Summer 8-2017

Performance on Monitoring Basic Skills Progress - Computation Probes in First, Second and Third Grade: Is It a Predictor of Pennsylvania System of School Assessment Mathematics Achievement in Third Grade?

Adelle C. Campbell

PERFORMANCE ON MONITORING BASIC SKILLS PROGRESS - COMPUTATION PROBES IN FIRST, SECOND, AND THIRD GRADE: IS IT A PREDICTOR OF PENNSYLVANIA SYSTEM OF SCHOOL ASSESSMENT MATHEMATICS ACHIEVEMENT IN THIRD GRADE?

A Dissertation
Submitted to the School of Graduate Studies and Research
in Partial Fulfillment of the
Requirements for the Degree
Doctor of Education

Adelle C. Campbell
Indiana University of Pennsylvania
August 2017

Indiana University of Pennsylvania
School of Graduate Studies and Research
Department of Educational and School Psychology

We hereby approve the dissertation of

Adelle C. Campbell

Candidate for the degree of Doctor of Education

_____	_____ Timothy J. Runge, Ph.D. Associate Professor of Educational and School Psychology, Advisor
_____	_____ Joseph F. Kovaleski, D.Ed. Professor of Educational and School Psychology
_____	_____ Courtney L. McLaughlin, Ph.D. Assistant Professor of Educational and School Psychology
_____	_____ Michelle Ludwig, D.Ed. Adjunct Faculty, York College of Pennsylvania

ACCEPTED

_____ _____

Randy L. Martin, Ph.D.
Dean
School of Graduate Studies and Research

Title: Performance on Monitoring Basic Skills Progress – Computation Probes in First, Second, and Third Grade: Is It a Predictor of Pennsylvania System of School Assessment Mathematics Achievement in Third Grade?

Author: Adelle C. Campbell

Dissertation Chair: Dr. Timothy J. Runge

Dissertation Committee Members: Dr. Joseph F. Kovaleski
Dr. Courtney L. McLaughlin
Dr. Michelle Ludwig

This study examined the predictive relationship of a brief computation measure administered in the fall, winter, and spring of first, second, and third grade with the mathematic portion of a state-mandated academic achievement test administered in the spring of third grade. The relationship between mathematical achievement and resource availability and sex was also explored.

Multiple linear regression analysis and Pearson correlations indicate the brief computation measure from the winter of first grade through the spring of third grade has a strong predictive relationship with mathematical achievement on the state-mandated academic achievement test administered in the spring of third grade. The brief computation measure in the fall of first grade had a moderate predictive relationship with outcomes on the state-mandated math assessment in third grade. Sex was not found to be an adequate predictor variable. Resource available was a weak predictor of mathematical outcomes, but became more relevant in third grade.

ACKNOWLEDGEMENTS

I've come to the conclusion the most difficult part of this process is trying to appropriately articulate how thankful and appreciative I am for all the support and encouragement I have received. I would like to start off by thanking my dissertation chair, Dr. Runge. Thank you for reminding me how finishing was not a question of ability, but a question of will. This sediment always resonated with me in the moments I contemplated hitting the snooze button and going back to bed. Your guidance has been invaluable. Thank you.

Dr. Kovaleski and Dr. McLaughlin thank you for your knowledge, time, and support. You have taught me much through your example as committee members, professors, and practitioners.

Dr. Ludwig, thank you for answering my never ending questions, being a sounding board for ideas, and for acting as a champion of students. Thank you for encouraging me to "eat a sandwich" and for not crushing my dreams of "read and feed". Your support as a member of my dissertation committee, but more importantly as a friend, is sincerely appreciated.

Thank you to my amazing family, friends, and co-workers for a seemingly endless supply of caffeine, kind words, and comic relief. Thank you for forcing me to relax, scheduling girls' night around my writing schedule, and listening to all my dissertation ramblings. Your thoughtfulness and understanding speak to the wonderful people you are.

Mom, thank you for instilling in me a love of learning through your example. You were my first teacher and by far the most influential. I do not think it is a coincidence four of your five children are involved in education in some form or another. You make us proud, love you. Dad, thank you for the gift of "Remmel Standard Time" and for never asking, "Are you done your dissertation yet?"

Lori, Vicki, Liz, and Russell, you are the best siblings a person could ask for. I could fill several pages thanking you. It would probably make sense only to the five of us, so I won't, but thank you. Vicki, I am so glad "grammar is your thing" because we both know (so does Dr. Runge) it is not mine. Your gracious offer to proofread likely saved me months of time. Liz, thank you for the "gift of time" and encouragement along the way. Also, please be prepared to provide me with a list of good books I have missed out on over the past two years.

Nan and Pap Pap, you have been cheering me on from day one. Thank you for welcoming me into your family and home. I miss evenings sitting on the porch, watching deer, and listening to stories. Thank you for all your support and wisdom. I hope I have made you proud.

To my two sweet boys, Harrison and Jack, when the light at the end of this tunnel seemed particularly far away or especially close, I would think about what I would say to the two of you. This almost always resulted in me getting choked up because of what this experience has taught all of us. I suspect what you have learned extends well beyond earning a degree or writing a dissertation. I hope you learned about setting goals, working hard to reach them, finding balance, and most importantly, how families love and support each other. On several occasions I was given the advice to wait to have children until my doctorate was complete. I do not for one second regret the course I took. It would have been easier, but would not mean nearly as much. Thank you for the all the pictures, candles, and dissertation chap stick. Harrison, thank you for the quiet moments in the wee hours of the morning "working" together and the artwork decorating my office walls. Jack, thank you for infinite hugs and kisses. You have no idea how much you crawling up on my lap and asking which buttons to push so you can help means to me. I love you both so much.

Craig, you have been my biggest supporter not just with my doctorate, but everything since the moment we met. I know this has not been easy for you (understatement). Thank you for tolerating two years of 3:30 am alarm clocks and watching movies in installments because I can't keep my eyes open. I love you even more for it. It is no exaggeration when I say I could not have done it without you. I promise not to take on any big projects for at least a couple months (other than the pile of unfolded laundry in the basement and maybe a marathon…).

Please know, even if my words can't convey it, I am truly appreciative from the bottom of my heart. All of the small acts and words of kindness were not lost on me. Thank you, thank you, and thank you.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

CHAPTER I

INTRODUCTION

Over recent years, the accountability of educational institutions and practices has been pushed into the limelight. There is a significant gap between the national standards for mathematical proficiency and how students are performing in actuality (Kelley, 2008). The 2001 reauthorization of No Child Left Behind (NCLB) increased pressure on schools to have 100% of students performing proficiently on state standards measured by performance on state-generated academic achievement tests in addition to achieving Adequate Yearly Progress (AYP). AYP is measured by several factors (e.g., attendance) but most significantly achievement. School performance on state assessments is linked directly to federal education funding at the state level (Braden & Schroeder, 2004). The end-of-the-year state assessments required by NCLB are considered high stakes testing due to the potentially serious implications these tests may have on school systems (Bell, Taylor, McCallum, Coles, & Hays, 2015; Braden & Schroeder, 2004). Recent legislative changes, specifically the December 2015 signing of the Every Student Succeeds Act (ESSA), seek to lessen the negative ramifications of high stakes testing. However, it is yet unknown how these changes will impact the current educational climate and culture (White House Office of the Press Secretary, 2015).

As educators question and reflect on how to improve educational practices, problem-solving models, which have long been supported in the medical field, have gained increased attention and support (Sansosti & Noltemeyer, 2008). The 2004 reauthorization of the Individuals with Disabilities Education Improvement Act (IDEIA) and ESSA requires use of periodic standardized assessments to inform curriculum implementation and school improvement. This legislation encourages schools to adopt a problem-solving approach to

meeting the needs of all students (Glover & DiPerna, 2007). The recognition that education systems in the United States were not meeting the academic, social, and behavioral needs of all students coupled with this legislation prompted school systems to more readily adopt tiered instruction systems that embraced problem-solving approaches. Most common are the three-tiered versions of Response to Intervention (RTI) and Positive Behavior Support (PBS).

In the most basic sense, RTI is defined as the change in behavior or performance as a function of intervention (Gresham, 2002). RTI systems require ongoing evaluation of student response to instruction and intervention (Daly, Martens, Barnett, Witt, & Olson, 2007). In an RTI model, eligibility and decisions regarding educational programming are made based on a student's response to an intervention that is matched to skill deficits, is research-based, and is implemented with integrity. RTI models are supported by legislation, and research prioritizes the need for an effective core curriculum and early identification of students in need of additional instruction or intervention. Despite the increasing prevalence of RTI implementation, there are limited regulations on how to specifically implement RTI. Therefore, RTI models vary significantly from state to state and school district to school district. A three-tiered model is the most prevalent in both practice and research, with the intensity of assessment and intervention increasing at each tier.

RTI is frequently thought of as a special education initiative because IDEIA supports its use to identify students as having a specific learning disability (Walker & Shinn, 2010). In actuality, many educational systems are implementing tiered intervention models but not as a means of identifying students for special education services. These systems, instead, view RTI as a service-delivery model in which universal instruction is improved so that very few students require tiered supports. Additionally, assessments are regularly administered to identify

emerging skill deficits and provide remedial support in an efficient manner. Communities, education systems, families, and students see the benefit of RTI systems in literacy education and would like to reap the benefits in all academic areas and behavior as well. However, there is limited research regarding best practices of RTI implementation in math in comparison to the body of research available in reading (Clarke, Doabler, & Nelson, 2014). Within recent years, more attention has been given to implementation of problem solving models in mathematics and behavior. The behavior and social/emotional needs of a system is commonly addressed through PBS.

PBS is a global term used to describe an approach for addressing behavior to improve quality of life and decrease maladaptive behaviors that impact functioning. The definition of PBS has changed significantly and rapidly since its emergence in the mid-1980s. PBS has most recently been defined as:

> An approach to behavior support that includes an ongoing process of research-based assessment, intervention, and data-based decision making focused on building social and other functional competencies, creating supportive contexts, and preventing the occurrence of problem behaviors. PBS relies on strategies that are respectful of a person's dignity and overall well-being and that are drawn primarily from behavioral, educational, and social sciences, although other evidence-based procedures may be incorporated. PBS may be applied within a multi-tiered framework at the level of the individual and at the level of larger systems (e.g., families, classrooms, schools, social service programs, and facilities). (Kincaid et al., 2016, p. 71)

Based on this broad definition, PBS can range significantly in intensity and implementation. PBS can be intensive behavioral plans applied to an individual after completion of a functional

behavioral assessment to behavior approaches implemented across an entire system to prevent mild to moderate behavioral problems from emerging. School-wide positive behavior support (SWPBS) or positive behavioral interventions and supports (PBIS) are terms used to describe PBS when applied to entire educational systems. SWPBS is typically used when referring to the multi-tiered systems implemented in kindergarten through twelfth grade education systems (Kincaid et al., 2016). SWPBS focus on creating a safe learning environment for students by providing social and behavioral instruction to promote desired social and behavioral outcomes. Cornerstone features of successful SWPBS systems are embedding evidence-based practices, responsive changes to discipline practices, data-driven decision making, and maintenance of these practices over time (Frey, Lingo, & Nelson, 2010).

Both RTI and SWPBS utilize a multi-tiered method of service delivery. However, the language used to define these systems does not fully encompass their inclusive nature. As a result, RTI is often thought of as an academic, assessment-based system (Stoiber, 2014) whereas SWPBS is behavior-centric. Attempting to implement two separate, multi-tiered systems simultaneously places undue stress on schools and often results in ineffective systems.

Multi-Tiered Systems of Support (MTSS) combines the function of RTI and SWPBS into one dynamic system responsive to the needs of all students. It is important to note that many scholarly works and published text continue to refer to comprehensive multi-tiered service delivery models as RTI or SWPBS. There is a gradual shift, however, toward use of MTSS as a more general term to describe these systems to decrease the misconception that problem-solving models are solely a method of identifying students for special education services or are unique to academic or behavioral concerns. For the sake of clarity in this dissertation, MTSS will be used

to describe any multi-tiered service delivery model. RTI will be used to describe the process of identifying students for special education services within MTSS models.

MTSS models aim to improve student outcomes in academics and social/emotional development. Reported benefits of a MTSS model include a significant reduction of students being referred to and qualifying for special education programs, a decrease in office discipline referrals, and, with each year of implementation, an increase in the number of students demonstrating proficiency on state assessments (Lillenstein, Fritschmann, & Moran, 2012; Michigan Department of Education, 2012; Stoiber, 2014).

MTSS models of service delivery emphasize high-quality core instruction provided to all students. Educators rely on evidence-based practices to differentiate instruction for students and facilitate a positive learning environment (Stoiber, 2014). Supplemental instruction and intervention are provided in the areas of reading, writing, mathematics, and behavior with increasing intensity based on student need. The intensity of the intervention should be matched with the severity of student need. Students are able to receive appropriate levels of support in a time-efficient manner, with or without a special education label. While there are variations within MTSS models, most states have adopted a three-tier approach, with intensity of instruction increasing at each tier. Tier 1 is the foundation of the education system. Essential features of Tier 1 include high quality instruction, appropriate differentiation of instruction, alignment with state standards of essential learning, and universal screening. Tier 2 incorporates targeted instruction and intervention for students who have been identified as needing additional academic, social, or behavior support in addition to high quality instruction provided through Tier 1. Tier 3 is intensive, individualized intervention for a small percentage of students (1-5%) who have not made adequate progress when provided support at Tiers 1 and 2 intensity levels.

Tier 3 services are provided in addition to those at Tier 1 and Tier 2 (Albers & Kettler, 2014; Stoiber, 2014).

The National Association of School Psychologists (NASP) outlines six key features of successful MTSS systems: (a) differentiated instruction within a high quality core curriculum, (b) universal screening, assessment, and monitoring progress; (c) focus on prevention and intervention, (d) fidelity of interventions, (e) evidence-based practices, and (f) professional development (Cowan, Vaillancourt, Rossen, & Pollitt, 2013; Stoiber, 2014). Universal screenings are one of the key components of MTSS systems, and the focus of the present study.

Universal screening is defined as "the systematic assessment of all children within a given class, grade, school building, or school district, on academic and/or social emotional indicators that the school personnel and community have agreed are important" (Ikeda, Neessen, & Witt, 2008, p. 103). The use of universal screening is supported in federal and state education legislation. The 2004 reauthorization of the IDEIA and NCLB required the use of periodic standardized assessments to inform curriculum implementation and school improvement (Glover & DiPerna, 2007). While ESSA seeks to minimize the amount of time students are spent engaged in standardized testing, it continues to require periodic standardized testing of students. Additionally, ESSA has allowed several states to pilot use of local assessments to evaluate student outcomes and teacher effectiveness in place of a statewide assessment (National Education Association, 2015).

Universal screenings are generally administered to all students three times a year to access critical academic skills (Gerzel-Short & Wilkins, 2009). Universal screening data are used for three essential functions. Primarily, universal screening data are used to identify students who may need remediation or enrichment and acceleration. Classroom teachers are

encouraged to use universal screening data, in conjunction with other assessment data, teacher observations, and work samples, as a resource when grouping students for differentiated instruction (Parisi, Ihlo, & Glover, 2014). Secondly, universal screening data are used to calculate local normative data and evaluate student growth over time. The third essential function of universal screening is to evaluate core curriculum and effectiveness of school systems.

When making decisions about universal screenings, educational systems need to consider (a) what general outcome measures to be assessed, (b) the use of a broadband or narrowband screening tool, (c) who will be assessed, and (d) whether or not to employ a gated evaluation system (Albers & Kettler, 2014). District and school teams should reflect on these issues before making a decision about the screener(s) to be used, lest an inefficient screener is implemented.

General outcome measures represent the curricular content or specific skills students are expected to learn which represent global learning outcomes. General outcome measures utilize standardized administration procedures and "produce critical indicators of student performance" (Fuchs & Deno, 1991, p. 493). General outcome measures can represent a sampling of skills directly linked to a curriculum. A second approach to general outcome measurement is to assess global skills that require students to apply knowledge they would be expected to master by the end of the school year or indicators of growth (e.g., oral reading fluency; Foegen & Deno, 2001). Both approaches to general outcome measurement can use broadband or narrowband screening instruments. Broadband instruments assess multiple skills simultaneously (e.g., concept and application probes), whereas narrowband instruments assess a specific area of functioning (e.g., number identification; DiPerna, Bailey, & Anthony, 2014).

Decision-makers within education systems also need to determine who will be assessed and how frequently. Typically, universal screening is completed three times a year with all students, in the fall, winter, and spring. Alternatively, a gated evaluation system may be employed instead of traditional universal screening methods. Gated evaluation is the process of "involving multiple assessments that cost efficiently identify a subset of individuals from a larger pool of target participants with a combination of methods and measures generally arranged in sequential order" (Walker, Small, Severson, Seeley, & Feil, 2014, p. 47). In such a gated system, for example, all students may complete a simple, broadband screener. A small proportion of those students, as identified on the initial broadband screener, are gated into the next stage of screening. In this next stage of the gated system, the small number of students complete a narrowband assessment. Such gated assessment practices are more common in PBIS (Walker et al., 2014), and increasingly recommended in MTSS models.

The practice of universally screening students to provide early intervention for academic and behavior concerns is more common at the early elementary level than at the secondary level. This is in part due to the significance of early intervention to remediate learning and behavioral challenges. Duncan et al. (2007) found a strong relationship between early learning and later academic outcomes in mathematics, further highlighting the need for research to support effective identification of mathematical deficits.

Examination of universal screening logistics is an important consideration for school teams. Equally important, however, is consideration of the quality of the screening instrument used. The reliability and validity of an instrument are primary indicators of their quality. Reliability indicates the accuracy and consistency of a measure, or how close an observed score is to a true score (DiPerna et al., 2014). Validity is the "extent to which a score on a measure

represents the construct of interest" (DiPerna et al., 2004, p. 234). For a universal screening instrument to be effective, it needs to measure what it is purported to measure. In other words, universal screening instruments need to be reliable and correlate with a future learning outcome or have strong validity.

Tier 1 instruction of an MTSS model should be aligned with state standards and utilize data from state assessment measures to evaluate and ensure high quality instruction for all students. One of the crucial requirements of universal screeners is to be reflective of the curriculum so schools can use data to evaluate core instruction. Hence, the need for universal screening tools in primary grades that have a strong correlation with later outcomes on state assessments. Universal screeners do not act as global indicators of the effectiveness of an educational system unless they are aligned with the curriculum. Universal screening measures are described as "essentially worthless" (Ikeda et al., 2008, p. 103) when they are not reflective of the curriculum or robust indicators of student success. Conversely, universal screening instruments that reflect the curriculum or are robust indicators of student success provide an efficient, effective, and cost-effective means to evaluate how well a system is meeting the needs of all students and identifying students who may be in need of additional intervention.

Many academic, social, and behavioral problems can be addressed within the core curriculum and regular education setting through an early, proactive approach (Elliott, Haui, & Roach, 2007; Stoiber, 2014). Many initial decisions regarding student intervention are based on universal screening data within MTSS service delivery models, highlighting the need for valid and reliable universal screening instruments.

As described earlier, reliability is the accuracy and consistency of an instrument to measure a person's true score. Methods of measuring reliability include internal consistency,

test re-test reliability, and interrater reliability. Each of these reliability methods addresses different aspects of reliability and should be used in combination with each other (DiPerna et al., 2014). Moderate to strong reliability is necessary for academic universal screening instruments because data from instruments need to be consistent and stable over time to accurately conclude difference in scores are due to differences among students or systems, not inconsistencies with the screening instrument. Reliability is also thought to be a prerequisite for validity. An instrument is unlikely to demonstrate adequate validity if it is not reliable (Christ & Nelson, 2014).

Universal screening measures should demonstrate moderate to strong validity. Validity can be measured in terms of content-related, construct-related, and criterion-related. This study focuses on the criterion or predictive validity of a math computation measure.

Predictive validity is the extent to which an instrument is able to predict a future outcome, and this indicator of technical adequacy forms the basis of the overall purpose of this study. Specifically, predictive validity measures how strong the relationship is between the universal screening instrument and a future learning outcome or criterion measure (Christ & Nelson, 2014; DiPerna et al., 2014). In this case, the state-mandated academic achievement test is used as the criterion measure because it represents a set of skills students are expected to master by the end of third grade. It is essential that universal screening instruments have strong predictive validity to identify and intervene with students who may be at risk for future learning difficulties in an early and proactive manner. Predictive validity of universal screening instruments generates four possible outcomes described as classification accuracy.

Education systems should consider classification accuracy and decision rules when choosing and employing universal screening measures. Classification accuracy, one indicator of

technical adequacy, refers to an instrument's likelihood of correctly identifying students in need of additional intervention (true positive), correctly identifying students who are not in need of additional intervention (true negative), incorrectly identifying students as being in need of intervention (false positive), or failing to identify students in need of intervention (false negative). Sensitivity and specificity are calculated based on the four classification accuracy features of a screening instrument. Sensitivity is the percentage of true positives detected by a test. Specificity accounts for the percentage of true negatives detected by a test (Christ & Nelson, 2014; VanDerHeyden, 2011). Sensitivity and specificity are used to establish appropriate cut-off scores for a screening tool by determining at what score point is a student likely, or not, to reach future learning outcomes while minimizing the likelihood of over- or under-identifying students as being in need of additional interventions. These cut-off scores are then used within decision-making rules regarding providing students with the appropriate intensity of instruction (Parisi et al., 2014). Universal screening tools should demonstrate good technical adequacy to increase the likelihood of true positives and true negatives and decrease instances of false positives and false negatives. In basic terms, instruments that have been designed and validated for universal screening help to identify the appropriate students for tiered supports while minimizing inaccurate predictions of which students may or may not need those tiered supports.

