 =
26 + 2 =	32 + 4 =	43 + 3 =	2 + 6 =	8 + 9 =	15 + 4 =	5 + 8 =
31 + 2 =	21 + 8 =	13 + 3 =	22 + 5 =	34 + 3 =	26 + 3 =	24 + 2 =
12 + 7 =	14 + 2 =	1 + 2 =	47 + 1 =	22 + 6 =	40 + 2 =	49 + 0 =
23 + 4 =	42 + 5 =	44 + 1 =	8 + 1 =	14 + 5 =	46 + 2 =	16 + 1 =
41 + 1 =	16 + 3 =	30 + 4 =	9 + 2 =	36 + 2 =	50 + 0 =	38 + 1 =

STEEP Math Computational Fluency CBM Probe, Grade 3



APPENDIX C

IRB DOCUMENTATION

ASU Knowledge Enterprise Development

Office of Research Integrity and Assurance

To:	Linda Caterino Kulhavy EDB
From:	Mark Roosa, Chair Soc Beh IRB
Date:	04/09/2012
Committee Action:	**Exemption Granted**
IRB Action Date:	04/09/2012
IRB Protocol #:	1203007660
Study Title:	The relationship of curriculum based measurement probes on standardized achievement tests

The above-referenced protocol is considered exempt after review by the Institutional Review Board pursuant to Federal regulations, 45 CFR Part 46.101(b)(1) .

This part of the federal regulations requires that the information be recorded by investigators in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects. It is necessary that the information obtained not be such that if disclosed outside the research, it could reasonably place the subjects at risk of criminal or civil liability, or be damaging to the subjects' financial standing, employability, or reputation.

You should retain a copy of this letter for your records.