There is a plethora of information regarding screening and intervention in the area of reading; however, there is significantly less research to provide educators direction in the area of mathematics (Clarke, Haymond, & Gersten, 2014; Methe, 2009). As multi-tiered models of service delivery continue to grow in popularity and implementation, it is vital that research-based practices are also applied to the area of mathematics (Gersten et al., 2012; VanDerHeyden,

2010). Research exploring the technical adequacy of universal screeners is especially critical within a MTSS model because multiple assessment instruments are used to make educational decisions with more frequency than in the traditional discrepancy model (VanDerHeyden, 2011). Typically, this is done by correlating one instrument with another instrument previously shown to measure the same construct or with some meaningful outcome, including end-of-the year tests, referred to as the outcome criterion (VanDerHeyden, 2010).

Curriculum-based measures (CBM) have shown to be appropriate universal screening instruments assuming they demonstrate sufficient reliability and validity evidence, measure constructs that are meaningful and reflective of the school curriculum, and are developed using universal design. Universal design refers to the ability of an assessment instruments to be sensitive to the characteristics of all potential test takers. Instruments with universal design typically have clearly identified constructs; bias-free content; been formatted to be accessible to all students; allow for accommodations; simple, clear administration and scoring procedures; appropriate readability; and legibility of text and graphics (Anderson et al., 2011). CBM math probes are relatively easy to score; cost effective; can typically be group administered; and can be used for subsequent progress monitoring (Albers & Kettler, 2014; Clarke et al., 2014; Howell & Hosp, 2014). CBM are standardized and have been shown to have moderate to strong predictive accuracy and validity (Clarke et al., 2014).

CBM have also been shown to have outcome utility, another criterial factor when evaluating universal screening instruments. Outcome utility is the ease at which (a) a system, teachers, administrators, and parents can understand the implications of the screening data; (b) the data are useful in guiding instruction/intervention; and (c) the data are able to have a positive impact on student outcomes (Glover & Albers, 2007). The technical adequacy and outcome

utility of universal screening instruments are often based on how well the data generated can predict future outcomes for students (i.e., their predictive power). This is done by correlating screening data with a future outcome, such as demonstrating proficient performance on a standardized test of academic achievement. This future outcome is referred to as the predicted criterion or outcome criterion. The lack of consistent predictor criteria when validating these instruments and unknown classification agreement have been identified as areas in need of more research (Gersten et al., 2012; VanDerHeyden 2010). This study focuses on determining the predictive power of a math computation probe administered in first, second, and third grade with the state assessment given in the spring of third grade. The Pennsylvania System of State Assessment (PSSA) serves as the predictor criterion in this study.

Mathematics achievement is measured by the math composite score on the PSSA mathematics assessment, in addition to scores on the five subtests that yield a composite score. The state assessment reflects Pennsylvania Common Core learning outcomes, therefore, the content assessed has been deemed valuable by Pennsylvania's state department of education. Mathematics achievement is divided into five subtests, established by the PSSA: Numbers and Operations, Measurement, Geometry, Algebraic Concepts, and Data Analysis and Probability.

Numbers and Operations is a subtest of the PSSA-Math (PSSA-M) in which students demonstrate an understanding of numbers, ways of representing numbers, relationships among numbers and number systems, meanings of operations, understanding and application of operations and how they relate to each other, the ability to compute accurately and fluently, and the capacity to make reasonable estimates. The Measurement subtest assesses understanding of measureable attributes of objects and figures, and the units, systems and processes of measurement. Students are also required to apply appropriate techniques, tools and formulas to

determine measurements. This subtest includes calculation of time and elapsed time, length, area, volume and weight of objects, and use of a ruler. Geometry is a subtest of the PSSA-M which measures a student's ability to analyze characteristics and properties of two and three dimensional geometric shapes. The Geometry subtest also requires students to "demonstrate understanding of geometric relationships and identify and/or apply concepts of transformations or symmetry" (Data Recognition Corporation, 2014, p. B-2). Algebraic Concepts is a subtest of PSSA-M which measures a student's ability to "demonstrate an understanding of patterns, relations, and functions and represent and/or analyze mathematical situations using numbers, symbols, words, tables and/or graphs" (Data Recognition Corporation, 2014, p. B-2). The Data Analysis and Probability subtest of PSSA-M requires students to formulate or answer questions that can be addressed with data and/or organize, display, interpret or analyze data (Data Recognition Corporation, 2014).

Statement of the Problem

Research on mathematical learning trajectories indicates that students who demonstrate math skills within the bottom 10th percentile in kindergarten have a 70% likelihood of remaining below the 10th percentile five years later when in fifth grade (Martin et al., 2012; Morgan, Farkas, & Wu, 2009, 2011). Educational practice should emphasize a preventative, proactive approach to potential math difficulties given the stability of learning trajectories over time. However, there is limited consensus about what constitutes well-developed universal screening instruments to support the early identification of students who would benefit from additional math intervention. Glover and Albers (2007) identified several key requirements of universal screening tools. Universal screening instruments need to reflect the standards of importance to the education system, be based on universal design, demonstrate technical adequacy, and have

outcome utility. Minimal research has been conducted to determine the predictive validity of universal screening instruments in mathematics with high-stakes testing, such as state-mandated tests of academic achievement. The limited availability of validated measures of a system's effectiveness and individual student performance is problematic when systems are striving to provide high quality mathematics instruction to all students and be proactive in remediation of potential mathematics deficits.

The purpose of this study is to further the research about what instruments may constitute effective universal screeners in mathematics. To that end, contemporary research indicates that there are two potentially significant factors outside of the purview of educational systems which may affect mathematical outcomes. There is a growing body of research that indicates students living in low socio-economic status (SES) homes are likely to demonstrate difficulty with mathematical learning, and these deficits are more persistent when compared to students who are not living in poverty. These findings suggest students living in low SES homes many benefit significantly from early, intensive math intervention (Reardon, 2013).

There is a mixed body of research on whether or not sex has an impact on mathematical learning outcomes. Some research indicates a significant difference in mathematical learning between males and females, but other research disputes an achievement gap between males and females (McGraw et al., 2006; Stoet & Geary, 2013). Therefore, this study also explores what, if any, impact SES and sex have on predicting mathematical outcomes.

As the world becomes more focused and reliant on technology, the quality of our science, technology, engineering, and mathematics (STEM) education becomes increasingly important for the United States to remain competitive in a global economy (The President's Council of Advisors on Science and Technology [PCAST], 2011). According to a 2011 report produced by

The President's Council of Advisors on Science and Technology (PCAST), the United States is lagging significantly behind other nations in STEM education at the elementary and secondary level. According to the National Assessment of Educational Progress (NAEP), a significant number of children from low-income homes do not reach basic levels of mathematical proficiency and are under-represented in STEM related professions (NAEP, U.S. Department of Education, 2015). These findings support the need for further research regarding the impact of SES on the development of mathematical skills. SES is included in this study to determine if it is a predictive factor of performance on PSSA-M unrelated to performance on a universal screener in mathematics.

Women are also under-represented in the field of STEM. The increased percentage of women entering the workforce with higher education and advanced degrees observed in other fields of study is not reflected in STEM fields (National Science Foundation, 2015; U.S. Department of Education, 2012). Some hypothesize this is due to sex-differences in mathematical learning, with men outperforming females on mathematical achievement tests (McGraw, Lubienski, & Strutchens 2006). However, there is mixed research on whether a significant difference exists between male and female mathematical achievement (Else-Quest, Hyde, & Linn, 2010). McGraw, at al. 2006 reviewed the United States National Assessment of Educational Progress (NAEP) data from 1990 to 2003 and found a small but statistically significant difference in mathematical achievement between male and female mathematical achievement across multiple years. Other research, contrarily, has not indicated statistically significant differences in mathematical performance between males and females (Else-Quest et al., 2010; Hyde, Fennema, Ryan, Frost, & Hopp, 1990; Scheiber, Reynolds, Halovsky, & Kaufman, 2015). Due to conflicting findings regarding a mathematical achievement gap coupled

with an under-representation of females in STEM related fields, sex is studied as a potential predictive factor on PSSA-M achievement.

MTSS models function on the principle of adjusting instruction and intervention based on responsiveness to the needs of systems and individual students, including students who historically perform poorly in mathematics. The dynamic nature of MTSS models require education systems and educators to be responsive to student and community needs. Universal screenings are the first step when determining which students may be at risk of academic, social, or emotional deficits (Albers & Kettler, 2014). Students who are identified early for potential deficiencies can receive supplemental intervention at varying degrees of intensity to limit the future impact of these difficulties (Albers & Kettler, 2014; Kettler, Glover, Albers, & Feeney-Kettler, 2014). After an extensive analysis of mathematics instruction studies, Slavin and Lake (2008) highlighted the importance of adult responsiveness when young children are developing mathematical skills. To support responsiveness, there is a need to assess children's mathematical skills to promote learning (Salvin & Lake, 2008); however, this is problematic given the complexity of mathematical development and relative dearth of research in mathematics instruction and assessment (Fisher, Dobbs-Oats, Doctoroff, & Arnold, 2012; Murphy, Mazzocco, Hanich, & Early, 2007; Methe, 2009).

Within MTSS service delivery models, the use of gated evaluation procedures are becoming more prevalent to reduce unnecessary strain on schools' financial and personnel resources by reducing the number of false positives (Albers & Kettler, 2014; Fuchs et al, 2011; VanDerHeyden, 2010). Gated evaluation systems, also referred to as multiple-gate models or Smart RTI, involve a series of evaluations to an increasingly smaller number of students to ensure the students who are being identified as being at risk of an academic, social, and/or

emotional difficulties are truly in need of intervention. The number of evaluation gates is determined by the educational system; however, three gate systems are prevalent in current practice. Data utilized in gated evaluation systems can include parent and teacher input, direct observation, and review of work samples in addition to direct assessment. These systems also have the potential to aid educators in identifying specific skill deficits for driving instructional decisions when used in data-analysis teaming. It is critical to have a measure with strong predictive validity at gate 1 to avoid administering the more robust and time consuming measures used in subsequent gates (Albers & Kettler, 2014; Clarke et al., 2014; Walker et al., 2014).

Identification of mathematical deficits by universal screening for the purpose of early and effective intervention has been identified as an area in need of more research (VanDerHeyden, 2010). There is a need for universal screening instruments to predict a future outcome or have high predictive validity. Predictor criteria are usually another instrument already shown to measure the skills being screened or an assessment of the desired skills. Examples of predictor criteria would be a standardized academic test such as the Wechsler Individual Achievement Test, Third Edition (WIAT-III) or a state-mandated test of academic achievement. State-mandated tests of academic achievement are frequently referred to as high stakes testing due to the significant consequences, intended and unintended, that can impact communities, schools, teachers, and students. School districts have a vested interest in being able to predict students who are likely to perform below proficiency on state assessments. Students correctly identified as being at-risk for academic deficits through universal screening are able to benefit from additional instruction and intervention in an attempt to ameliorate academic deficits on a systemic and individual level. This study examines the predictive power of a math computation

probe given in the fall, winter, and spring of first, second, and third grade with the state math assessment given in the spring of third grade. The strength of the relationship between SES and sex with future math outcomes is also explored. It is hoped that this study will contribute to a developing body of research on universal screening for mathematical deficits.

Significance of the Problem

Poor mathematical skills have been linked to truancy, increased disciplinary referrals, risk of unemployment, involvement in the criminal justice system, and increased health risks (Every Child a Chance Trust, 2009). As previously noted, students who initially placed in the bottom 10th percentile of a criterion measure when entering kindergarten but were performing above the 10th percentile upon exiting only had a 30% chance of performing below the 10th percentile five years later while in fifth grade (Morgan et al., 2009, 2011). These findings support the positive impact and necessity of early intervention for academic difficulties. When effective intervention is not made available, early math difficulties correlate highly with poor outcomes on future indicators or mathematics performance.

Students from low SES homes are among this population. According to the Trends in International Mathematics and Science Study (TIMSS), schools with 50% or more of their students living in poverty, determined by number of students accessing free and reduced lunch, performed below average in relation to schools with lower poverty levels (Gonzales et al., 2009). Students living in poverty have demonstrated resistance to improved mathematical education, as evidenced by a slow rate of growth in comparison to other population groups (Aud, Fox, & KewalRamani, 2010).

The predictive relationship with sex and future math outcomes is explored due to conflicted research regarding an achievement gap between males and females. Recent research

does not indicate significant difference in mathematical performance between male and female students. This challenges whether an achievement gap in mathematics between the sexes exists. (McGraw et al., 2006; Stoeb & Geary, 2013). Hyde et al. (1990) found that sex differences in math is based on a stereotype that mathematical thinking is a male-dominate domain, rather than actual differences in male and female math achievement.

There is a substantial amount of research that supports the practice of universally screening all students for the early identification and intervention of academic, social, and emotional difficulties in an educational setting (Kettler et al., 2014). Student outcomes in academics, social skills, and behavior improve significantly when deficits are identified and intervened upon earlier rather than later. Research suggests that if students do not acquire and master basic reading and mathematical skills by the end of third or fourth grade, there is a high likelihood they will continue to struggle throughout their school career (Elliot, Huai, & Roach, 2007; Morgan et al., 2011). Increased academic, social, and/or behavioral difficulties are considered significant predictive factors for poor outcomes such as high school dropout, drug and alcohol abuse, and future or exacerbated mental health issues and highlight the need for early, proactive intervention (Every Child a Chance Trust, 2009; Walker & Shinn, 2010). Poor student outcomes, in turn, impact schools, communities, and society by effecting the local economy and job-market. Ultimately, this becomes a factor when the United States is competing with other nations to remain a leader in the global economy (PCAST, 2011). MTSS models give particular attention to developing the ecological factors that promote positive student outcomes, including home-school relations, school climate, and skill-level of school personnel (Stoiber, 2014).

Over the past decade, several precipitating events have led to an increased focus in the area of mathematics. The National Council of Teachers of Mathematics (NCTM) released their *Focal Points* (2006), effectively ending the debate regarding math curriculums, dubbed the Math Wars, by establishing the importance of automaticity and problem solving skills (Davison & Mitchell, 2008; NCTM, 2006). Two years later, the National Mathematics Advisory Panel released *Foundations for Success* (2008) which identified significant weaknesses in mathematical education and made recommendations to remediate current math education practices. These recommendations are reflected in the 2010 adoption of the Common Core State Standards.

As education systems adapt to provide the more rigorous Common Core State Standards with less financial and personnel resources, the prevalence of MTSS implementation is increasing (Stoiber, 2014). MTSS models promote a more efficient and effective use of personnel and financial resources by addressing the diverse needs of all students in one cohesive system as opposed to many dueling or parallel systems. This allows systems to work smarter, not harder, decreasing financial and personnel strains while improving student outcomes. Universal screening is an essential component of multi-tiered systems (Albers & Kettler, 2014; Kettler et al., 2014; Stoiber, 2014). However, there is limited research in universal screening measures and procedures, especially when screening for mathematical skill deficits (Clarke et al., 2014; Methe, 2009). There are not specific regulations regarding the technical adequacy of assessment measures used to make instructional decisions, but there are guidelines for best practices. Professional organizations such as the National Association of School Psychologists (NASP) and the American Psychological Association (APA) offer guidelines under ethical implementation of best practices (Jacob & Hartshorne, 2003). Practical application and ethical

issues require that more research is needed to improve the practice of universal screening in the area of mathematics, especially in the areas of decision rules and technical adequacy.

Math screeners demonstrate higher specificity than sensitivity indicating they are better at detecting students who will emerge with math difficulties as opposed to those who will develop adequate math skills (Fuchs et al., 2007). This suggests that when using Math-CBM (CBM-M) as a universal screening tool, the probability of false positives, or over identifying students as being at-risk, is higher than the probability of false negatives. This could be problematic when evaluating core curriculum. The core curriculum may appear to be ineffective for an inflated number of students. This also presents a problem when identifying students who are at-risk and in need of supplemental instruction or intervention. While universal screening instruments should err on the side of caution, over-identification of students who are in need of additional support may unduly burden school resources, including money and intervention personnel time.

Gersten and Jordan (2005) suggested there is research to validate screening instruments for later math difficulties, but there is limited research in early math skill screenings. More recent research has found a strong relationship between number sense skills in kindergarten and first grade, by administering CBM-M measures of early numeracy skills (Missall, Mercer, Martinez & Casebeer, 2012). Jordan, Glutting, Ramineni, and Watkins (2010) found early numeracy skills predict future math outcomes, indicating good predictive validity. There was a strong correlation between performance on a brief screen of number sense administered in kindergarten and first grade with math outcome while in third grade. Early numeracy or number sense skills require the understanding of numbers and relationship between numbers such as differentiating between number magnitudes, counting, and making sets of numbers (Jordan, Kaplan, Ramineni, & Locuniak, 2009).

After an extensive literature review, early numeracy skills and the extent in which they predict future math outcomes was again identified as an area of need by Gersten et al. (2012). Although there are no agreed upon general outcome measures in mathematics, there is a fair amount of literature correlating poor math fact retrieval fluency with math deficits in students who have been identified as low achieving, having a specific learning disability in math, and students with learning disabilities in both reading and math (Geary 2004; Geary et al., 2012; Jordan & Hanich, 2003; Martin et al., 2012). These deficits appear to be relatively stable in both elementary and secondary students (Jordan & Hanich, 2003; Martin et al, 2012).

Shapiro, Keller, Lutz, Santoro, and Hintze (2006) examined the predictive validity of CBM-M with PSSA performance and found a moderate relationship between students' performance on CBM-M in the winter of third grade with PSSA performance in the spring of third grade. These findings were further supported by Keller-Margulis, Shapiro, and Hintze (2008) who found evidence to suggest CBM-M is a valid predictive measure for student performance on PSSA when administered in first and second grade. There is some evidence to suggest word-problem solving measures are able to predict future math problem solving outcomes (Sisco-Taylor, Fung, & Swanson, 2015). However, reading and/or listening comprehension likely has a significant effect on student performance on these measures.

The proposed study will expand on current research by exploring the predictive validity of Monitoring Basic Skills Progress, Computation probe (MBSP-C) in the fall and winter of first, second, and third grades with PSSA-M administered in the spring of third grade. Knowing the extent to which MBSP-C predicts PSSA-M will facilitate more efficient and effective early intervention practices. The contribution of SES on PSSA-M achievement is examined given students in low-income homes have been identified as a population consistently under-

performing in mathematics. Sex as a predictive factor of PSSA-M is also studied in an attempt to provide some clarity among conflicting research whether or not a gender gap in mathematics continues to exist.

Research Question and Hypotheses

One broad research question was generated given the statement of the problem and problem significance previously reviewed: To what extent does a universal mathematics screening, MBSP-C in first, second, and third grade, sex, and SES predict math achievement as reported on the five subtests of the PSSA-M in third grade? It is hypothesized that MBSP-C scores in first, second, and third grade will be moderately predictive of math achievement as measured by the five subtests of the PSSA-M in third grade. Based on previous research, it is hypothesized the correlation between MBSP-C and PSSA-M scores will be moderate to strong. It is hypothesized that student performance in the fall of first grade will have the weakest correlation with PSSA performance and student performance in the spring of third grade will have the strongest correlation with third grade PSSA-M achievement due to time proximity between MBSP-C and PSSA-M administration. It is further hypothesized that SES will account for a significant amount of variance on PSSA-M achievement, with the potential to decrease the longer students are in a high quality educational setting. However, previous research has indicated students living in poverty are more resistant to improvement in mathematics instruction, so there is potential for the amount of variance accounted for by SES to remain the same or increase the longer a student is in an educational setting.

It is also hypothesized that MBSP-C will have the strongest correlation with the Numbers and Operations subtest of the PSSA-M. The Numbers and Operations subtest of the PSSA-M asks students to demonstrate an understanding of numbers, ways of representing numbers,

relationships among numbers and number systems, an understanding of the meanings of operations, use of operations and understanding how they relate to each other, the ability to compute accurately and fluently, and the capacity to make reasonable estimates. These skills closely resemble those assessed on the MBSP-C probes; therefore, it is predicted the strongest prediction of the MBSP-C will be to the Number and Operations subtest of the PSSA-M.

It is hypothesized that sex and resource availability will have a moderate association with math achievement, based on highlights from the 2007 TIMSS (Gonzales et al., 2009). A secondary hypothesis is offered that if a gender gap is present, it will be among high achieving students as opposed to low performing students (Stoet & Geary, 2013). It is hypothesized that SES and sex will not have a significant interaction with each other.

Definition of Terms

The following definitions are provided to clarify the very specific way terms are used within the context of this study. Terms are defined to ensure uniformity of meaning throughout this study.

Formative Assessment

Formative assessments are instruments which attempt to capture data regarding student learning of a particular skill or behavior at a specific point in time to inform instructional practices (Burns, 2010). Appropriate formative assessment instruments should be able to measure growth over time, be instructionally diverse, and practical to administer. Formative assessment instruments need to demonstrate technical adequacy and treatment sensitivity (Clarke & Shinn, 2004; Fuchs & Fuchs, 1999; Stecker & Fuchs, 2000). CBM are frequently given as examples of formative assessment.

Curriculum Based Measures

CBM are defined as a "set of standardized and specific measurement procedures that can be used to quantify student performance in the basic skill areas of reading, spelling, mathematics computation, and written expression" (Hintze, Christ, & Methe, 2006, p. 51). CBM are used to assess students on content directly related to the curriculum. CBM are frequently used for progress monitoring student learning and as universal screening instruments. Teachers can use data from CBM to inform instructional decisions.

Summative Assessments

Summative assessments are used to assess student learning after instruction. Summative assessments evaluate the extent to which a student has learned what was taught. An example of a summative assessment is a state-mandated academic achievement test administered in the spring of the school year, assessing students on the core standards they were expected to learn while in that particular grade.

High Stakes Testing

High stakes testing is the term given to assessments that have significant consequences tied to test results. State-mandated achievement tests are often considered a high stake test (Braden & Schroeder, 2004).

Specific Learning Disability

IDEIA (2004) recognizes 13 disability categories for which students can receive specially-designed instruction through an Individualized Education Program (IEP). Specific learning disability (SLD) is one of the disability categories recognized by federal law. Eligibility for a SLD considers underachievement, evidence of learning difficulties and exclusions of other factors which may impact learning such an intellectual disability (Lichtenstein, 2014). Eligibility

for a specific learning disability is determined by a multi-disciplinary evaluation. Parental consent for a multi-disciplinary evaluation is required. Students with a specific learning disability typically demonstrate poor academic achievement in one or more area that cannot be explained by socio-economic status, limited opportunity to learn, English as a second language, an intellectual disability, truancy, and/or an emotional disturbance (Huefner, 2006). Once a student is found to be eligible and in need of specially designed instruction an IEP is developed. Students are re-evaluated a minimum at least once every three years to determine if they still meet criteria of a SLD and are in need of specially designed instruction. Approximately 5 – 8% of the student population is thought to have a SLD in mathematics (Clarke et al., 2014).

Individualized Education Program

An IEP is required under IDEIA for any student who meets criteria for one of the 13 disability categories and demonstrates a need for special education services. The purpose of the IEP is to outline goals, specially-designed instruction, and secure resources for the benefit of the student. The IEP is developed by a student's educational team which includes parents, general education teachers, special education teachers and specialists, a representative from the local education agency, and depending on age, the student. IEPs are reviewed and revised a minimum of one time per calendar year, but should be fluid in nature to reflect the changing needs of the student (Huefner, 2006).

Resource Availability

Resource availability is defined as the economic contributing factors of the child as reported by whether or not the child qualifies for the school's free or reduced meal program. Resource availability is used as a measure of socio-economic status.

Early Intervention

Early intervention services are composed of four key principles. First, early intervention requires effective screening instruments to identify students who may be at risk for future academic, social, emotional, or behavioral problems. Second, once these students are identified, the intensity of instruction changes through additional intervention and/or differentiation in the classroom. Third, the student's response to intervention is progress monitored to determine growth over time (Clarke, Baker, Smolkowski, & Chard, 2008). Lastly, progress monitoring data, along with other available formative and summative assessments are used to inform instructional decisions regarding whether the intervention should be modified, continued, or discontinued.

Multi-Tiered Systems of Support

MTSS are problem-solving systems of service delivery in schools focused on improving school outcomes for all students through high quality and differentiated instruction. The intensity of instruction is matched to a student's learning needs. Key features of MTSS systems include evidence-based practices, data-based decision making, proactive early intervention of academic and behavior deficits, universal screening, coordination of intervention and intensity of instruction based on student need, and fluidity of programming based on student need (Stoiber, 2014).

Response to Intervention

RTI is method of determining whether or not a student is in need of special education services. In an RTI model, eligibility and decisions regarding educational programming are made based on a student's response to an intervention that is matched to identified skill deficits, is research-based, and is implemented with integrity. RTI models are supported by legislation,

and research prioritizes the need for an effective core curriculum and early identification of students in need of additional instruction or intervention. RTI is the preferred method to determine SLD in a MTSS model (Kovaleski, VanDerHeyden, & Shapiro, 2013).

Gated Evaluation System

A gated evaluation system, also referred to as a multiple-gate model, is an assessment model for identifying students at risk for school failure (Albers & Kettler, 2014; VanDerHeyden, 2010). A different instrument is used at each stage, therefore eliminating students who were not identified as being at-risk. An example of a gated evaluation system would be an entire grade completing an oral reading probe (i.e., Gate 1). Students who perform below the 20th percentile on the Gate 1 assessment would then be administered a Gate 2 assessment, for example, a reading comprehension CBM. Students who perform below a pre-established percentile rank on the Gate 2 measure would then complete a more comprehensive measure, such as CORE's Multiple Measures Reading Assessment (Arena Press, 2008), which not only identifies below grade level skills, but identifies specific areas of deficit for intervention purposes. While more research is needed, initial findings support gated evaluation procedures as an accurate method for identifying students at risk (Albers & Kettler, 2014; Fuchs, Compton, et al., 2011; VanDerHeyden, 2010).

Universal Screening

Universal screening is "the systematic assessment of all children within a given class, grade, school building, or school district, on academic and/or social emotional indicators that the school personnel and community have agreed are important" (Ikeda, Neessen, & Witt, 2008, p. 103). Data from universal screenings are used to identify students who would benefit from

additional intervention and to evaluate the effectiveness of an educational system, curriculum, or program (Kettler et al., 2014).

Predictive Validity

Predictive validity is a type of criterion validity, or the degree to which an instrument accurately projects a specific outcome. Within the context of MTSS, predictive validity is the extent to which an instrument can predict a future skill or behavior (Jackson, 2003; Leary, 2001).

Outcome Criterion

The outcome criterion is the term used to describe a future measure that serves as the index to which all tests are compared, such as state-mandated tests of academic achievement and nationally normed academic achievement tests (VanDerHeyden, 2010). Outcome criterion are also referred to as predictor criterion. The outcome criterion used in this study is the PSSA-M.

Classification Accuracy

Classification accuracy is the extent screeners can accurately identify students based on their performance on both the screener and some external criterion (Gersten et al., 2012; VanDerHeyden, 2010, 2011). There are four outcomes of classification accuracy, (a) true positive, (b) true negative, (c) false positive, and (d) false negative. Screening tools should maximize sensitivity and specificity by setting accurate cut-scores (Kamphaus, Reynolds, & Dever, 2014). Classification accuracy was founded in the medical field, were a positive result indicates the presence of a problem and a negative result indicates no problem or typical outcomes.

Sensitivity

Sensitivity refers to the portion of true positives correctly identified by a test or instrument (VanDerHeyden, 2011). Sensitivity is a measure of a test's power to correct identify the presence of a deficit.

Specificity

Specificity refers to the proportion of true negatives detected by a test or instrument (VanDerHeyden, 2011). It is a measure of a test's power to correctly identify the absence of a problem.

True Positive

A true positive outcome indicates the predictor instrument correctly detected the presence of a problem, meaning the predictor instrument and the criterion measure are in agreement (Christ & Nelson, 2014). In relation to universal screening, a true positive means the universal screening data correctly identified students as being at-risk for academic, social, or behavioral deficits. Within the context of the present study, a true positive means the predictor instrument correctly indicates that a math deficit is present.

True Negative

A true negative is when a predictor and criterion measure agree a problem is not present (Christ & Nelson, 2014). In other words, a true negative is when universal screening instrument correctly identifies when a disease is not present. Within the context of the present study, a true negative means the predictor instrument correctly indicates that a math deficit is not present.

False positive

A false positive is when the predictor incorrectly identifies the presence of a problem. The predictor measure indicates a problem that is not confirmed on a criterion instrument,

resulting in an inaccurate classification (Christ & Nelson, 2014). A false positive on universal screening data means the screener incorrectly suggests the presence of a problem when no problem is present. Within the context of this study, a false positive indicates that the predictor instrument incorrectly indicates that a math deficit is present.

False Negative

A false negative is when the predictor fails to identify that a disease or condition is present (Christ & Nelson, 2014). An example of a false negative is when universal screening does not identify a problem when a problem is, in fact, present. Within the context of this study, a false negative occurs when the predictor instrument incorrectly indicates that a math deficit is not present. This is considered the most problematic classification error because students are not identified to receive intervention necessary to remediate areas of need.

Assumptions

There are several assumptions made regarding the instruments and procedures used in this dissertation. It is the assumption of this study that MBSP-C meets the requirements to be used as a universal screening measure within the sample school district. To meet the criteria of an appropriate universal screening the instrument should be able to (a) identify potential academic and/or behavioral concerns at a system and individual level, (b) provide information related to how students are responding to core instruction, (c) provide data to teachers to inform instruction and align instructional resources, and (d) demonstrate adequate reliability and validity (Albers & Kettler, 2014; Kettler et al., 2014).

It is assumed MBSP-C and PSSA-M meet all professional standards, generate tests scores that distinguish among students, are able to identify students performing above and below the expected level of achievement, are composed of items sensitive to change in performance, and

have had sources of bias been eliminated. It is further assumed adequate training is given to educators in scoring, administration, and interpretation of test results for both MBSP-C and PSSA-M (Ikeda et al., 2008). It is assumed MBSP-C were developed in accordance with universal design. Universal design is the ability of an assessment instruments to be accessible and sensitive to the characteristics of all potential test takers (Anderson et al., 2011).

Universal screening is not used with select students, but is administered individually or in a group to an entire classroom, grade, school, or district. It is the assumption of this study that MBSP-C was administered to all students attending the three elementary schools in the sample population.

It is assumed that all screenings and state assessments were administered and scored in accordance with standardized procedures. Due to the use of archival data, protocols to ensure standardized administration of assessments could not be implemented. However, administration of the math screening measure was completed by the same person using scripted standardized directions. Standardized administration procedures of MBSP-C were overseen by two building principals who lead the math curriculum committee. PSSA-M was administered by classroom teachers, school counselors, reading specialists, intervention support teachers, and learning support teachers. Classroom teachers and school staff who proctor the state achievement test have to complete yearly training to ensure correct testing procedures are followed. Administration of PSSA-M was monitored by building principals. Administration procedures are monitored by state compliance officers who conduct random unscheduled visits to schools during the PSSA testing window.

The math curriculum taught to the sample population is assumed to be similar to other school districts within Pennsylvania, as they were in the process of adopting PA Common Core

standards. Secondly, it is assumed students received appropriate instruction provided by qualified teaching professionals.

Limitations

Limitations to this study stem primarily from the use of a convenience sample and archival data. Additional limitations are variations in the implementation of math curriculums and technical adequacy of universal screening instruments.

Limitations to this study include the use of a convenience sample. All data were gathered from a rural school district in Pennsylvania. This sample population may not be representative of the population as a whole, which limits generalizability of the findings to more diverse student populations. This study is also limited by the screening measure used within this school district. MBSP-C was the only mathematical screening instrument used, therefore, there is no basis for comparison between it and another universal screening instrument.

The use of archival data created several limitations. Standardized administration of MBSP-C is assumed but not guaranteed. Secondly, the school district did not retain completed MBSP-C probe sheets, therefore, scoring accuracy and correct data entry into the data warehouse cannot be verified.

The large amount of variability that occurs naturally within math curriculums from school district to school district is another limitation as is variation in quality of instruction from classroom to classroom (Davison & Mitchell, 2008; Kelley, 2008). Use of archival data did not allow for quality control checks regarding instruction and accurate implementation of the district's math curriculum. This limitation is minimized by the adoption of the PA Common Core, which standardized required instructional content and curriculum expectations. Math screeners demonstrate higher specificity than sensitivity indicating they are better at detecting

students who will emerge with math difficulties than those who will develop adequate math skills (Fuchs et al., 2007). This suggests when using CBM-M as a universal screening instrument, the probability of a false positive is high. Given the primary function of universal screening is to identify students who are at-risk for deficits, over identification rather than under identification of students for tiered support is preferred.

Summary

This chapter provides a brief historical context for this study, specifically how MTSS require the utilization of universal screeners to identify students in need of additional instruction and intervention beyond the core curriculum. Fundamental considerations of universal screeners were noted with an emphasis placed on empirically-validating the predictive strength of math screeners to appropriately identify students at risk for failing high-stakes testing outcomes. Lastly, the research question, definition of terms, assumptions, and study limitations were outlined and discussed.

CHAPTER II

REVIEW OF RELATED LITERATURE

Math Domains

The development of mathematical skills is so complex that experts in the field have not been able to generate clearly defined general outcome measures and subsequently what skill deficits constitute a math disability. Therefore, in order to give meaning to the constructs frequently assessed in math, it is necessary to review the development of mathematical skills and mathematical knowledge domains.

While mathematics processes are understudied, with many unanswered questions (Fisher, Doctoroff, Dobbs-Oats, & Arnold, 2012; Kelley, 2008; Mazzocco, 2003), mathematical proficiency is currently thought to be based on four mathematical knowledge domains that are used to guide mathematical instruction, (a) conceptual knowledge, (b) strategic or procedural knowledge, (c) factual or declarative knowledge, and (d) problem-solving skills or application knowledge (Kelley, 2008). These domains are also referred to as instructional domains and, according to the National Council of Teachers of Mathematics (NCTM), should each be represented in the content areas of numbers and operations, algebra, geometry, measurement, data analysis, and probability.

Conceptual knowledge is defined as, "a deep understanding that allows categorization of examples from non-examples and the critical attributes from the noncritical attributes of the concept" (Kelley, 2008, p. 421). Conceptual knowledge is considered the primary goal of mathematics (Hudson & Miller, 2006; Miller & Hudson, 2007). It is believed to be a necessary skill to progress through mathematics curriculum, apply skills in other content areas, and generalize to real-world applications. It is also important to note that conceptual knowledge

encompasses knowing definitions, rules, routines/problem-solving procedures of a concept (Kelley, 2008; Miller & Hudson, 2006).

Strategic or procedural knowledge is the sequential steps used to do something, in this context, solve a math problem (Hudson & Miller, 2007; Kelley, 2008; National Research Council, 2002). These skills are required to solve computation problems and apply math skills outside of the classroom.

Factual or declarative knowledge describes all the basic facts required to solve mathematical problems (Kelley, 2008; Miller & Hudson, 2007; National Research Council, 2002). These include basic math computation facts, terminology, counting, number identification, and symbol identification (Kelley, 2008). Factual knowledge skills should be taught to mastery to allow for automaticity when utilizing this information to problem solve (Kelley, 2008; Miller & Hudson, 2007).

Application knowledge or problem-solving skills are viewed by some mathematics instructors as a knowledge domain, but others argue it is the application of the three knowledge domains described above (Kelley, 2008; National Research Council, 2002). Application knowledge is required to apply mathematics skills across settings. It can be described as the component of math instruction that makes it functional. It is recommended application knowledge be explicitly taught and embedded in instruction when developing conceptual knowledge, strategic knowledge, and factual knowledge (Kelley, 2008).

While not considered one of the mathematical knowledge domains, the National Research Council (2002) also recognizes Engaging as an important feature of mathematical learning. Engaging is defined as, "Seeing mathematics as sensible, useful, and doable – if you work at it – and being able to do the work." (National Research Council, 2002, p. 9). It is

important to include Engaging as a component of high-quality math instruction given attitude toward math significantly effects mathematical learning outcomes. Students need to develop the skills to approach mathematical learning in an engaging manner (McGraw et al., 2006; PCAST, 2011).

The four mathematical knowledge domains guide formal mathematical instruction students receive while in school. However, the development of mathematical understanding continues to develop and evolve across the life span.

Development of Mathematical Skills

Three primary neural pathways have been identified in mathematical processing and development, (a) linguistics, (b) spatial attention, and (c) quantitative (Dehaene, Molko, Cohen, & Wilson, 2004; Dehaene, Spelke, Pinel, Staneseu, & Tsivkin, 1999; LeFevre et al., 2010). Figure 1 represents the three primary pathways involved in the development of math skills and corresponding early numeracy skills and mathematical outcomes. As depicted in Figure 1, cognitive skills associated with mathematical learning offer both shared and independent contributions to the development of early numeracy skills and later mathematical outcomes.

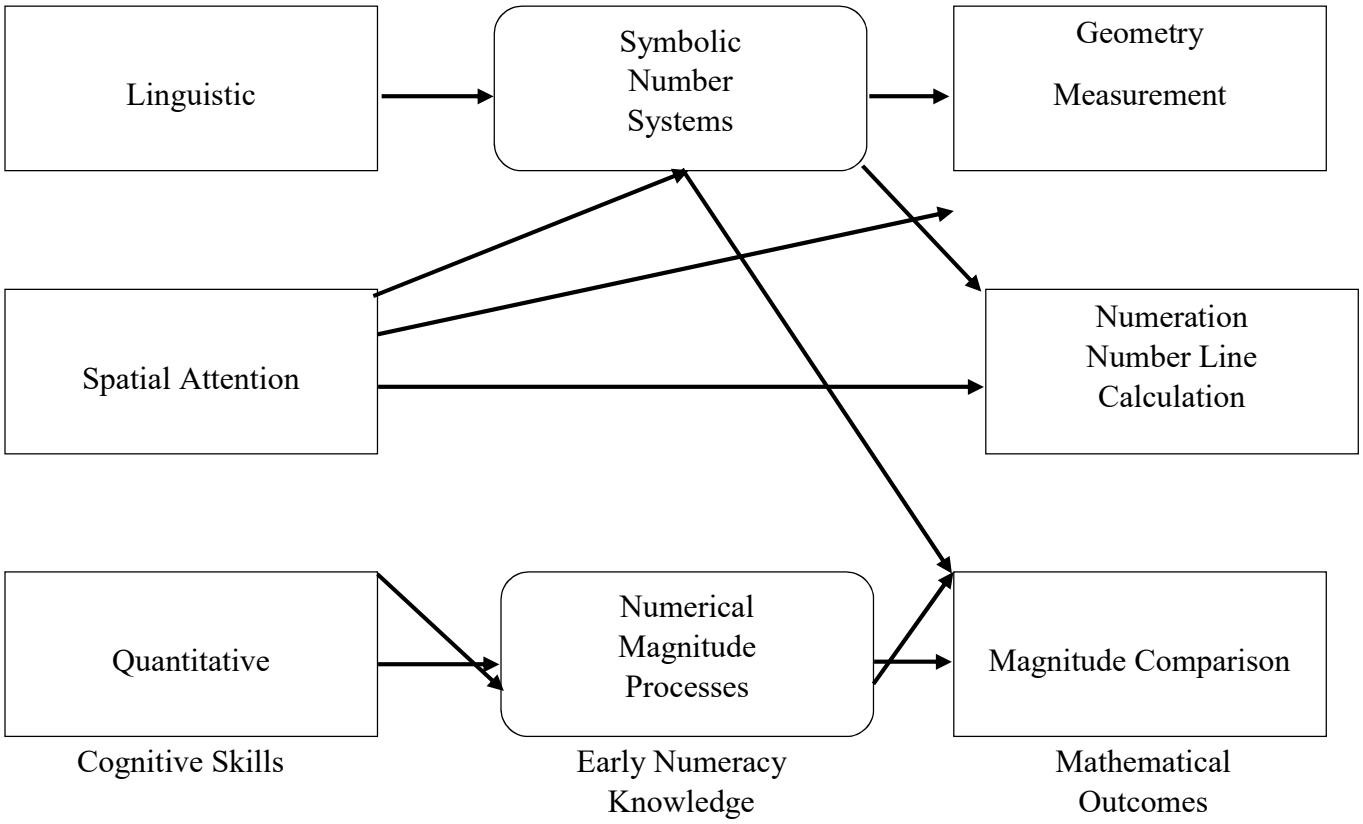


Figure 1. Three pathways model of mathematical development with corresponding early numeracy knowledge and mathematical outcomes. Adapted from "Pathways to Mathematics: Longitudinal Predictors of Performance," by J.-A. LeFevre, L. Fast, S.-L. Skwarcuk, B. L. Smith-Chant, J. Bisanz, D. Kamawar, and M. Penner-Wilger, 2010, *Child Development, 81,* p. 1755. Copyright 2010 by Society for Research in Child Development, Inc. Reprinted with permission.

Mathematical Development and Linguistics

The linguistic pathway encompasses general language and language processing skills, such as phonological awareness, vocabulary, verbal reasoning, and listening comprehension. The linguistic pathway is thought to be the strongest and most consistent predictor of early numeracy skills (LeFevre et al., 2010; Purpura & Reid, 2016; Vukovic & Lesaux, 2013). LeFevre et al. (2010) found measures of linguistic accounted for a significant portion of variance in elision, vocabulary, and number naming.

Purpura and Reid (2016) expanded on the existing research connecting linguistics to mathematical learning outcomes. The authors explored whether or not individual differences in mathematical language were a stronger predictor of numeracy skills than differences in general language. A secondary research question investigated group differences in mathematical language performance based on parental education and age. Mathematical language or math-specific language uses vocabulary that is content-specific and required to understand mathematical tasks.

In mathematics, vocabulary is frequently quantitative and spatial in nature. Quantitative language includes words to describe quantities and make comparisons between numbers (e.g., more, less, many, and fewer). Spatial vocabulary refers to words used to talk about relationships between numbers and physical objects (e.g., before, over, above, near, and far). The findings of Purpura and Reid (2016) support previous research that linguistic skills are a strong predictor of numeracy skills in children. The authors extended beyond previous research and determined mathematical language was a much stronger predictor of numeracy skills than general language. Mathematical language was such a strong predictor of numeracy skills that when added to a regression model, general language skills was no longer a significant predictor of numeracy skills (Purpura & Reid, 2016).

There is a strong positive correlation between parent education level and linguistic development in children (Chu, vanMarle, & Geary, 2015; Lehrl, Kluczniok, & Rossbach, 2016). Children with at least one college-educated parent performed higher on mathematical measures than children in families who were not college-educated. This is attributed to an increased use and exposure to mathematical language for children who had at least one college-educated parent (Purpura & Reid, 2016).

A similar topic was explored by Vukovic and Lesaux (2013). The authors examined the impact of language development on mathematical skills, longitudinally from first through fourth grade, in native English speakers and language-minority learners. The authors also controlled for visual-spatial skills, sex, and socio-economic status (SES). They concluded that language ability was predictive of data analysis/probability and geometry. Language ability was not found to be predictive of arithmetic or algebra. These finding were consistent in native English speakers and in English language learners, which implicates language in the acquisition of mathematical cognition and understanding. However, language did not influence mathematical learning with Arabic or abstract symbols used in arithmetic and algebra (Vukovic & Lesaux, 2013).

Based on this research, it can be concluded early language experiences are key in the development of mathematical understanding, regardless of language background. Children should be exposed to language and play that is rich in quantitative and spatial vocabulary. This is especially important for children in low-income families, as SES is a significant indicator in the development of mathematical language which affects future mathematical learning outcomes (LeFevre et al., 2010; Lehrl et al., 2016; Purpura & Reid, 2016; Vukovic & Lesaux, 2013).

Mathematical Development and Spatial Attention

Spatial attention is typically defined and measured as a function of working memory. Working memory involves the simultaneous storage and processing of information (Geary, Hoard, & Bailey, 2012; Toll & Van Luit, 2014). It is comprised of verbal working memory and visual-spatial working memory. Verbal working memory involves the phonological loop which processes and rehearses short-term verbal information. Visual-spatial working memory involves the visuo-spatial sketchpad, which processes and rehearses visual and spatial information

(Lukowski et al., 2014). Visual-spatial working memory is the "storing and processing of information related to shape, color, brightness, and static visual layout properties" (De Santana & Galera, 2014, p. 399). Both verbal and visual-spatial working memory have been implicated in the development of early numeracy skills.

Spatial attention is related to the development of number naming and numerical magnitude in children (LeFevre et al., 2010). The general attentional processes of working memory, or the ability to hold information in short-term memory in order to process it, is necessary to complete the complex and multi-step requirements of many mathematical tasks (Geary et al., 2012; LeFevre et al., 2010; Toll & Van Luit, 2014). Verbal working memory is a significant predictor of mathematical tasks that require more advanced problem solving and computation, such as word problems. Verbal working memory, however, is not predictive of relatively simple mathematical tasks, such as basic computation (Lukowski et al., 2014). Visual working memory is predictive of arithmetic reasoning, number writing, and symbolic magnitude (Lukowski et al., 2014; Toll & Van Luit, 2014). Visual working memory is heavily used when learning a novel mathematical task whereas verbal working memory is utilized once the learning has been established (Toll & Van Luit, 2014).

Mathematical Development and Quantitative Knowledge

Quantitative knowledge is an understanding of quantities and numbers. Quantitative knowledge is required for magnitude discrimination and magnitude comparisons, or the ability to distinguish between amounts. The ability to discriminate between magnitudes develops rapidly throughout infancy (Landerl & Kölle, 2009). Quantitative knowledge has been observed in infants as young as 6-months-old (Wynn, 1995). It is thought to be the core knowledge base underlying all other mathematical processes and development (Landerl & Kölle, 2009); however,

LeFevre et al. (2010) found quantitative knowledge to be the least predictive of future math outcomes in comparison to the linguistic and spatial attention pathways.

While more research is needed, initial findings indicate a breakdown in one or more of these precursor cognitive skills could be the etiology of innate mathematical learning disabilities, or math deficits that cannot be explained by poor instruction or environmental factors. It is important to note Lukowski et al. (2014) found both genetic and environmental factors played a significant role in the development of mathematical skills. Therefore, it is important to account for both genetic and environment factors when identifying and remediating mathematical deficits. These factors are not included in the scope of the present study due to the use of archival data.

Early Numeracy Skills

Children enter school with varied skill levels, but the majority of students have some understanding of numbers and number concepts. This is referred to as early numeracy skills or number sense. Similar to early literacy skills, play and observation promote development of math competencies long before children are school-aged (Landerl & Kölle, 2009; Mazzocco, 2003; Purpura & Reid, 2016). The innate, nonverbal ability to understand non-symbolic quantities is termed number sense (Price & Fuchs, 2016; Toll & Van Luit, 2014). There is some evidence suggesting that a child's understanding of magnitude begins in infancy and number sense is "a basic capacity of the human brain" (Dehaene, Molko, Cohen, & Wilson, 2004, p. 218; Landerl & Kölle, 2009). Wynn (1995) found infants were able to differentiate between small magnitudes of dots, points of light, sound, physical action, and household items by six months of age. These innate skills generally facilitate the learning and development of early numeracy

skills that are considered pre-requisites to the development of the four domains of mathematical understanding.

Early numeracy skills are a series of pre-requisite mathematical skills necessary for more complex mathematical understanding and problem solving. Early numeracy skills include verbal counting, recognition of number symbols and quantities, distinguishing between number patterns, comparing magnitudes, and estimating quantities (Fuchs, Fuchs, & Compton, 2012; Mazzocco & Thompson, 2005; Toll & Van Luit, 2014; VanDerHeyden & Burns, 2009). Development of mathematical skills is complex and affected by many non-mathematical factors such as language, parental education, SES, executive functions, and intelligence (Bailey, Watts, Littlefield, & Geary, 2014; Chu et al., 2015; LeFevre et al., 2010). Most children transition from innate number sense to the development of early numeracy skills through informal play in both the home and pre-school environment. Early numeracy skills are typically established during pre-school years (3-5 years old). Differences and deficits in early numeracy skills can be detected by 5 years of age (Toll & Van Luit, 2014).

Mathematical Development from Childhood to Adulthood

Young children develop early numeracy skills by expanding on innate number sense skills. As children progress through school into adulthood, their mathematical thinking is thought to transition from formal procedural thinking to more abstract thinking or application of formal knowledge. This is challenged by some theories of mathematical development in which the opposite is proposed. Tall (2008) and Braithwaite, Goldstone, van der Mass, and Landy (2016) propose young children are more abstract in their thinking and mathematical processes become increasingly formal the more one is exposed to mathematical information. These theories rationalize that the more rehearsed a skill is, the more automatic it becomes, which in

turn decreases the use of abstract thought. Both theories of development have support in the research. From an educational perspective, the outcome is the same: Children require explicit instruction in math concepts, procedure, and increased opportunity to apply formal procedural knowledge in order to develop problem-solving skills as represented in the four domains of mathematical knowledge (conceptual knowledge, strategic or procedural knowledge, factual knowledge, and application knowledge [Braithwaite et al., 2016; Kelley 2008; Tall, 2008]). The development of mathematical literacy is dependent upon the development of each inter-connected knowledge domain. Due to the interdependent nature of the domains, difficulty with one or more of them is likely to impact mathematical learning and future outcomes.

Math Learning Disabilities

Historically, educational researchers and practitioners struggled to determine whether performance deficits are due to a lack of effective instruction, a true learning disability, or a combination of the two. Differentiating between a learning disability and deficits due to poor instruction is especially difficult in ineffective educational systems, which can result in over-identification of specific learning disabilities.

Approximately 5% to 8% of children are believed to have some form of a mathematical learning disability (MLD; Geary, 2004; Shin & Bryant; 2015). The complexity of math development has led to inconsistent definitions of what constitutes a math learning disability, which, in turn, affects how to best identify a math disability (Mazzocco, 2003). Some of the most notable challenges to the identification of a learning disability are the non-sequential development of math skills, wide ranges of math curriculums, and inconsistencies in math instruction.

The research reviewed below provides information regarding frequent characteristics of math disabilities, potential academic and cognitive indictors of a MLD, and the stability of MLD over time. These all support the necessity of early intervention of math deficits through universal screening practices.

Characteristics of MLD

The components of mathematical understanding are not learned in isolation. Features required for mathematical proficiency are interdependent on each other and interwoven; one area of deficit can have a cascading effect on learning and achievement (National Research Council, 2002). MLDs can present as deficits in one or more of the mathematical domains or as one or many individual skills within a single domain (Geary, 2004). For example, many children with MLD demonstrate average number processing skills in isolation (factual knowledge), but demonstrate frequent, persistent errors when applying these skills to complete mathematical processes, i.e., arithmetic (strategic or procedural knowledge). These students may not be identified as at-risk on early numeracy skill universal screening instruments; however, deficits in complete mathematical processes are more apparent on timed assessments, which require automaticity (Geary, 2012).

When comparing differences in performance between students with MLD and non-learning disabled (NLD) peers (based on age and/or grade normative data), Shin and Bryant (2015) found NLD peers outperformed students with MLD in the areas of mathematical calculations, word problem solving, arithmetic strategies, and number sense skills at both the elementary and secondary level. The most sizable group differences were observed on measures of mathematical calculation and arithmetic strategies. Students with MLD had significantly weaker counting strategies than NLD students when engaging in problem solving. Of significant

note, the authors found both groups, those with and without a MLD, struggled when asked to prove problems as a means of assessing mathematical reasoning. This suggests a need for improved instruction to facilitate the development of mathematical reasoning skills for all students. Geometry was another area of relative weakness for both peer groups with and without MLD (Shin & Bryant, 2015).

Geary (2004) found that students with MLDs frequently had difficulty with retrieval of basic arithmetic facts from long-term memory. This research was furthered by Geary, Hoard, and Bailey (2012). The researchers compared fact retrieval fluency in children identified as low achievers in mathematics and those identified as a having a specific learning disability in mathematics. Over the course of the three-year longitudinal study, the researchers compared fact retrieval deficits in students identified as having a specific learning disability in mathematics and students identified as low achieving in mathematics. Students with MLDs were defined as those who demonstrated achievement below the 10^{th} percentile in mathematics on standardized tests of achievement for multiple years who also demonstrated low average reading, working memory, and general IQ. Students identified as low achieving were described as having average reading achievement, IQ, and working memory, while their performance on standardized mathematics achievement tests fell within the 10^{th} and 25^{th} percentile. The low achieving (LA) students were separated into two subgroups, LA-mild fact retrieval and LA-severe fact retrieval. MLD and LA students were compared and contrasted with each other and with typically achieving peers. The fact retrieval of 231 students were studied over the course of three years, from kindergarten through third grade.

The findings of this study suggest that fact retrieval deficits are persistent in both children with MLDs and LA-severe fact retrieval, supporting the use of a fact-based measure as a

universal screening tool. Geary et al. (2012) noted the deficits of children identified as LA-severe fact retrieval were not accurately represented on untimed standardized measures. It is hypothesized that because children with mathematical difficulties do not demonstrate automaticity of basic math facts and/or possess adequate conceptual understanding, they require more time to problem solve, which significantly impacts performance on timed measures (Fuchs et al., 2005; Geary et al., 2012; Shin & Bryant, 2015). These findings support the use of a timed measure for the purpose of universal screening.

Namkung and Fuchs (2012) investigated potential differences in early numerical competencies for students who demonstrate deficits in either computation or word problem-solving in an attempt to develop early identification of math disability subtypes. The researchers compared early numerical competencies or early numeracy skills of numerical magnitude, counting knowledge, numerical value of small quantities, between second grade students with computational difficulty, word problem solving difficulty, students demonstrating difficulty in both areas, and children demonstrating typical development of math skills.

Namkung and Fuchs (2012) found that typically-developing students outperformed students with computational difficulty, word problem solving difficulty, and students with skill deficits in both computation and word problem solving on measures of precise representation of small quantities and large magnitudes. Students categorized as having computational difficulty or word problem solving difficulty performed comparably, with both groups outperforming students who demonstrated deficits in both math skills. On assessments of counting knowledge, students with computational or word problem solving difficulties performed comparably to students with typical development. All three groups outperformed students who demonstrated both computational and word problem solving difficulty. No discernable differences between

subtypes of math disabilities were identified until a subtraction probe was administered (Numkung & Fuchs, 2012). Students categorized as demonstrating typical math development once again outperformed all three subgroups of students with math deficits. On the subtraction measure, students with computational difficulties outperformed students with word problem solving difficulties and those with computational and word problem solving difficulties. This suggests differences between young students with deficits on computational tasks and those with deficits in solving word problems or deficits in both areas are not captured with measures of early numerical competencies. The findings of this study provide support for use of multiple-skill measures as universal screening instruments because they are able to identify students who may have one or more mathematical deficits who would not be detected as at-risk with a single-skill computation universal screening measure.

Cognitive deficits characteristic of MLD. In addition to low mathematical achievement, students with mathematics disabilities frequently demonstrate cognitive deficits (Geary, 2004; Murphy et al., 2007). Students with MLDs demonstrate cognitive deficits in one or a combination of executive functioning, which encompasses working memory, information representation and manipulation in the language system, and visual-spatial systems (Geary, 2004).

Toll and Van Luit (2014) explored cognitive deficits which may indicate poor early numeracy skills in young children ($N = 990$, M age = 4.55). The authors found verbal working memory, symbolic comparison, and math language correlated with deficits in early numeracy skills. Children with poor numeracy skills at the start of that study demonstrated significantly high rates of growth once receiving formal instruction throughout their kindergarten year (Toll & Van Luit, 2014).

When evaluating predictive factors through multiple linear regression, the researchers found reading, inattention, broad computation, and specific skill fluency tasks (initial score only) to be uniquely predictive. However, basic math calculation skills, such as those assessed with Monitoring Basic Skills Progress, Computation (MBSP-C) probes, were found to be a better predictor of mathematical problem solving skills than cognitive factors that impact math performance including visual spatial processing, fluid reasoning, working memory, processing speed, crystalized intelligence, auditory processing, and long-term retrieval (Decker & Roberts, 2015). This research indicates that while students with MLD frequently demonstrate deficits in one or more areas of executive functioning including working memory, information representation and manipulation in the language system, and visual-spatial systems, instruments that detect cognitive deficits are not effective universal screeners for mathematical deficits.

Distinguishing MLD from co-morbid reading disabilities. Much attention has been given to the strong correlation between reading skills and academic achievement in mathematics. Research has indicated reading universal screening measures are strong predictors of math deficits, especially when students have comorbid reading and math disabilities (Codding, Petscher, & Truckenmiller, 2015; Toll & Van Luit, 2014). Therefore, it is important to explore measures that can differentiate between reading deficits and mathematical deficits.

Computation deficits are prevalent in both students who are low-achieving and those identified as having a MLD; however, computation deficits are not prevalent in students who demonstrate typical achievement or those with a reading disability (Geary, 2004; Shin & Bryant, 2015). This suggests that computation fluency measures are better able to distinguish between students with MLD and those with stand-alone or comorbid reading disability with more accuracy than other measures such as word problem measures. Subsequent research reviewed in

this section provides support that math CBM are more predictive of future math outcomes than reading measures. Math measures are also able to better distinguish between students with deficits solely in mathematics and those with co-morbid reading deficits.

Jordan, Hanich, and Kaplan (2003) conducted a longitudinal study to investigate mathematical competencies of students with mathematical difficulties verses students with comorbid reading and mathematical difficulties. Participants included 190 students who were followed from 2nd to 3rd grade. The students were identified as belonging to one of the following groups: math difficulty only (MD), math and reading difficulties (MD-RD), reading only difficulties (RD), or normal achievement (NA). The authors found similar rates of growth across all four groups. Students with both reading and mathematical difficulties were consistently outperformed by students in the NA and RD groups. Students with MD performed higher than MD-RD on measures of math problem solving, but not in the area of calculation. Of particular interest to the current study, calculation fluency and fact mastery deficiencies were persistent in both the MD and MD-RD groups. This supports calculation fluency and instruments requiring automaticity of basic arithmetic skills be considered as general outcome measures.

A second thorough review of the literature over a decade later generated similar findings regarding the importance of math fact automaticity with future math outcomes. Shin and Bryant (2015) synthesized the literature regarding the mathematical and cognitive performances of students with math difficulties and those with comorbid reading difficulties. The authors searched for all published articles from 1975 to 2011 that focused on the relationship between mathematics and cognitive functioning in students with MLD. The initial search yielded 538 studies, 105 of which were selected for more thorough review. Of the 105 reviewed, 23 met inclusionary criteria. The authors focused on studies which compared the performance of MLD

to students with MLD/RLD and students with no learning disabilities (NLD). Based on this literature review, there were no significant differences between MLD and MLD/RLD students on measures of mathematical calculation. However, students with MLD performed significantly higher than those with both a MLD/RLD on word problems and arithmetic fact strategy.

Fact retrieval fluency deficits were pervasive and persistent in both elementary and secondary students with a MLD and MLD/RLD (Geary, 2004; Shin & Bryant, 2015). Students with MLD demonstrated significantly better scores on instruments assessing completion of word problems than those with reading and math deficits. However, the authors found insignificant differences between the two student groups on instruments assessing mathematical calculation skills. This suggests the need to focus on skills such as basic fact fluency and approaches to solving math word problems (Shin & Bryant, 2015).

Both studies support the role of math computation knowledge and automaticity as strong predictors of future math outcomes. This, in turn, supports the use of a computation-based fluency measure for the purpose of universally screening students who would benefit from additional math intervention. Computation deficits are prevalent in both students who are low-achieving and those identified as having a MLD; however, computational deficits are not prevalent in students who demonstrate typical achievement or those with a reading disability. This suggests that computation fluency measures are better able to distinguish between students with MLD and those with independent or comorbid reading disabilities with more accuracy that other instruments such as word problem screeners.

Stability of MLD. After second grade, math proficiency is relatively stable, supporting the necessity of early detection and intervention of math deficits (Jordan & Hanich, 2003; Jordan, Hanich, & Kaplan, 2003). The stability of math proficiency was further explored by

Martin et al. (2012). The researchers sought to further define MLDs and outcomes by looking at the severity of the deficits in relation to categorical change and continuous change over time. The authors identified categorical change as changes in a students' educational program as it related to their need for specially designed instruction. In other words, categorical change is used to describe whether or not a student continues to be in need of special education services. Continuous change is defined as student growth over time, regardless of how they are educationally identified or labeled. The method of MLD determination is taken into account because of potential differences in access to intervention and lack of consensus regarding best practices when identifying MLD. The identification measures reviewed include the following: IQ-achievement discrepancy, performance below a percentile cutoff score, intra-individual differences, and response to intervention.

Previous research on the stability of MLD in terms of categorical change indicates that MLD are relatively dynamic based on what skills are measured at different stages in a student's education (i.e., students previously identified as meeting criteria of a MLD may no longer do so at a later point in time, or students originally identified as being low risk meet the criteria of a MLD at a later date). Stability of math deficits in terms of continuous change in a naturalist environment is explored in Martin et al. (2012). Typically, continuous change is studied to determine the effectiveness of a particular intervention. The growth of students receiving intervention is compared to growth of students in the control group. The researchers calculated a reliable change index for students outside of the context of a specific intervention. It was hypothesized that students with MLD would demonstrate more categorical change than students without a learning disability. The researchers separated student with MLD into two categories based on severity. Students with more severe deficits demonstrated more positive continuous

change but less categorical change than students with a less severe disability. Conversely, students with less severe MLD demonstrated more categorical change and less positive continuous change. It is important to note, categorical change is more prevalent in younger students (Martin et al., 2012).

Research indicates that some populations are more likely to enter school with deficits in mathematical learning. These populations include children living in low-income families and to a lesser extent females. However, research regarding a mathematical achievement gap between males and females is conflicting and inconsistent. Regardless of the root cause of early mathematical deficits, proactive intervention has been shown to be beneficial, necessitating the need for universal screening practices in mathematics.

Sex Differences in Math Achievement

As the world becomes more focused and reliant on technology, the quality of our science, technology, engineering, and mathematics (STEM) education becomes increasingly important for the United States to remain competitive in a global economy. According to a 2011 report produced by The President's Council of Advisors on Science and Technology (PCAST), the United States is lagging significantly behind other nations in STEM education at the elementary and secondary levels. When ranked in comparison to international counterparts, the United States is consistently falling at or below the middle of the group. According to the National Academy of Sciences (2005), there is a culture of dislike toward mathematics instruction in the United States. The authors suspect this is due to an overemphasis on rules and procedures and not enough explicit instruction on real-life application of skills. The culture toward math in the United States also perpetuates the myth that you are either good at math or you are not. This

culture differs from that of other developed countries, where people view success in math as being directly related to the effort put into learning it.

Many minority groups and women are under-represented in STEM professions. The number of women seeking higher education and joining the workforce continues to increase but is not reflected in STEM professions. The National Science Foundation (2015) reports 23% of doctorates in mathematics, 40.6% of master's degrees, and 43.1% of Bachelor's degrees are awarded to women. This is disproportionate considering women represent over half of all academic degrees earned. Women are awarded 57.4% of bachelor degrees, 62.6% of master's degrees, and 53.3% of doctorates (U.S. Department of Education, 2012).

According to PCAST (2011), there is a troubling lack of interest in STEM-related fields. Causes for the under-representation of women in STEM professions have been debated. Several areas of explanation have been explored and include biological and socio-cultural causes, with a multi-causal explanation being the most likely (Stoet & Geary, 2013). Some hypothesize this is due to poor education of STEM. Criticisms include a lack of focus on high achieving students, unenthusiastic educators of STEM, teachers who lack proper training, poor systems of support for STEM education, and antiquated STEM curriculums.

McGraw, Lubienski, and Strutchens (2006) reviewed the United States National Assessment of Educational Progress (NAEP) data from 1990 to 2003. The authors examined achievement trends in mathematics as they related to sex, race/ethnicity, and socioeconomic status. Specifically, they investigated whether gender gaps maintained from 1990 to 2003, whether gender gaps in mathematical achievement changed by mathematical strand (numerical operations, geometry, etc.) and achievement level; and whether males and females differed in their attitude toward mathematics. Results indicated a small, but statistically significant, and

persistent difference in mathematical achievement between male and female mathematical achievement from 1990 to 2003. It is noteworthy that gender differences remained consistent despite overall improvements in both male and female mathematical achievement in over a decade of data. The researchers found males outperformed females in four of the five mathematical strands. Negligible differences were observed between low achieving males and females (10^{th} percentile and below). However, the achievement gap increased directly with achievement, with the largest achievement gap occurring between the 75^{th} and 90^{th} percentiles. This suggests it is unlikely for sex to play a significant role in mathematical achievement for students who demonstrate mathematical deficits in math. However, it may be more relevant in students who are high achieving.

Stoet and Geary (2013) analyzed a decade of Programme for International Student Assessment (PISA) data to further investigate sex differences in mathematics and reading achievement. The PISA is funded by the Organization for Economic Co-Operation and Development and includes close to 1.5 million students from 75 different countries. The assessment content is the same for all countries and focuses on the problem-solving and application skills of mathematics, reading, and science. Similar to McGraw, Lubienski and Strutchen's (2006) findings, results of this analysis showed small but stable sex differences in mathematics performance across the four administrations of the PISA. The difference between male and female achievement became more pronounced among students who are high achieving and statistically insignificant among students who low achieving. These findings are in contrast with Else-Quest, Hyde, and Linn's (2010) meta-analysis of TIMSS and PISA data, who found gender difference in math achievement to be negligible.

Scheiber, Reynolds, Halovsky, and Kaufman (2015) challenged the significance of the achievement gap in their study examining the data gathered from the Kaufman Test of Educational Achievement – Second Edition, Brief Form (KTEA-II Brief) normative sample. The researchers included all data from students in 1st through 12th grades, 793 females and 781 males ranging in age from 6 to 21. The results did not yield a significant difference between math achievement in males and females, in contrast to the results of a PISA and other researching suggesting a significant sex gap (McGraw et al., 2006; Stoeb & Geary, 2013).

Hyde, Lindberg, Linn, Ellis, and Williams (2008) analyzed gender differences in mathematical achievement on state-administered academic achievement tests. The researchers requested gender and performance data from all 50 states, ten of which supplied data. After comparing data from participating states with NAEP data, it was concluded the ten states constituted a representative sample of the United States. Effect size for gender differences in mathematics was found to be insignificant (Scheiber et al., 2008). A meta-analysis completed in 1990, yielded similar results (Hyde, Fennema, Ryan, Frost, & Hopp, 1990). The authors reviewed studies from 1967 through 1988 that focused on sex differences in math and attitude toward math. The authors concluded the differences in mathematical performance and attitude toward math between males and females were small. The only substantial difference between male and females identified in this meta-analysis was in the stereotyping of math as a male-dominate domain (Hyde et al., 1990). This suggests that differences in male and female mathematical performance is rooted in a perception of mathematical achievement rather than an actual achievement gap.

Given the conflicted findings of mathematical differences between males and females, it is unclear whether a sex gap in mathematical achievement exists. Therefore, sex is included in

this study to determine if it is a contributing factor to the prediction of PSSA-M performance above and beyond that provided by the MBSP-C measure.

Socio-Economic Status and Math

According to the NAEP, a significant number of children from low-income homes do not obtain basic levels of mathematical proficiency (NAEP, U.S. Department of Education, 2015). While the United States has demonstrated overall improvements in mathematical skills based on the Trends in International Mathematics and Science Study (TIMSS), rates of growth are not as strong among some student populations (Aud, Fox, & KewalRamani, 2010). Math education in the United States is described as being overdetermined. According to the National Resource Council (2002), this happens when, "a large number of pressures exert forces on these systems, making them remarkably stable and resistant to change." (p. 33). Students from low SES homes and students with disabilities are among these more resistant populations.

This is especially concerning given that income achievement disparities have increased over recent years (Reardon, 2013; Reardon & Bischoff, 2011). Analysis of state-mandated academic achievement testing data suggests the achievement gap between children in low-income families and those in high income families has grown over the past three decades. It is of substantial relevance that an achievement gap already exists when students enter kindergarten and remains relatively stable as they progress through school. This suggests that the academic achievement gap between students in low- and high-SES is the result from out-of-school socio-economic factors rather than school practices and policies (Reardon, 2013).

Reardon (2013) examined the data of approximately 25,000 students from kindergarten to eighth grade made available through the Early Childhood Longitudinal Study-Kindergarten Cohort (Tourangeau et al., 2009). The authors found that while students were in school, the

academic achievement gap between children in low- and high-income homes decreased. However, the gap was re-established over summer months when children are not in the school setting. It is recommended that schools allocate additional resources to provide early intervention to students while in kindergarten and first grades because of the self-perpetuating nature of achievement gaps.

Bachman, Votruba-Drzal, Nokali, and Heatly (2015) explored the impact of SES on opportunity to learn procedural and conceptual math skills in elementary schools. The researchers used a robust amount of longitudinal data (N = 1,364) from multiple sites to examine the impact of SES on opportunities to learn and practice math while in elementary school. The authors did identify significant SES disparities in math achievement in first grade which decreased slightly but remained present and significant through fifth grade. It was hypothesized by the authors that students in low income families experienced less opportunity for mathematical learning in the home environment. It was further hypothesized that students in low income families receive less opportunity for high-order and conceptual math instruction, but high exposure to basic, procedural instruction in the school setting, which would perpetuate SES achievement disparities.

This hypothesis, however, was not supported by the data. The Bachman et al. (2015) found that students from low SES families received comparable or more high-order, conceptual instruction than their middle- to higher-SES peers. In complete contrast to the hypothesis, the authors concluded children in low-income families, especially with parent or caregivers who are less educated, would benefit from increased procedural instruction, especially of calculation skills. The impact of these findings on the present study is twofold. First, it supports the need to determine what impact SES as on future math outcomes. Secondly, it supports the need for early

identification and intervention for students who demonstrate math computation deficits. This can be accomplished through multi-tiered educational models such as MTSS.

Given these findings, the present study will look at the amount of variance in academic achievement that can be accounted for by a student's SES. If a significant amount of variance can be attributed to SES, school systems may be able to provide additional academic support to minimize the impact of SES on learning outcomes through additional intervention. As this study suggests, intervention should target summer months for substantive intervention for students from low SES environments.

Multi-Tiered Systems of Support

MTSS is a data based decision-making model designed to support the needs of all students with a dynamic problem solving approach. Initially, MTSS models were commonly referred to as Response to Intervention (RTI). However, it is important to clarify RTI is a process used to determine whether or not a student has a specific learning disability and is in need of specially designed instruction, requiring an Individualized Education Program (IEP; Kovaleski, VanDerHeyden, & Shapiro, 2013). RTI is the preferred method for special education identification within MTSS models, but is not the framework itself. This has resulted in some confusion of terminology within school districts and communities. The legislation that validated RTI as a method to qualify students for special educational services, also limited its scope for many educators: The term RTI became synonymous with special education and the identification of specific learning disabilities. Therefore, the field has adopted the term MTSS to describe the system-wide model (Walker & Shinn, 2010).

MTSS is defined as "a multicomponent, comprehensive, and cohesive school-wide and classroom-based positive support system through which students at-risk for academic and

behavioral difficulties are identified and provided with evidence-based and data-informed instruction, support, and intervention" (Stoiber, 2014, p. 45). Multi-tiered models are characterized by tiered instructional practices and structure, with instructional intensity increasing with each tier.

The majority of MTSS models consist of three tiers, with Tier 1 being the foundation. All students receive the academic, social, and behavioral curriculum provided in Tier 1. Instruction in Tier 1 is differentiated, high quality, and utilizes evidence-based practices. Approximately 80% of students should be successful in Tier 1 (Kettler et al., 2014). The more effective Tier 1 is, the fewer students in need of the intensity of intervention provided at Tier 2 or Tier 3 levels (Walker & Shinn, 2010). Universal screening data, in addition to formative and summative assessment data, are used to make decisions regarding movement within tiers at this initial level (Albers & Kettler, 2014; Kovaleski & Pederson, 2014; Parisi, 2014; Stoiber, 2014). Universally screening students is a cornerstone of multi-tiered service delivery models. Some scholars suggest that without universal screeners, multi-tiered systems are another example of a wait-to-fail education model (Berninger, 2006).

Tier 2 provides additional academic, social, and behavioral supports to students who do not meet performance targets within Tier 1 (Johnson, Carter, & Pool, 2012). Approximately 15% of students require Tier 2 services (Kettler et al., 2014). Interventions provided in Tier 2 are generally designed to target specific skill deficits and are provided in a small group setting in or out of the general education classroom (Johnson et al., 2012). Duration of Tier 2 interventions can vary significantly, but generally occur 3-5 times per week for 30 to 40 minutes for 6 to 20 weeks (Stoiber, 2014). Structured, explicit interventions are recommended at Tier 2 because

they have been shown to significantly improve outcomes for struggling learners (Johnson et al., 2012).

Approximately 5% of students will not respond to Tier 2 instruction and require Tier 3 levels of intervention (Kettler et al., 2014). Tier 3 represents the most intensive set of interventions and supports for students and in some models is synonymous with special education services (Johnson et al., 2012). In other models, Tier 3 is not considered special education but rather the most intensive interventions available in the general education setting (Walker & Shinn, 2010). Tier 3 interventions occur in addition to the general education instruction (i.e., Tier 1) due to an increased need for repetition and opportunities for practice (Denton, 2012). Tier 3 interventions have a low teacher-to-student ratio as a way to increase instruction intensity. The ratio of teachers to students is typically 1:3 but can be as low as 1:1 (Denton, 2012; Stoiber, 2014). Students in Tier 3 receive intervention from 30 to 120 minutes, 5 days a week for 10 to 30 weeks (Stoiber, 2014). Tier 3 intervention is typically provided by qualified general education teachers, special education teachers, and reading specialists (Denton, 2012; Kovaleski et al., 2013).

When differentiating between instructional tiers in a MTSS model, the word intensity is frequently used. There are several ways to manipulate the intensity of instruction. Intensity can be increased or decreased based on the format of the intervention being used. For example, a scripted intervention program is much more intense than a less structured intervention such as repeated readings or a drill sandwich. Student-teacher ratio, duration of intervention, and frequency of intervention are other common means of changing intervention intensity.

Regardless of intensity level, interventions implemented within MTSS models should be evidence-based or empirically-supported. Empirically-based interventions have a sound body of

scientific research to support their effectiveness. Implementation of empirically-based interventions require treatment fidelity, which means educators apply intervention programs or instructional practices in the same manner as the research validating their effectiveness. Interventions that are evidence-based demonstrate substantial positive effect size, outcomes that can be replicated by others, and gains that are maintained over time (Walker & Shinn, 2010). Evidence-based interventions are those whose specific procedures and materials have not been validated via scientific study. However, evidenced-based interventions use strategies or techniques that have been validated through empirical research.

One potential challenge to implementing MTSS systems is determining which students should receive what supports and the duration of these supports (Stoiber, 2014). As previously noted, MTSS focuses on data-based decision-making. Instructional decisions and movement through tiers is based on student data. Students are universally screened in order to identify who may be at-risk for poor learning outcomes and could benefit from additional academic and/or behavioral intervention. Once students begin intervention, progress monitoring data provide educators with information regarding students' response to intervention. It is recommended that schools develop data analysis teams to engage in systematic data analysis teaming to aid in instructional decision making and movement between tiers (Kovaleski & Pederson, 2014; Nellis, 2012).

Data analysis teams use universal screening data, in conjunction with other available data, to guide large-group instructional planning. Data analysis teams typically meet a minimum of three times a year, after tri-annual (fall, winter, and spring) universal screening measures are administered. Data are disseminated and reviewed prior to data analysis meetings. Meetings begin with a review and summary of universal screening data. First, areas of weakness within a

grade-level are identified and goals are set for the next universal screening administration. Once weaknesses are identified and goals are set, the team generates a list of possible Tier 1 strategies to improve student outcomes. The most appropriate strategies to target deficit areas are selected and a plan for implementation and support is developed. At this point, the data analysis team identifies students who are potential non-responders, or students who are likely to need more intensive support than Tier 1 strategies (Kovaleski & Pederson, 2008).

At this point, data analysis teams look at individual student data as opposed to data for an entire grade level. Students with similar areas of need are identified and intervention groups are formed. Similar to data analysis teaming procedures at the Tier 1 level, a list of intervention packages is generated, with the team selecting the intervention best suited to the instructional needs of the students. A plan for implementation and progress monitoring is then developed (Kovaleski & Pederson, 2014).

Data analysis teams continue to use existing data when reviewing which students may need Tier 3 intensity of instruction. Data typically available at this level include universal screening, local and state assessments, disciplinary referrals, attendance records, and progress monitoring data from intervention that may have been provided at Tier 1 and Tier 2 intensities. At this point, the data analysis team may also require additional assessment to further support instructional planning for individual students. After additional data are collected, appropriate interventions are identified and selected. A plan for implementation and progress monitoring is developed and put into action. If a student is not responsive to the intensity of intervention at Tier 3, the data analysis team may recommend an evaluation to determine eligibility for special education services (Kovaleski & Pederson, 2014).

MTSS recognizes the need for a continuum of services to meet the educational needs of all students with early, proactive intervention. There is an acknowledgement that not all students are successful within the core curriculum, but lack of success can be addressed outside of special education via increased intensity of intervention. MTSS attempts to address systemic program issues as opposed to attributing poor learning outcomes to student deficits (Ikeda, Paine, & Elliott, 2010; Walker & Shinn, 2010). It is seen as a vehicle for educators to embed effective, research-based academic and behavioral programs and utilize financial and personnel resources more efficiently to improve student outcomes (Ikeda et al., 2010).

Benefits of MTSS include providing early, research-based intervention instead of waiting until students are performing significantly below age or grade expectations before receiving help, or a wait-to-fail model. A related benefit is decreasing the number of students receiving special education services by meeting the needs of most students within the general education setting. MTSS methods decrease the likelihood of students who are culturally diverse, are from low socio-economic environments, and / or are non-native English speakers from being over-identified in special education (Klotz & Canter, 2007).

For MTSS models to function efficiently and improve student outcomes, universal screening instruments with adequate technical and classification accuracy must be administered to all students. There is a plethora of research to support universal screening for reading within MTSS. However, MTSS is not meant to be reading-centric, but to support the whole child. Therefore, more research is need in universal screening for math, writing, and social/emotional deficits to provide early proactive intervention. The focus of this study is to add to the research base regarding the predictive validity of a math computation universal screening instrument.

Secondly, this study will provide information regarding the impact of SES and sex on mathematical achievement.

Response to Intervention

RTI is a tiered process to determine eligibility for special education services. The scope of RTI is limited to the identification of students for special education services. RTI gained legislative support with the 2004 revision of IDEIA. It identified RTI as a method for determining the presence of a specific learning disability (SLD) and need for special education services. This grew out of a recognition that the ability-achievement discrepancy model was not an effective method for identifying SLD (Gresham, Reschly, & Shinn, 2010).

The common alternative to an RTI model is the discrepancy model (Glover & Albers, 2007). The ability-achievement discrepancy model has been appropriately dubbed the wait-to-fail model because it is neither predictive in nature nor supportive of early, proactive intervention. First, this method of identifying students for support works against early proactive supports, which have been shown to improve student outcomes. Students need to be performing significantly below age or grade achievement levels prior to referral to qualify for support services, perpetuating learning problems by blocking access to appropriate intervention (Glovers & Albers, 2007). Secondly, there is not a universal or systematic method of referral under the ability-achievement model. Schools are inconsistent in applying discrepancy model rules when determining which students have a specific learning disability resulting in an over or under-identification of students requiring special education services (Gresham et al., 2010). In addition to blocking access to intervention and inconsistencies regarding criteria of a SLD, research exploring the effectiveness of the ability-achievement model does not validate it as an accurate means of identifying SLD. Studies examining the classification accuracy of the ability-

achievement model indicate a very high rate of false positives (33% to 45%; Gresham et al., 2010). Despite this, the ability-achievement method remains prevalent in practice.

The underpinnings of RTI for identification of an SLD focus on the effective use of the school system to provide proactive, preventative, and early interventions as opposed to the wait-to-fail model. Members of the multi-disciplinary evaluation team review data from multiple measures given over time, including progress monitoring data, to determine if the student demonstrates dual discrepancies. Dual discrepancies require that a student demonstrates below expected grade or age level performance levels in addition to a slow rate of improvement or lack of progress in relation to peers (Kovaleski et al., 2013; Lichtenstein, 2014). Students who demonstrate a low level and rate of academic growth in relation to same age and/or grade peers are considered as having a dual discrepancy. Students who demonstrate a dual discrepancy frequently have more severe academic deficits than children with an IQ-achievement discrepancy and are more resistant to intervention. However, children, even those with dual discrepancies, have been shown to benefit from early intervention when identified as at-risk with universal screening practices (Speece, Case, & Molloy, 2003; VanDerHeyden, 2011). Universally screening students in order to provide early intervention is a cornerstone of both MTSS and RTI.

Universal Screening

The importance of early intervention for academic skills, social/emotional development, and behavior difficulties has been clearly identified. Yet there are persistent inconsistencies about how to best identify students who may be at-risk and how to best address these deficits. Prior to widespread acceptance of problem-solving models in education, such as MTSS, teacher referral was the primary means by which students were nominated for additional intervention.

While teachers provide valuable information regarding how a student functions in the classroom, there is little empirical evidence to support teacher referral as the sole method to identify students who need additional intervention (Elliott, Haui, & Roach, 2007).

To address the inefficient and ineffective teacher-referral process, educational systems borrowed from medical models and adopted universal screening practices for a more systematic and comprehensive approach to early identification of students in need of additional intervention. Universal screening is defined as a "process that generally consists of administering measures or collecting other data to allow broad generalizations to be made regarding the future performance and outcomes of all students, both at the individual level and at the group level" (Albers & Kettler, 2014, p. 121). The practice of universally screening students is considered an essential requirement to provide students at risk for emotional, behavioral, and academic difficulties access to early intervention (Glover & Albers, 2007). Educational decisions made by child study or data analysis teams that allow students early access to academic, social, or emotional interventions are more likely to be consistent when screenings are given universally, or to all students (Kettler, Glover, Albers, & Feeney-Kettler, 2014).

Function of Universal Screening

The function of universal screening is to inform educational systems and educators about students' academic and behavioral needs and then intervene on those needs. Universal screening data serve this function in two ways. Screening data are used first, at an individual level, to identify students who would benefit from additional intervention or increased intensity of instruction. Second, universal screening data are used to evaluate the effectiveness of a system and aid in the identification of curricular areas that may need more robust instruction, or areas which require a higher quality of instruction (Kettler et al., 2014).

Within a three-tiered MTSS model, the first step to identifying students who are at-risk of not achieving learning goals is to utilize a universal screener. When used in this capacity, universal screening practices identify students for the purpose of proactive intervention rather than screening for identification of an already establish deficit. This shift in practice is focused on providing students with the appropriate intensity of instruction in an early, proactive manner to minimize or negate long-lasting negative outcomes associated with academic difficulties (Albers & Kettler, 2014).

There is a robust body of research supporting this practice in reading, with well-established general outcome measures, in stark contrast to a lack of agreed upon general outcome measures in mathematics (Jenkins, et al., 2007; Mazzocco, 2003). As education systems successfully implement multi-tiered intervention programs in reading, more attention has been placed on establishing similar procedures in math (Clarke, Haymond, & Gersten, 2014; Clarke, Nese, et al, 2011; Gersten, Jordan, Flojo, 2005; Methe, Briesch, & Hulac, 2015). This has proved difficult given math's interwoven conceptual knowledge skills, non-sequential development, and lack of clearly defined general outcome measures and is compounded by significant variations in curriculums and instructional strategies (Clarke et al., 2011; Kelley, 2008). Despite these challenges, some promising developments have occurred in recent years regarding how to proactively identify students in need of more intense math instruction within a multi-tiered service delivery model.

A secondary function of universal screening is to determine the effectiveness or benefit of the curriculum or instructional practices for the majority of the student population. Universal screening measures are designed to be administered to all students to identify those who may be at risk and in need of intervention but also evaluate the effectiveness of the core curriculum

(Albers & Kettler, 2014). When used as part of an MTSS, it is assumed the instruction received by students in Tier 1, or the core curriculum, is of high quality and differentiated to student needs (Jenkins, et al., 2007). MTSS models encourage on-going evaluation of core practices. Universal screening data are used in conjunction with other available data to assess the efficacy of the core curriculum and instruction. Similarly, these data can be used to identify potential deficits and develop a systems-level plan to remediate those deficits. After implementation of a systems-level plan, universal data are used to help determine its effectiveness (Parisi et al., 2014).

Teachers and interventionists are also able to employ universal screening data as a starting point for differentiating instruction (Albers & Kettler, 2014). In addition to laying a strong instructional foundation for all students, high-quality core instruction includes ongoing opportunity to increase the intensity of instruction for students who may be struggling and opportunities for students who are high achieving to be accelerated (Clarke, Doabler, & Nelson, 2014). Teachers are able to use data from the fall universal screening and spring of the previous year to start differentiating Tier 1 instruction early in the academic year by determining which students may benefit from an increased intensity of instruction within the classroom. Interventionists are able to use universal screening data, along with other sources of information, to determine what intervention will be a good instructional match for a student.

Features of Universal Screening Measures

While screening measures can vary significantly based on what is being monitored, there are several consistent features of universal screening measures. Universal screening data are used for systemic program evaluation or to determine if instruction is effective in a class, grade, school, or district and to identify students who would benefit from a greater intensity of

instruction (Albers & Kettler, 2014). From a functional standpoint, universal screening measures should lend themselves to group administration a minimum of three times per year (Albers & Kettler, 2014; Kettler et al., 2014; Kovaleski & Pedersen, 2014).

High quality universal screening instruments should reflect general outcome measures such as state and/or national curricular standards, be sensitive to small increments of change over time, be capable of differentiating between students performing within expected ranges and those who are not, and be easily administered to the majority of the student population or be developed with universal design (Anderson, Lai, Alonzo, & Tindal, 2011). Assessments developed with universal design take into consideration the characteristics of all test takers, are composed of explicit constructs and bias-free content, are accessible, allow for accommodations, have simple, clear administration and scoring procedures, use appropriate readability, and utilize legible text and graphics (Anderson et al., 2011). Another important but often overlooked requirement for universal screening instruments is the data gathered should be easily disseminated to and understood by school personnel for effective data-analysis teaming (Kovaleski & Pedersen, 2014; Messick, 1995; Nellis, 2012).

In addition to usability and appropriateness for intended use, technical adequacy is an essential requirement of a universal screening tool. Appropriate technical adequacy of an assessment tool for decision-making purposes is one of the preliminary criteria that should be considered when evaluating an instrument for the purpose of universal screening (Christ, Johnson-Gros, & Hintze, 2005; Glover & Albers, 2007; Jenkins, Hudson, & Johnson, 2007). Technical adequacy for purposes of universal screening encompasses criterion validity, classification accuracy, sensitivity, specificity, and ability to measure growth over time (Glover & Albers, 2007; Jenkins et al., 2007; VanDerHeyden, 2010, 2011).

The paradigm shift from a reactive to a proactive approach toward student learning and behavioral difficulties are reflected in the use of universal screening data. Previously, these data were used to identify deficits in curriculums and individual students. The data are now being used as a method for identifying students who may struggle academically, socially, or behaviorally in the future to provide early intervention and for matching those students with the appropriate intensity of instruction (Albers & Kettler, 2014; Kettler et al., 2014; Kovaleski & Pedersen, 2014). Universal screening instruments provide preliminary information regarding potential skill strengths and weaknesses. Universal screening measures are frequently used as the first gate in a multiple-gate or gated screening system.

Gated Evaluation System

Gated evaluation systems, also referred to as multiple-gating, is defined as a "generic process involving multiple assessments that cost efficiently identify a subset of individuals from a larger pool of target participants with a combination of methods and measures generally arranged in sequential order." (Walker, Small, Severson, Seeley, & Feil, 2014, p. 47). For example, when oral reading fluency probes are administered as a universal screening tool, fluency and accuracy data are collected. These initial data could be used to separate students into different intervention groups or determine what Gate 2 assessment should be utilized.

Recent research has suggested promising results when using curriculum-based measures (CBM) as the first step in a gated evaluation system. Gated evaluation systems have been proposed as one potential solution to the problem of false positives (Fuchs, et al., 2011; VanDerHeyden, 2010, 2011). It is recommended that additional universal screening measures be introduced only when relevant and can improve identification accuracy (VanDerHeyden, 2013).

This procedure may be especially relevant when identifying students in need of math intervention because of the interwoven skills and non-sequential nature of mathematical learning (Missall et al., 2012). Fuchs et al. (2011) explored the use of a two-stage screening to identify students with math problem solving difficulties. The authors first administered a low-cost group-administered universal screener in math followed by a Dynamic Assessment (DA). DA has been shown as a strong predictor of word problem skills (Seethaler et al., 2011). Dynamic assessment "involves structuring a learning task, providing feedback or instruction to help the examinee learn the task, and indexing responsiveness to the assisted learning experience as a measure of the examinee's capacity to profit from future instruction" (Seethaler et al., 2011, p. 224). The researchers added a DA to the fourth year cohort of a previously established sample from a longitudinal study already in process. The specificity increased from 48.0% to 70.4% with the addition of a second screening tool. Sensitivity remained consistent at 87.5% (Fuchs et al., 2011). Therefore, these findings support the use of a gated evaluation system.

High Stakes Testing as Predictor Criterion

For universal screening data to be meaningful, they have to correlate with a relevant future outcome such as a high-stakes state-mandated academic achievement test (Clarke et al., 2014; VanDerHeyden, 2010, 2011). No Child Left Behind Act of 2001 (NCLB) and Individuals with Disabilities Education Improvement Act of 2004 (IDEIA) have acted as catalysts for educational reform in the United States that was initiated 20 years prior with the 1983 publication of *A Nation At Risk* by the Commission on Excellence in Education.

NCLB, a reauthorization of the Elementary and Secondary Education Act (ESEA), was the largest federally-funded education program in the United States. The majority of funds were designated to Title 1 programs, which focus on reducing the impact of poverty on learning. In

addition to Title 1, NCLB set standards for teacher quality, instruction of English Language Learners, school safety, assessment, and educational innovation. The legislative guidelines in NCLB have resulted in increased accountability within school systems and the adoption by many states of the rigorous National Common Core curriculum.

NCLB was re-authorized and called Every Student Succeeds Act (ESSA) in 2015. Similar to NCLB, the focus of ESSA is providing all students with a high quality education. ESSA eliminates annual yearly progress as an accountability measure and permits states to determine their own measure of accountability. Subsequently, states are then able to direct more support toward the lowest-performing schools, schools with a high incidence of dropout, and schools with achievement gaps.

ESSA mandates that states administer an annual reading and mathematics assessment in grades three through eight and once during high school. Science is assessed once while in elementary school, middle or junior high school, and high school. While one intention of ESSA is to decrease the amount of time students spend being assessed, state assessments continue to be mandated. Ideally, the state academic achievement tests reflect the academic skills that are valued by the federal and state education system, local community, local education system, families, and students. The state assessments are direct measures of the required standards. Therefore, they are important indictors of how well students are mastering content the state has determined to be valuable academic skills. Because state-mandated academic achievement tests are considered high-stakes and the direct measure of critical learning skills, they are often used as predictor criterion for the validation of universal screening instruments.

The primary intended consequence of high-stakes testing is to improve student academic achievement and overall student outcome. This results in an increased alignment between

instruction and standards; improved instructional efficiency; more pointed resource allocation; heightened student, teacher, and parent motivation to perform, teach, and support; an increased reliance on evidence-based instructional methods; and reducing the achievement gap between the majority of students and underserved or low-achieving students (Braden & Schroeder, 2004).

Braden and Schroeder (2004) identified several intended and unintended consequences of high-states testing. Unintended consequences of high-stakes include a narrow curricular focus, academic demoralization in low performing schools, test anxiety in students and school personnel, cheating, inappropriate resource allocation, and use of a single data source to make high-stake educational decisions. High-stakes test results may impact grade promotion, class placement, graduation, and teacher performance ratings, hence the term high-stakes testing (Braden & Tayrose, 2008). While the impact of ESSA is yet unknown, the intention is to decrease some of the negative unintended consequences of state-mandated academic achievement testing.

In order to further mitigate negative consequences of high-stakes tests, Braden and Schroeder (2004) suggested using test results to help inform instruction for subsequent years. They further recommended the use of sound research methods when making decisions regarding curriculum, instruction, and intervention and ensuring opportunities to learn. These practices increase the probability of the positive intended consequences of high-stakes testing (Braden & Schroeder, 2004; Braden & Tayrose, 2008). A universal screener with high predictive validity with high-stakes testing will help students to access supplemental intervention while continuing to participate in the general education curriculum.

The second piece of legislation that has played a significant role in education reform is the 2004 reauthorization of IDEIA, originally known as the Education for All Handicapped

Children Act of 1975. This reauthorization aimed to improve early intervention services and pre-referral intervention programs (Braden & Tayrose, 2008). This is represented by the legislative support provided in IDEIA for the use of a multi-tiered system as a method of educational service delivery and to identify students as being in need of specially designed instruction. IDEIA also re-directed funds from other programs in order to provide fiscal backing for pre-referral programs.

IDEIA required an increased inclusion of students with disabilities in state accountability systems (Braden & Tayrose, 2008). The inclusion of students with disabilities in state accountability data served as a motivating factor for schools to improve the quality of education students identified as needing specially designed instruction received.

Large scale analysis of strengths and needs of STEM instruction in this country, which has been facilitated in part by high stakes testing and the legislation that supports them, have consistently shown a need for and benefit of ongoing formative assessment to drive instructional decisions in math (Gersten et al., 2009; NMAP, 2008). These government-driven changes have a reciprocal relationship with a paradigm shift in school systems to move toward a more problem-solving, differentiated models of instructional delivery. These changes aim to ensure high quality education for all students, one driving the other.

Potential barriers to universal screening. Parisi, Ihlo, and Glover (2014) identified common barriers to effective universal screening practices. Common barriers include a lack of professional development with school personnel about how to use screening data to link at-risk students with appropriate interventions and engage in program evaluation. In addition to providing professional development to staff, Parisi et al. (2014) suggest that school administrators and general educators should be included on the data analysis team. Successful

implementation of a multi-tiered model without the support of administration/leadership to advocate for the collection and use of universal screening data is very difficult. Members of leadership teams must be well-trained in data-based decision-making and should include general education teachers, specialists, and special education teachers. Excluding general education teachers from data analysis teams perpetuates the belief that students need special education to be successful and takes away from a culture of shared ownership regarding the performance of all students.

Waiting to employ universal screening data until all staff are committed to a multi-tiered service delivery model is another potential barrier. Guidelines for introducing MTSS models suggest securing the buy-in of 80% of school personnel prior to implementation. However, universal screening instruments can be introduced prior to high levels of staff support. When done properly, universal screening procedures can increase staff support of multi-tiered service delivery models (Parisi et al., 2014). One possible solution to mitigate these barriers is the development of a data-analysis team to engage in systematic review of student data, commonly referred to as data-analysis teaming (Kovaleski & Pedersen, 2008; Kovaleski & Pedersen, 2014; Nellis, 2012). Data analysis teams are composed of "teachers, school psychologists, administrators, and other educators who meet to conceptualize how data inform instructional decision making" (Kovaleski & Pedersen, 2014, p. 100).

A potential barrier to implementation of universal screeners is the high rate of false positives (Fuchs et al., 2011; Mazzocco, 2003; VanDerHeyden, 2010, 2011, 2013). False positives identify students as being at-risk when they would develop proficient academic skills without supplemental intervention. High rates of false positives create financial and personnel strain and decrease the intensity of intervention for students who are truly at-risk, which is why

technical adequacy of universal screening tools is so paramount (Fuchs et al., 2011; VanDerHeyden, 2010, 2011). Despite these criticisms, universal screening practices are preferred over conventional methods used to identify students for additional intervention, such as teacher referral (Fuchs et al., 2011). Universal screening data are more systematic and demonstrate better reliability and validity than conventional methods, resulting in more accurate identification of students who may be at-risk for academic difficulties.

A high percentage of students failing to meet performance benchmarks on a universal screening measure indicates a systemic problem, not a student-specific problem (Albers & Kettler, 2014; Kettler et al., 2014). When school systems chronically identify a large percentage of students as being at-risk, system-wide interventions are recommended over individual student interventions. When systemic problems have been identified, universally screening for individual deficits may no longer be functional.

VanDerHeyden (2013) cautioned against universally screening all students when other data sources indicate a majority of students are demonstrating academic, emotional, and/or behavioral difficulties. In educational systems that have a large percentage of students identified as at risk, universal screening systems are no longer as effective and efficient. Typically, when over 20% of the student population is demonstrating a need, it is considered a systemic deficit and systemic interventions are recommended. The use of threshold decision making is promoted to take into account contextual factors that impact student outcomes.

Threshold decision-making models are prevalent in the medical field but have not yet translated into educational practice. Medically, threshold decision-making is used to determine whether screening and/or intervention should be initiated based on the probability of being asymptomatic, probability of negative side effects for any given age for participation or lack of

participation in screening, and probability of death (Hoffman, Wilkes, Day, Bell, & Higa, 2006). According to VanDerHeyden (2013), three options should be considered when making decisions about the need for intervention:

> (a) to provide intervention without screening, (b) to withhold intervention without testing, (c) to conduct the test to determine whether the intervention is needed for the 'in-the-middle' students or students who are neither clearly at-risk or clearly not at-risk. (p.406)

In an education setting, threshold decision-making would require educators to consider the probability of a false negative, probability of a false positive, probability of a false result for students who will not fail (specificity), benefit of intervention for students who will fail, risk of intervention for students who will not fail, and risk-of-test. Please refer to Figure 2 for a visual representation of this model. In this model, screening is only conducted with students for whom it is unsure whether or not they are in need of intervention.

Potential benefits of this model include less strain on school resources and instructional time to conduct universal screenings with all students and decreased risk of flooding Tier 2 intervention with false positives. Potential downfalls include loss of universal screening data to evaluate the effectiveness of the educational system and potential for false negatives (VanDerHeyden, 2013). Empirical and longitudinal research is need to determine whether or not threshold decision-making is applicable within an educational setting.

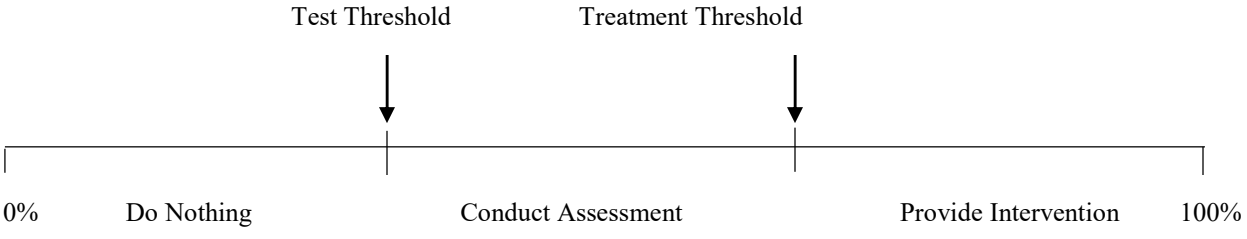


Figure 2. Representation of logic behind threshold decision making

While more research is being conducted into alternative methods, CBM are the most prevalent universal screening tools. CBM "represents a set of standardized and specific measurement procedures that can be used to quantify student performance in the basic skill areas of reading, spelling, mathematics computation, and written expression" (Hintze, Christ, & Methe, 2006, p. 51).

Curriculum Based Measures as Math Universal Screeners

When discussing high-stakes testing and assessment, it is important to be cognizant of the multifunctional role of assessments. Assessments serve one or more functions. They can be used for placement or selection, accountability purposes, diagnosis, and in the support of learning (Berry, 2008). Within the current high-stakes educational climate, it is beneficial for school systems to choose assessment instruments that serve multiple functions to minimize the amount of instructional time lost to assessment.

In the literature, the terms CBM and universal screener are often used interchangeably. CBM refers to a type of assessment, while universal screener refers to how that assessment is functioning or being used. The unique characteristics of CBM make them ideal for use as a universal screening instrument. CBM are developed in direct reflection of local curricula and measure multiple math constructs (Deno, 1985). Unlike classroom-based assessments or curriculum-embedded assessments (e.g., chapter or unit tests), the same constructs are assessed over time to allow for measurement of growth as opposed to the extent of skill mastery (Fuchs, et al. 2005). Curriculum-based measures have been shown to be a reliable and valid method for assessing a student's knowledge at any given point of time (Fuchs et al., 2005). A primary function of CBM is progress monitoring of student learning or to measure academic growth over time (Codding, Petscher, & Truckenmiller, 2015).

Two broad approaches to CBMs have been identified, curriculum sampling and robust indicators (Christ et al., 2008; Foegen et al., 2007; Fuchs, 2004; Fuchs, et al., 2008). Curriculum sampling measures represent skills students are expected to learn throughout each school year and relate to a specific curriculum (Foegen et al., 2007; Fuchs, 2004; Fuchs et al., 2008). Christ and Vining (2006) differentiate between Curriculum Based Measures of Math (CBM-M) that are subskill mastery measures and those that are general outcome measures. Subskill mastery measures are used to assess a specific skill that is expected to be acquired over a brief period of time. The curricular-sampling approach to CBM is an example of subskill mastery measures. Curricular-sampling CBM is composed of skills representative of what a student is expected to learn by the end of the school year and a direct reflection of students' curriculum (Fuchs, Fuchs, & Zumeta, 2008).

Robust indicator measures are composed of skills that represent proficiency in mathematics but are not representative of a specific curriculum (Christ et al., 2008; Foegen et al., 2007; Fuchs, 2004). Robust indicators are considered curriculum-based despite not being tied to a particular curriculum because they are intended to measure core competencies which students are expected to know at the end of a certain period of time or grade level (Christ et al., 2008). Robust indicators are also referred to as general outcome measures. General outcome measures assess students on skills they are expected to master over an extended period of time, such as the course of an academic year (Christ & Vining, 2006; Christ et al., 2008). It is important to note that, although the authors refer to math computation as a general outcome measure, there is a significant amount of debate in the field regarding what skill sets constitute a general outcome measure in mathematics (Mazzocco, 2003). At the elementary school level, robust indicators focus predominately on basic fact fluency. However, assessment of early mathematical

competencies have gained increased attention (Foegen et al., 2007; Lago & DiPerna, 2010; Polignano & Hojnoski, 2012).

CBM-M can be used to assess computation and/or concepts and application skills (Christ, Scullin, Tolbize, & Jiban, 2008). Computation probes assess basic arithmetic skills. Measures can assess multiple skills (addition and subtraction without regrouping) or focus on a specific skill (e.g., multiplication facts to 12). The skills represented on the computation probe should reflect the curriculum or specific grade-level skills. Concept/Application probes focus on the application of arithmetic skills to the problem solving process. They may consist of word problems and more complex mathematical operations (i.e., solving for an unknown or interpreting charts and graphs). Probes can be scored as digits correct or correct problems. Digits correct scoring has been shown to be more sensitive to growth over time (Stevens-Olinger, 2014).

CBM differentiates itself from mastery measurement tools and other forms of curriculum based assessments (CBA) in several ways. CBM assesses students on broad learning objectives as opposed to short-term learning objectives (Fuchs, 2004; Hintze, Christ, & Methe, 2006). Secondly, CBM assesses students repeatedly on the same set of skills. Another distinction between CBM and CBA is use of standardized administration and scoring procedures. It is important to distinguish between these two instruments because CBM is frequently used as universal screeners and CBA is not.

Clarke et al. (2014) identified two types of established universal screening tools in mathematics: single-proficiency measures and multiple-proficiency measures. CBM can be used to evaluate student performance on a specific skills area (single-skill computation measures) or a broad range of skills (mixed-skill computation or concept/application measures). Both types of

screening measures have been shown to be good predictors for future math outcomes. Research in the fields of developmental and cognitive psychology supports the use of single-proficiency measures. Single-proficiency measures typically focus on early numeracy and basic arithmetic skills. They are typically brief (1-2 minutes) and easy to administer (Albers & Kettler, 2014; Clarke et al., 2014). Multiple-proficiency measures, contrarily, combine several skills into one screening probe (Clarke et al., 2014). For example, Number Sense Brief (Jordan, Glutting, & Ramineni, 2008) combined multiple early numeracy skills, strategic counting, magnitude comparison, and number identification into one brief measure. Preliminary research suggests multiple-proficiency measures assess a more comprehensive range of mathematical skills, which results in a slightly stronger predictive power than single-proficiency measures (Clarke et al., 2014).

Psychometric Adequacy

There is a growing body of research examining the utility of both single and multiple-proficiency measures. Fuchs (2004) identified three stages of programmatic research on CBM. Stage 1 revolves around the technical features of each individual or static score. Stage 2 investigates the technical features of slope and if growth captured on CBM measure translates to increased achievement within the academic domain. The intent of Stage 3 research is to determine the instructional utility of a CBM. In other words, can the data gathered from that particular measure be used to inform instructional decisions resulting in improved student outcomes (Fuchs, 2004)? The majority of research relating to curriculum-based measures of mathematics (CBM-M) would be categorized as Stage 1 and to lesser extent Stage 2.

Within the past decade, more emphasis has been placed on looking at CBM-M from the standpoint of instructional utility. This is in part due to the MTSS/RTI paradigm shift.

Educators are now looking to transfer established RTI reading practices to math instruction, especially as MTSS models, which emphasize a cohesive system of instruction, have gained increased support. It is important to note that there is a significant amount of research validating CBM in reading at all three stages; however, the depth of research in mathematics is not present (Fuchs, 2004; January & Ardoin, 2015). Another significant factor in this research trend is the necessity of a tool being validated at Stage 1 and Stage 2 levels for progress monitoring prior to the instructional utility being explored at Stage 3 (Fuchs, 2004).

The technical adequacy of CBM has recently been revisited as a topic of study and concern. Claims have been made that many students may have been inaccurately identified as being learning disabled based on CBM that was not technically sound for educational decision-making purposes (Methe et al., 2015). Effective universal screening measures are required to demonstrate adequate technical adequacy. Generally, for universal screening instruments, this is defined as having a reliability of .70 or higher, being able to differentiate between different groups of students, exhibiting a strong correlation with a criterion measure, and demonstrating appropriate levels of specificity and sensitivity. In an RTI model, CBM can be used as part of the eligibility for the special education decision making process. It is important to note, high-stakes testing which could possibly result in life-changing special education classification require much stronger reliability of .90 or higher (Christ & Nelson, 2014). The validity of a screening instrument is frequently established by correlating the researched instrument with a previously established measure. Correlation coefficients of .40 - .50 or higher are generally considered acceptable if the criterion measure has well-established psychometric properties (Burns et al., 2014).

Reliability of CBM. The reliability of a measure is its ability to be consistent over time, across alternate forms and raters. Types of reliability most relevant to this study and CBM include test-retest reliability and alternative form reliability. Test-retest reliability describes the stability of an instrument over time. For the purpose of universal screening, a correlation of .70 or higher between two separate administrations of a measure is considered good test-retest reliability. Alternative form reliability is a measure of stability of performance across multiple different but equivalent versions of an instrument (Jackson, 2003; Leary, 2001; Strait et al., 2015). Multiple forms of CBM-M including measures of early numeracy skill, computation and basic math fact fluency measure, and concept/application measures, have been able to consistently demonstrate adequate reliability coefficients (Foegen et al., 2007; Gersten et al., 2012).

Validity of CBM. Validity is an instrument's ability to measure what it claims to or is intended to measure. Predictive validity is a measure's ability to predict a future outcome. Predictive validity is a type of criterion related validity, which means a measure is correlated with a previously established measure representative of the desired outcome. The predictive validity of MBSP-C with PSSA-M is the primary focus of this study. In other words, to what extent can a MBSP-C probe administered in the fall, winter, and spring of first, second, and third grade predict student performance on the PSSA-M administered in the spring of third grade? If the relationship between the two measures is strong, then MBSP-C is likely a good instrument for universally screening students to identify those in need of additional math intervention.

The predictive validity of a measure results in four possible categories upon which screening decisions are based: true positive, true negative, false positive, and false negative. A true positive is when a student is correctly identified as being at-risk. A true negative is when a

student is correctly identified as not being at-risk. False positives, also referred to as Type I errors, are when students are inaccurately identified as being at-risk. False negatives, or Type II errors, occur when a measure fails to identify a student as being at-risk. False negatives are considered the worst possible outcome in regards to predictive validity because these students would benefit from additional intervention but are not identified as being at-risk. This could result in missed opportunity to provide the appropriate intensity of instruction a student may need for positive learning outcomes (Kettler et al., 2014; VanDerHeyden, 2010). The accuracy of a particular score will fit into one of these reporting categories based on screening data. When the validity outcomes are analyzed as a group, the predictive validity of a measure can be further calculated through specificity, sensitivity, positive predictive power, and negative predictive power (Kettler et al., 2014).

Specificity is the likelihood a screening measure will correctly identify a student as not being at-risk. The sensitivity of an instrument is the likelihood that a screening measure will correctly identify a student as being at-risk. Specificity and sensitivity address the predictive power of the screening instrument. Positive predictive value and negative predictive power relate to the accuracy of cut-off scores used to determine whether or not a student is at-risk or the proportion of students screened who ultimately perform successfully or poorly on the criterion outcome. Positive predictive power is the percentage of students identified as at-risk on the screen who later failed the criterion measure. Negative predictive power is the percentage of students identified as not at-risk on the screen who successfully passed the criterion measure (Kettler et al., 2014; Petscher, Kim, & Foorman, 2011; VanDerHeyden, 2010).

The specificity and sensitivity of CBM has become a focus of research as these measures are used more frequently to make educational decisions. Receiver Operating Characteristics

(ROC) analysis is a practical option for educators to find a balance between specificity and sensitivity. ROC analysis is a statistical procedure used to gauge an assessment's capacity to predict an outcome or differentiate between groups. This is done by establishing cut points to dichotomize a continuous scale into typical and atypical groups (Streiner & Cairney, 2007). ROC analysis allows for comparison between variables when their relationship is not linear.

The Area Under the Curve (AUC) is the primary statistic of ROC analysis. The AUC is the probability a measure will predict subjects to fall within the typical or atypical group. For example, when validating their brief measure of number sense (NSB) as a predictor of future math outcomes, Jordan et al. (2010) found an AUC of .88 in the winter of first grade. This suggests if we take two random students, one struggling in math and one not, it is 88% probable the student with typical achievement will perform higher on the NSB measure. An AUC between 0.50 and 0.70 is considered low, 0.70 and 0.90 is moderate, and over 0.90 is high. It is important to note AUC is an estimate. Therefore, a standard error should be generated. For universal screening instruments, this is typically done by calculating a confidence interval. The confidence interval should then be taken into account when analyzing AUC (Streiner & Cairney, 2007).

Diagnostic accuracy and technical adequacy address an instrument's specificity and sensitivity. Cut scores are developed, with the consideration of acceptable levels of specificity and sensitivity. Universal screening instruments can afford to have higher rates of false positives to decrease incidences of false negatives. This is because universal screening is meant to serve as a robust predictor of true positives. False positives can be identified by taking into account supplemental sources of data such as parent and teacher input, additional assessment data, and permanent product.

The validity of CBM-M was reviewed in Christ et al. (2007) under the context of Messick's (1995) framework which redefines validity in terms of practical application by combining construct validity, relevance-utility, social validity, and consequential validity. The authors determined CBM-M has adequate validity to be used as a decision making tool. However, it is noted, no studies that directly addressed the ethical implications of CBM-M or consequential validity, were found. Consequential validity deals with the potential negative social ramifications of an assessment or instrument. For an assessment to demonstrate strong consequential validity it would not cause any atypical social consequences (Christ et al., 2008; Messick, 1995). Instead, the majority of math universal screening research focuses on more classical notions of validity, such as predictive validity and classification agreement. Single and multiple skills instruments have been shown to produce equally sound classification agreement in the spring and fall. Overall the instruments were less predictive of future procedural math difficulties than future conceptual difficulties. It is hypothesized that this is due to limited instructional focus on numerical operations and high focus on early numeracy concepts in kindergarten and first grade (Seethaler & Fuchs, 2010).

Both curricular-sampling and robust indicator approaches to CBM-M have been shown to demonstrate high criterion-related validity (Christ et al., 2008; Fuchs et al., 2008). In addition to adequate correlations with criterion measures, robust indicators and curriculum sampling CBM are both able to distinguish between students of varying mathematical performance and growth over time (Foegan et al., 2007; Fuchs et al., 2007).

Strengths of CBM

CBM can be used to evaluate student performance on a specific skills area (single-skill computation measures) or a broad range of skills (mixed-skill computation or

concept/application measures). CBM assesses students on broad learning objectives as opposed to short-term learning objectives (Fuchs, 2004; Hintze, Christ, & Methe, 2006). This allows student growth to be measured over the course of an academic year. The expectation of growth over time is based on the skills students are expected to gain within that academic year.

CBM employs standardized administration and scoring procedures. This allows for inter- and intra-individual comparisons. The use of standardized measures allow data teams to make comparisons between students in addition to evaluating student growth (Hintze, et al., 2006; Shinn, 2008). Additional factors that have contributed to CBM's popularity include relatively low cost, efficiency, broad assessment of global skills as opposed to specific skills deficits, and access to alternative forms of the assessment (Fuchs & Fuchs, 1999). Secondly, since CBM is assessing students repeatedly on the same set of skills, comparisons can be made regarding retention and generalization of learning (Hintze et al., 2006).

These factors have led to CBM's identification as a more appropriate screening and progress monitoring tool than commercial standardized achievement tests. For example, fact retrieval (Geary & Hoard, 2001) and number sense (Jordan et al., 2009) have been shown to be more relevant for early identification of math deficits than standardized norm-referenced academic achievement tests (Martin et al., 2012).

There are strengths and weaknesses to both curriculum sampling and robust indictor CBMs, and educational systems need to be aware of these when choosing an instrument that best meets their universal screening needs. Benefits of a curriculum sampling approach include a direct association to instruction children receive in the classroom. Instruments developed with a curriculum sampling approach are directly linked to the curriculum students receive. Therefore, teachers are able to relate CBM performance back to specific skills represented in the curriculum

to address deficits and aid in instructional planning (Fuchs et al., 2008; VanDerHeyden & Burns, 2005). Benefits of robust indicator CBMs include allowing for comparisons across years and a focus on core skills as opposed to a particular curriculum. Robust indicators can be used for multiple years, which saves both time and money. Data generated from robust indicator measures aid in program evaluation because data can be compared over multiple cohorts of students.

Whether curriculum-sampling or robust indicator CBM are used, the skills assessed are reflective of student learning objectives. This results in the appropriate use of CBM as both a universal screener and as a progress monitoring tool. CBMs are frequently used to identify students who may be in need of additional intervention and monitor the effectiveness of intervention through progress monitoring (Fuchs et al., 2005). According to best practices, the universal screening instrument and progress monitoring instrument should be aligned. When the two instruments are aligned they measure the same constructs and function off of the same scale. This allows for data analysis teams to make more comprehensive conclusions and comparisons regarding growth over time which translates into better coordination of services and fluidity between tiers (Clarke et al., 2014).

Potential Weaknesses of CBM

While recent research is promising, weaknesses of CBM as universal screening instruments include limited research regarding their technical adequacy for educational decision-making (Christ, Johnson-Gros, & Hintze, 2005). Another potential weakness of CBM as universal screening instruments and an area which requires more research is whether or not they are appropriate for diverse populations. School systems serve increasingly diverse populations in terms of race, linguistics, and culture. School systems need to be aware of the concept of

universal design when choosing an instrument and more importantly, interpreting universal screening data (Albers & Kettler, 2014). For example, school systems should consider if a universal screening tool is appropriate to use with English language learners, those who are frequently truant, have a fine motor deficit, or speech and language deficits.

Both the curriculum-sampling and robust indictor approach to CBM have their unique differences which school systems need to be cognizant of when considering what universal screening instruments best meet their needs. Curriculum-sampling measure's direct reflection of a curriculum can be viewed as a strength regarding validity but a weakness in terms of practical application. A new measure needs to be generated anytime the curriculum is modified to continue to be reflective of the mathematics curriculum. This process can be time consuming and also requires new local normative data to be generated. Secondly, comparisons of student growth cannot be made from year to year because of the dynamic nature of curriculum sampling CBMs (Foegen et al., 2007).

The development of robust indicator measures can be difficult due to the dearth of general outcome measures in mathematics, especially after the development of early numeracy skills and mastery of basic computation (Kelley, 2008; Mazzocco, 2003). Because robust indicator CBMs are not directly tied to the curriculum, it is more difficult for teachers to use data for instructional decision making (Foegen et al., 2007).

Types of CBM-Math

Gersten et al. (2012) conducted a literature review of universal screening of mathematics. The researchers included articles from 1996 to 2011 on ERIC and PsychINFO electronic databases in addition to a manual review of the *Journal of Special Education Exceptional Children,* the *Journal of Educational Psychology, and* the *Journal of Learning Disability*.

Research included in the review focused on children ages birth to 12 years. The authors identified 48 total studies; 21 were selected for further review. Sixteen of the studies met inclusion criteria. Eleven of the 16 selected focused on single proficiency measures. Four studies included the use of multiple proficiency measures. Of the 16 studies selected, five predicted MLD or low achievement with diagnostic utility statistics. Three of the studies used a combination of single skills, multiple skill measures, and diagnostic utility statistics to predict mathematical deficits. All of the studies focused on one or more of the four skills that compose number sense/number competence: magnitude comparison (Booth & Siegler, 2006), strategic counting (Geary, 2004), ability to solve simple word problems (Jordan et al, 2009), and automaticity of basic math facts (Jordan, Hanich, & Kaplin, 2003). The majority of universal screening measures being used in educational systems reflect one of more of the mathematical components reviewed in this 2012 literature review.

Measures of Early Numeracy

Measures of early numeracy skills have a rapidly growing body of research with promising technical adequacy. Given the primary function of universal screening is for early identification of students in need of additional intervention, early numeracy skills, also referred to as early mathematical competencies, is of high importance (Methe, Begeny, & Leary, 2011). Early intervention of math deficits coupled with progress monitoring with formative assessment have been identified as critical factors for improving learning outcomes (Clarke & Shinn, 2004).

Cognitive psychologists have identified four early numeracy skills as being significant in future math outcomes: (a) magnitude comparison, (b) strategic counting, (c) the ability to solve simple word problems, and (d) an understanding of basic math facts (Dehaene et al., 2004; Geary, 2004; Gersten et al., 2012; Geary, 2004). Failure to grasp basic mathematical concepts

such as understanding the meaning of numbers, counting, and magnitude discrimination can have serious implications on learning outcomes (Gómez-Velázquez, Berumen, & González-Garrido, 2015). The majority of early numeracy research focuses on some variation of these skills.

Reliability and validity. Curriculum based measures of early math skills (CBM-EM) instruments administered with pre-kindergarten and kindergarten students have demonstrated strong reliability coefficients. Validity coefficients fall within the low to moderate range (.40 to .60) with this very young population but improve significantly, to .70 and higher, when CBM-EM are administered to students in first grade (Foegen et al., 2007; Gersten et al., 2012).

All four skills of CBM-EM have demonstrated moderate to high predictive validity. Magnitude comparison measures have demonstrated validity coefficients ranging from .50 to.79, which are considered appropriate for the purpose of universal screening. Strategic counting measures have also demonstrated adequate validity for first grade students (.68), but relatively weak validity in kindergarten (.37). A measure composed of simple word problems administered in the fall of kindergarten demonstrated moderate validity with a computation measure given at the end of second grade (.51). Measures of basic fact retrieval have yielded concurrent and predictive validity coefficients ranging from .50 to .59 (Gersten et al., 2012).

Predictive adequacy. Recent research indicates tests of early numeracy skills administered in kindergarten and first grade have a strong relationship with math achievement in third grade (Jordan, Kaplan, Ramineni, & Locuniak, 2009; Geary, Bailey, & Hoard, 2009; Jordan, Glutting, Ramineni, & Watkins, 2010). Clarke and Shinn (2004) investigated four potential measures of CBM-EM, an oral counting measure, a number identification measure, a quantity discrimination measure, and a missing number measure. Two standardized, nationally-normed tests of academic achievement (Woodcock-Johnson Applied Problems subtest and the

Number Knowledge Test) and a first grade computation probe were used as criterion measures. All four CBM-EM demonstrated strong inter-scorer, alternate form, and test-retest reliability, ranging from .76 to .99. Number identification, quantity discrimination, and missing number demonstrated moderate to strong concurrent validity with criterion measures in the fall, winter, and spring, ranging from .74 to .79. Oral counting consistently demonstrated the relatively weakest concurrent validity with the other experimental measures, ranging from .55 to .79. Oral counting yielded the weakest correlation with criterion measures, ranging from .49 to .70. Quantity discrimination was found to be the best predictor of early mathematical skills, both in terms of concurrent validity and predictive validity, ranging from .68 to .93 (Clarke & Shinn, 2004). Clarke et al. (2008) expanded on this research by investigating the predictive power and growth over time of a CBM-EM previously found to be reliable and valid (Clarke & Shinn, 2004).

Clarke et al. (2008) made an argument that rate of growth should play a significant role as a feature of student learning when evaluating the effectiveness of an intervention or instructional program, especially since this information is used for instructional decision making in MTSS/RTI systems. Of the four measures studied, oral counting, missing number, number identification, and quantity discrimination, the rate of improvement, or slope, of the quantity discrimination measure was the only one to account for additional variance, outside of the variance as a screening measure, with a criterion measure (Clarke et al., 2008). This suggests quantity discrimination is a stronger progress monitoring and universal screening tool for young students in need of or receiving math intervention than measures of oral counting, number identification, and missing number. The authors note concerns with the relatively small sample

size of their study (n = 111) and suggest the need for replication of this study with a larger sample size.

Missall, Mercer, Martinez, and Casebeer (2012) studied the same early numeracy skills with a significantly larger sample size ($N = 535$) with similar results. The authors evaluated the predictive validity of the Tests of Early Numeracy Curriculum-Based Measurement (TEN-CBM) with the Indiana Statewide Testing for Educational Progress-Plus (ISTEP+). TEN-CBM is a paper and pencil, two-minute mixed computation probe. The TEN-CBM assess four areas of number sense, oral counting, missing number, quantity discrimination, and number identification. The results indicated several significant findings. The results indicted a decrease in scores from spring of kindergarten to fall of first grade, which suggested loss of skill over the summer. Oral counting and number identification screens did independently predict future math outcomes. Quantity discrimination and missing number were the strongest predictors of performance on the state exam ISTEP+. The authors' findings support previous research on screening for early numeracy skills. Students who did not demonstrate proficiency on quantity discrimination and missing number screens in kindergarten continued to demonstrate poor math achievement in third grade (Jordan et al., 2010; Missall et al., 2012). They suggested implementing a quantity discrimination and missing number screening in kindergarten to identify students who need additional intervention (Missall et al., 2012).

The Number Sets Test (Geary, Bailey, & Hoard, 2009) is a group-administered paper and pencil assessment taking less than 10 minutes to complete. During administration, students are given a sheet with a combination of number sets composed of shapes such as diamonds, triangles, squares, and Arabic numerals. They are instructed to circle all of the sets that add up to either 5 or 9 depending on which form of the test they are given. Students with the sums to 5

version have one minute per page. Students with the sums to 9 form are given 90 seconds. The predictive power of the Number Sets Test was demonstrated to be strong. When administered in first grade, the Number Sets Test was able to correctly identify 67% of students identified with a MLD in mathematics by third grade. This was one of the few math CBM studies which also explored a measure's ability to correctly identify students who were not in need of additional math intervention. When administered in first grade, data from the Number Sets Test were able to correctly identify 90% of students who were not at-risk of a MLD (Geary et al., 2009).

The CBM-EM research was expanded on by the validation of a brief number sense screener (NSB) in a longitudinal study (Jordan et al., 2010). The NSB is an untimed, 33-item screener composed of items to assess counting knowledge and principles, number knowledge, number recognition, nonverbal addition and subtraction, addition and subtraction story problems, and addition/subtraction number combinations (Jordan et al., 2008). NSB has previously been established as having a unique and significant contribution to the variance on the Woodcock-Johnson, Third Edition (WJ-III; McGrew et al., 2007) math achievement subtests administered in first grade. The NSB was administered to kindergarten and first grade students ($N = 204$) on six occasions throughout the beginning of kindergarten to the middle of first grade. Results of the NSB were then used to predict performance on a high-stakes assessment given 3-4 years later while in third grade. Repeated measures and ROC analysis indicated that the skills measured in the NSB while students are in kindergarten and first grade is predictive of math achievement in third grade with the AUC ranging from .78 to .88 (Jordan et al., 2010).

Despite these encouraging findings regarding CBM-EM, Methe, Begeny, and Leary (2011) identified weaknesses with these measures. The authors noted the presence of few CBM-EM focused on practical application skills. The authors explored the technical adequacy and

diagnostic accuracy of CBM-EMs intended to measure informal and conceptual knowledge in an attempt to provide more information regarding the practical application of CBM-EM. It is important to note, the measure in this study was also developed by the authors. The two kindergarten measures, equal partitioning and ordinal positioning, demonstrated strong reliability, validity, sensitivity to growth, and diagnostic accuracy. During the equal partitioning measure, students were shown two characters with small dots arranged under them or a series of dots between them. The students were then asked one of two questions. For an array item they would be asked: "How many cookies would they each get if they shared the cookies?" For a shared item students would be asked, "Does it look like they each have an equal share of cookies between them?" Students were asked to point or tell the examiner what place an object was in along a horizontal line on ordinal positioning measures. In terms of measuring student growth, equal partitioning and ordinal partitioning, student performance changed significantly over time and accounted for 40% of the variance in criterion measures when administered in isolation. The combined variance of these two measures accounted for over half of the variance on criterion measures. Two first grade measures, grouping by 5 and verbal facts, demonstrated promising technical and diagnostic accuracy, but more research is needed. The researchers suspect practical application measures may be less reflective in first grade because many first grade curriculums focus on mastery of basic computation skills (Methe et al., 2011).

Assessment tools with an emphasis on conceptual understanding of early numeracy were investigated by VanDerHeyden et al. (2011) with less positive outcomes. The researchers developed six new measures of math assessment to expand on previously established assessments of early numeracy skills. The newly developed measures assessed number sense (ordinality, subitivity, and cardinality), shape recognition, and patterning. Previously-established

measures included in the study were Missing Number (Clarke & Shinn, 2004) and Choose Number, Draw Circles, Write Number measures (VanDerHeyden, 2008). The previously-established measures demonstrated stronger reliability and validity than the newly-developed measures (VanDerHeyden et al., 2011).

Locuniak and Jordan (2008) investigated the relationship between number sense in kindergarten and calculation fluency in second grade, which has substantial implications for universal screening in mathematics. The researchers used block entry regression to examine the relationship between number sense and calculation fluency. In the first block, they accounted for age, reading, memory, and verbal and spatial recognition. The second block was comprised of number sense measures including counting, number knowledge, nonverbal calculation, story problems, and number combinations. Results indicated that, while all the measures correlated positively with each other, number sense measures were able to predict calculation fluency uniquely outside of the measures controlled for in block 1. Number combinations accounted for the most unique variance, indicating children who understand concepts of basic addition and subtraction in kindergarten are less likely to demonstrate math deficits in second grade (Locuniak & Jordan, 2008). These results support the earlier findings of Mazzocco and Thompson (2005). In their 2005 study, the authors found that mental math of basic addition and subtraction facts in kindergarten were predictive of math achievement and the presence of a specific learning disability in second and third grade. The longitudinal study included 209 students (103 males and 106 females) from one of seven elementary schools in a suburban school district. Participants were administered an individual tests of basic math, visual-spatial skills, and reading-related skills, two to three times per year in kindergarten through third grade. The authors found strong ROC values when using the composite score of all given measures

(.90). The four core subtests of the Test of Early Math Ability, Second Edition, yielded a similarly strong ROC value of .88. These findings support early deficits in numeric processing as being highly predictive of MLD and/or low math achievement in second and third grade.

There is some debate in the field whether automaticity of basic math facts should be included as a component of number sense. However, it is well established that poorly developed early numeracy skills impact learning and understanding of arithmetic (Geary, 2004; Mazzocco & Thompson, 2005).

Computation and Fluency

The need for ongoing research to identify general outcome measures in mathematics has been clearly acknowledged. Specific to mathematics, it is recognized there are not clearly defined general outcome measures (Mazzocco, 2003). However, computation skills, including basic math fact fluency, have a growing body of research to support their use as a general outcome measure.

Basic fact fluency has been a moderate indicator of future math outcomes (Geary et al., 2012; Keller-Margulis et al., 2008; Shapiro et al., 2006) and correlates with performance on higher level mathematics procedures (Price, Mazzocco, & Ansari, 2013). Therefore, it is applicable to a large range of ages and grades. The complexity of the arithmetic problems represented on computation CBM varies significantly based on the grade or instructional level being assessed. Computation fluency can be assessed with single or multiple skill probes or cloze procedures. The predictive validity of fluency based computation CBM is the primary focus of this study.

Cloze procedures. Cloze procedures in math were initially explored by Jiban and Deno (2007). The authors studied the predictive validity of high stakes testing in relation to other

commonly-used CBM measures, a 1-minute computation probe, and a 1-minute reading maze measure. Cloze math CBM require students to identify a missing portion of a number sentence (i.e., 2 + __ = 4; __ - 3 = 2). The authors' hypothesized cloze math CBM would have the desirable characteristics of previously established math CBM, but be a better measure of conceptual understanding and application than traditional computation focused probes.

Basic computation fluency. Automaticity of basic math facts is considered a bottleneck skill in mathematics, meaning deficits with fact retrieval can affect many other components of mathematical learning and achievement (Geary, 2004; Geary et al., 2012). Difficulty with automaticity of basic addition and subtraction facts is an early indicator of math learning problems. Retrieval of math facts remained a deficit for students identified as having a specific learning disability in math even when significant progress was made working with algorithms, procedures, and simple word problems (Gary, 2004; Geary et al., 2012; Jordan et al., 2003). Fluency with computational tasks indicates mastery and allows for application of these skills to higher level problem solving (Geary, 2004; Johnson & Layng, 1992; VanDerHeyden & Burns, 2009). Therefore, an argument can be made to support early screening for fact retrieval skills.

Computation measures are composed of basic arithmetic facts. Measures can be single- or multiple-skill measures and the complexity of arithmetic problems vary based on the grade-level. Typically, computation measures are timed to assess the rate and accuracy of performance. These instruments are scored as digits correct per minute or digits correct per the length of the probe when wishing to increase sensitivity (Stevens-Olinger, 2014).

Reliability and validity. Reliability coefficients of .70 or higher are generally thought to be acceptable for universal screening purposes (Burns et al., 2014; Christ & Nelson, 2014).

Foegen et al. (2007) found CBM-M computation to consistently demonstrate adequate internal consistency, test-retest reliability, and alternate form reliability, ranging from .73 to .98.

Christ et al. (2008) examined the reliability of CBM-M Basic Computation Fluency probes within the context of a literature review. Of the eight studies that met inclusionary criteria at the time, CBM-M demonstrated internal consistency coefficients of .80 or higher. The authors noted a significant amount of variation in interrater reliability, ranging from .60 to 1.00. At the time, no studies adequately examined the test-retest reliability of CBM-M.

Cloze CBM accounted for more variance than traditional math computation probes when predicting performance on the state academic assessment, Minnesota Comprehensive Assessment in Mathematics (Jiban & Deno, 2007). However, due to low reliability of cloze CBM, the results of this study should be interpreted with caution. When the authors combined the variance on the math cloze CBM and reading maze, it accounted for 52% of the variance on state academic assessments.

Stevens-Olinger (2014) expanded on the Jiban and Deno (2007) study with further examination of the technical adequacy, instructional effectiveness, and logistical application of cloze math CBM. Computation and cloze math CBMs were administered to 215 third grade students from 12 elementary schools, with varied administration times (1-, 2-, or 3-minutes). All combinations of the CBMs were administered to all students over the course of two days approximately three weeks before the state administered academic achievement test. In contrast to previous findings (i.e., Hintze, Christ, & Keller, 2002; Jiban & Deno, 2007), none of the 1-minute measures were found to be technically adequate. The reliability increased with administration time, with the 3-minute basic computation probe demonstrating the highest reliability (.78 for 1-minute, .81 for 2-minute, and .89 for 3-minute).

In terms of predictability with the state administered academic achievement test, cloze CBM demonstrated higher correlation coefficients than basic computation probes (.53 to .60). However, correlation coefficients with the state assessment were moderate at best for all measures (.43 to .60). Basic computation and cloze procedure accounted for a similar amount of variance on the state assessments but a weak to moderate amount on the computation probe (18.8% to 32.7%) and for the cloze CBM (27.9% to 36.4%).

These results indicate that cloze procedure CBMs are better predictors of future math outcomes. Stevens-Olinger had similar findings when comparing the predictive validity of basic math fact probes, including both one- and three-minute cloze procedure probes. The cloze procedure CBM were a stronger predictor of student performance on the state assessment. It is noted that three minute CBMs scored as digits correct had the highest reliability.

Strait et al. (2015) recognized that CBM are generally thought to be reliable tools for universal screening and progress monitoring students but drew attention to a lack of research examining test-retest and alternate form reliability of CBM math measures. The researchers examined the test-retest and alternate form reliability of the math CBM generated through interventioncentral.com. This information was sought out to determine if the free probes generated on interventioncentral.com have adequate reliability to be used as a progress monitoring tool for math computation intervention. The 283 participating sixth grade students were administered four alternative forms of the generated multiple skill probes, two times each over the course of the fall semester. Probes were completed every two weeks. Test-retest reliability, calculated with Pearson's correlation, ranged from .49 to .75 for fluency scores and .61 to .75 for accuracy scores. Alternate-form reliability, also calculated with Pearson's correlation, ranged from .41 to .81 for fluency scores and .47 to .78 for accuracy scores. A

multilevel linear model was used to calculate test-retest reliability to determine the stability of individual's scores in the context of each individual math class. Level 1 was set as Time 1 or 2 of probe administration, Level 2 as the individual random effect, and the classroom as the Level 3 random effect. Test-retest intra-class correlations (ICC) ranged from .49 to .73 for fluency scoring and .59 to .74 for accuracy scoring. The alternate forms reliability was generated using the same Level random effects as test-retest reliability. Separate alternative wave reliability was calculated for the first and second waves of probe administrations. The first wave ICC estimates for fluency scores were .58 and .57 for accuracy. The ICC estimate for the second series of probe administrations was .73 for fluency scores and .71 for accuracy scores. These results indicate moderate test-retest and alternate form reliability, which is acceptable for screening and progress monitoring purposes. The authors found when two or three measures were aggregated, test-retest and alternate form reliability increased above .80 for both fluency and accuracy scores. The findings of this study indicate the reliability of the fluency and accuracy measure administered improved greatly when two or three scores were analyzed as opposed to one score, suggesting the need to administer multiple probes to each student to increase reliability.

MBSP-C has demonstrated particularly strong reliability estimates. MBSP-C has been shown to demonstrate internal consistency ranging from .94 in third grade to .98 in second grade (Fuchs et al., 1994; Fuchs, Hamlett & Fuchs, 1999). There are 30 alternate forms of the MBSP-C probe. Alternate form reliability ranged from .73 to .93, with the majority of reliability coefficients falling above .80. The sample size of these studies were relatively small, ranging from 7 to 28 students at each grade level (Fuchs, Hamlett & Fuchs, 1998).

Jiban and Deno (2007) investigated the predictive validity of two brief one-minute math CBMs, one a traditional fact fluency probe and one cloze procedure, in third and fifth graders. It

is important to note that this study consisted of a relatively small sample size (third grade $N = 35$; fifth grade $N = 49$), therefore the results should be interpreted with caution. The third grade one-minute math fact fluency probe did not correlate with performance on the state math assessment and yielded a moderate correlation in fifth grade. The alternate form reliability of both one-minute math probes was moderate to moderately strong but did not meet the .80 reliability that is considered ideal. Similar to the findings of Strait et al. (2015), when the researchers aggregated two scores, the reliability of the obtained data improved to an acceptable level. However, this substantially increases the amount of instructional time used and progress monitoring data used to make instructional decisions. It is important to note previous research contrasts these findings. For example, Hintze et al. (2002) found one brief probe demonstrated adequate technical adequacy for the purpose of universal screening.

Predictive adequacy. Christ, Scullin, Tolbize, and Jiban (2008) determined that CBM in math computation was an appropriate universal screening measure. However, the authors recommended the need for more research regarding the validity of CBM computation and technical adequacy. A 2007 study found computation and concept/application measures are predictive of future math outcomes. Computation measures were found to demonstrate technical adequacy for progress monitoring and universal screening (Fuchs et al., 2007). One-minute CBM-M computation probes have demonstrated adequate technical adequacy for the purpose of universal screening when scored as digits correct (Christ, Johnson-Gros, & Hintz, 2005; Hintze et al., 2002). Longer administration times (12-13 minutes) were required to support high-stakes decision making (Christ et al., 2005).

Codding et al. (2015) examine the relationship between CBM and performance on state assessments administered in seventh grade. The researchers made the argument that although

CBM appear to have weak face-validity at the late elementary/secondary level, students still need to have mastered basic skills to access more complicated content knowledge (Codding et al., 2015; Daly et al., 2007). A four-minute, mixed-computation math probe CBM-M was group-administered to 249 seventh graders in the fall, winter, and spring. Data analysis included a multiple indicator parallel process latent growth model to evaluate growth trends. Findings indicated that the achievement gap between high and low achieving students widened as the school year progressed. When growth models were examined over the course of years, the achievement gap lessened in reading, but remained present in mathematics. CBM data were then correlated with performance on a state administered achievement test. The results reflect previous literature; CBM-M had a moderate correlation (.26 to .35) with math achievement on the end of the year state assessment. Spring CBM data correlated more strongly with performance on the state assessment than fall and winter data (Codding et al. 2015).

Early identification of mathematical deficits with measures of computation fluency were explored by Purpura, Reid, Eiland, and Baroody (2015). The researchers drew from previous research supporting the use of fluency-based screening tools to develop a brief measure of discrete skills for pre-school students. Discrete skill measures are generally fluency-based and designed to assess specific mathematical skills. These measures were previously shown to demonstrate good predictive validity and sensitivity to change over time (VanDerHeyden, Broussard, & Cooley, 2006). The researchers identified two potential concerns or limitations when discrete mathematical measures were used to determine future academic risk. Fluency-based mathematical measures correlate highly with non-math related measures such as reading fluency and school readiness measures. This suggests fluency-based measures are assessing non-mathematical constructs in addition to math skills. The second limitation identified by

Purpura et al. is the relatively small number of skills assessed on discrete measures. This is problematic given the absence of consistently agreed upon general outcome measures in mathematics. The findings of this study support the use of a brief measure of mathematical skills followed by a more in-depth broad measure and progress monitoring.

The use of a 5-minute computation probe as a universal screener for math disabilities was examined by Fuchs et al. (2005). The researchers sought to gather more information regarding the prevention, identification, and cognitive determinants of mathematical skills deficits and learning disabilities. Based on CBM data and teacher referral, 319 of the 667 first grade participants were identified as being at-risk for learning problems. The researchers provided high quality math intervention to further differentiate between student groups. Based on week 4 CBM data, the lowest 21% of students (n = 139) were assigned to one of two groups, control group or tutoring condition group. Students assigned to the tutoring group received 48 tutoring sessions with either a 1:2 or 1:3 teacher: student ratio. Math fact fluency CBM data generated unrealistically high prevalence rates of math disabilities, 9.40% of the general population for computation CBM and 6.38% for addition fact fluency. The high prevalence rates generated by the computation and addition fluency probe indicate a high rate of false positives. As previously noted, poor classification accuracy can have a significant impact on the utility of a universal screening measure. Measures with poor technical adequacy and classification accuracy can drain school fiscal and personnel resources (Clarke et al., 2014; VanDerHeyden, 2013).

Shapiro et al. (2006) explored the correlation between universal screening tools in reading and math with a state or standardized academic achievement test. The study took place in two different schools. Data were originally collected to develop local normative data. MBSP-C was only used at District 1. District 2 used a self-developed mixed computation math probe.

Math CBM data from the fall were correlated with a standardized academic achievement test in the spring of the same academic year. The researchers collected universal screening data for first through fifth grade. No statistical analysis were performed to determine correlations with first grade data and performance on the state assessment. Second and fourth grade math screening data were correlated with Stanford Achievement Test scores. Third and fifth grade screening data were correlated with the Pennsylvania System of School Assessment (PSSA). There are no reported correlation data for District 1's MBSP-C in second and first grade. Results indicated moderate correlations between MBSP-C data and PSSA performance in the winter and spring, ranging from .50 to .53. Fall data were the least predictive (.07 to .41). Winter data were the most predictive of PSSA performance in the spring. It is important to note students with an IEP for anything other than Speech/Language support or Gifted education were excluded from the study sample.

Keller-Margulis et al. (2008) sought to gain more information about the relationship between math computation screening tools and the PSSA administered one and two years later. They also evaluated growth rates and classification accuracy between math screening data and PSSA. The researchers used an archived data set originally collected during the 2002-2003 academic year to develop local normative data. Students with an IEP for anything other than Gifted Education or Speech/Language Support were excluded from the sample. This could be a potential weakness when determining the utility of a measure for identifying students in need of additional intervention. Excluding a portion of the general population a universal screener is meant to identify from the sample population will significantly impact analysis of predictive validity, technical adequacy, and diagnostic adequacy.

Results indicated a moderate correlation between MBSP-C in first grade (.50 to .59). The correlation between fall first grade MBSP-C data and PSSA was weak (.27) and PSSA administered in the spring of third grade. Second grade MBSP-C had a moderate correlation with PSSA administered in the spring of third grade (.52 to .60) but a significantly weaker correlation with PSSA administered in fourth grade (.14 to .58). Third grade MBSP-C data had a moderate correlation with the PSSA administered in the spring of the same year (.40 to .49). The correlation between concept/application probes and PSSA followed a similar pattern. The authors concluded that their results support the use of CBM as a universal screening tool for early identification of students in need of additional intervention (Keller-Margulis et al., 2008).

There are several factors which could have impacted the outcomes of the Keller-Margulis, et al. (2008) study. First, math curriculums have changed rather significantly since the data collection period of the Keller-Margulis et al. (2008) study with the adoption of a common core reflecting the NCTM focal points which acknowledge the importance of both automaticity and application of math skills. Data were collected during the 2002-2003 school year. Second, the data represented in the study were collected initially for the purpose of developing local normative data and students with IEPs were excluded, with the exception of Gifted and Speech/Language IEPs. Therefore, the sample is not representative of most school populations which in turn, decreases the relevance of the findings. Classification accuracy was further explored by Clarke et al. (2011) using a significantly smaller but more diverse population.

Clarke et al. (2011) examined the classification accuracy of easyCBM first grade mathematics screening measures. The first grade easyCBM measures are based on NCTM focal point standards and consist of three subsections: Number and Operations and Algebra, Number and Operations, and Geometry. The probes were administered to 145 first grade students from

four different schools in the fall, winter, and spring. Although the sample size of this study is relatively small, it is important to note participating students were recruited from a much larger nationally representative sample of schools who participated in a commonly-used data warehousing system.

The TerraNova 3 was administered in the spring of first grade and served as the criterion measure. The TerraNova 3 is nationally-normed standardized achievement test aligned with state standards, NAEP's framework, and the NCTM's focal standards. EasyCBM demonstrated strong reliability, .78 in the fall, .85 in the winter, and .87 for spring administrations of the assessment. Based on correlation with the TerraNova 3, easyCBM first grade math probes demonstrated adequate concurrent and predictive validity, ranging from .58 in the fall to .72 in the spring. Results of the ROC analysis indicated good levels of specificity and sensitivity at the 25th and 40th percentile (Methe et al., 2015). It is important to note the researchers found stronger positive predictive power when they used a higher cut-score than those published by easyCBM. The published cut-off scores generated a significant number of false positives. These findings highlight the need for local normative data and ROC analysis. Correlation with a previously established measure of mathematical achievement, in addition to sound technical adequacy, indicate easyCBM first grade probes are an appropriate screening tool when identifying students who are at-risk for math deficits.

Another study investigated the technical adequacy of easyCBM with older students. Anderson, Lai, Alonzo, and Tindal (2011) explored the utility of the fifth grade easyCBM math probes as both a universal screening measure and progress monitoring tool for students who were persistently low-performing. Students who were persistently low-performing are defined as those identifying as having a learning disability not classified as "severe" and those achieving

well below grade-level. The researchers administered the probes to students from two mid-sized school districts in the Pacific Northwest. The sample size ranged from 2,085 to 2,099 for each test item. The data were analyzed using a one-parameter Rasch model and graphed by percentage of students who responded correctly to evaluate how the item functioned for different student groups. Items ranged in difficulty from -2.56 to 2.58, with 24 items having a difficulty rating below zero and 24 above zero. This indicated an even sampling of questions students are likely to answer correctly and more difficult questions. This even sampling of questions is an ideal test composition. The difficulty level of the items were evenly distributed across the NCTM focal points, which indicates strong internal consistency. The researchers concluded the easyCBM probes show promise as a screening and progress monitoring tool but caution these results are based on how the items functioned at one point in time. They recommended further longitudinal research to determine how the items function in relation to student growth over time (Anderson et al., 2011).

While current research supports the use of CBM-M as universal screening instruments, Methe, Briesch, and Hulac (2015) question the technical properties of CBM-M for decision making purposes. The authors argue the majority of the research validating CBM-M is correlational in nature and more information is needed regarding the technical properties of CBM-M instruments. The authors recommend the technical adequacy of all commercially available CBM-M be reviewed to ensure classification accuracy. The authors developed math probes with more consistent item content than a commercially available CBM-M and found it to have more stability and better classification accuracy (Methe et al., 2015). Educators and researchers should be willing to revisit previously established measures and practices to minimize classification errors.

Initial research supports the use of computation fluency as a universal screening measure. However, the current research base is preliminary in nature. Further research is needed regarding the technical adequacy of computation measures for the purpose of universal screening and educational decision-making. This study aims to further investigate the predictive validity of MBSP-C when administered in the fall, winter, and spring of first, second, and third grade with the math portion of the state academic achievement test administered in the spring of third grade. If MBSP-C demonstrates a strong predictive relationship with the state assessment, further examination is warranted to determine sensitivity, specificity, positive predictive power, and negative predictive power.

Concepts and Applications

Concept and application CBM focus on a student's ability to use early numeracy skills and/or computation skills to problem solve. The skills assessed on concept and application CBM can vary significantly based on sources of development (robust indicator or curriculum sampling) and grade level. Concept and application probes generally encompass one or more of the following skills: counting, number concepts, names of numbers, reading charts and graphs, geometry, measurement, money, fractions, and word problems.

Reliability and validity. Concept and application measures demonstrate strong reliability, ranging from .81 to .98. Word problem solving measures demonstrate slightly lower reliability coefficients but still acceptable in the moderate range, varying from .60 to .83 (Foegen et al., 2007). Concept and application CBM demonstrate strong criterion validity when correlated with state assessments and nationally-normed standardized tests of academic achievement (Foegen et al., 2007; Jitendra, Dupuis, & Zaslofsky, 2014; Keller-Margulis et al., 2008; Shapiro et al., 2006).

Amselmo (2014) investigated the concurrent and criterion validity of the AIMSweb math concept and application probes (M-CAP) with the state required, North Carolina End-of-Grade Mathematics test for students in seventh grade. Results of this study indicate student performance on the M-CAP administered in seventh grade has a strong correlation with the state mathematics assessment administered at the end of seventh grade (r = .65). The criterion validity of the M-CAP (r = .66) indicate AIMSweb math concept and application probes are predictive of student performance on the End-of-Grade mathematics test. AIMSweb math concept and application probes demonstrated a much higher concurrent and criterion validity for secondary students than the AIMSweb computation probe, which accounted for 4.3% and 4.5%, respectively, of the variance on the North Carolina End-of-Grade Mathematics test.

Predictive adequacy. There is a growing body of evidence to suggest word problems account for unique variance on criterion measures, especially when differentiating between subtypes of math deficits. Shin and Bryant (2015) identified a significant difference between students with a math learning disability (MLD) and those with MLD and reading learning disability (RLD). Students with MLD only consistently outperformed students who had both mathematical and reading deficits on word problems. Although students with MLD scored higher on the word problem measure, there were minimal differences between the two student groups when performance on the mathematical operations test were compared.

There is evidence to support the use of word problem CBM to identify students at-risk of mathematical deficits. Word problem CBMs were found to have a unique variance, separate from math calculation and reading skills when predicting performance on the California Standardized and Reporting test (STAR; Sisco-Taylor et al., 2015). Results of a forced hierarchal regression indicated that two separate calculation measures accounted for 39% of the

variance. Word problem CBM accounted for 8% of the variance in student performance on a high-stakes math assessment. Although math calculation accounted for a larger percentage of variance for overall performance on the mathematical portion of the California STAR, word problem accounted for a significant and additional amount of variance on the criterion. Secondly, the AUC of word problem CBM and California STAR ranged from .80 to .83. AUC of .80 or higher are considered sufficient to identify student who are at-risk and in need of intervention.

Jitendra et al. (2014) also examined the reliability and validity of a word problem solving measure for the purpose of universal screening and progress monitoring. The word problem solving measure was administered to 136 third grade students every two weeks over the course of twelve school weeks. Reliability of the word problem solving measure was moderate (.67 to .71). The word problem solving measure demonstrated weak to moderate validity coefficients (.23 to .64).

Due to limited research studying word problems as a universal screening instrument, more research is recommended. However, word problem CBM will likely play a significant role as gated evaluation systems and universal screening and progressing monitoring practices for older elementary and secondary students become more prevalent. Another area which requires more research but is likely to play a significant role in universal screening procedure is Computer Adaptive Testing (CAT).

Computer Adaptive Testing

The topic of computer administered verse paper-pencil administration was explored by Shapiro, Dennis, and Fu (2015). The researchers compared student performance on a CAT assessment, STAR-Math, with performance on a paper-pencil math CBM, AIMSweb Math

Computation and AIMSweb Math Concepts/Applications (Shapiro et al., 2015). Both measures were administered to between 82-92 third graders, 71-84 fourth graders, and 64-74 fifth graders once a month for a seven month period. The student sample participating in the study included those included in the regular education setting and students identified as having a specific learning disability receiving specially designed instruction in a learning support setting. The researchers used Hierarchical Linear Modeling to compare the two measures. Results indicated all three measures were able to reflect student growth over the course of the seven month period. When given immediately preceding the PSSA, STAR-Math demonstrated the strongest correlation with PSSA performance, with the exception of fifth grade. The computation probes were shown to have the second highest correlation with PSSA, followed by the concept/application CBM (Shapiro et al., 2015). Although more research is needed, initial research supports the use of CAT for universal screening and progress monitoring. This study also supported the use of computation based CBM and concept/application focused CBM for universal screening and progress monitoring purposes.

Summary

Mathematical deficits remain persistent in students who are low-achieving, and the performance gap widens as students continue through school if not addressed with robust instruction and intervention. Early identification and intervention to address mathematical difficulties is paramount. Research indicates students who initially place in the bottom 10th percentile when entering kindergarten but were performing above the 10th percentile upon exiting only had a 30% chance of performing below the 10th percentile five years later while in fifth grade (Morgan et al., 2009, 2011). Without intervention in kindergarten, however, students who demonstrate math skills within the bottom 10th percentile in kindergarten have a 70%

likelihood of remaining below the 10th percentile five years later (Martin et al., 2012; Morgan et al., 2009; 2011). This highlights the needs for early identification and intervention. MTSS/RTI systems employ universal screening measures to identify students who may be at-risk for developing deficits for the purpose of early intervention.

Universal screening measures should be able to identify potential academic, behavioral, or emotional concerns in need of additional assessment or identify students who are at-risk of difficulty. Ideally, universal screening measures should be able to answer the following questions: How is each student responding to core instruction? How many students are at-risk for failure? Is core instruction effective? Which students are in need of additional assessments? What levels of resource support might be needed to promote criterion-level performance?

In terms of data use, effective universal screening tools generate data that are accessible to teachers and can be used to differentiate instruction. Effective implementation of universal screening practices requires an expectation or school culture that teachers use data to align instructional resources. Staff should be provided with training to administer, score, and interpret the results of universal screening. Universal screening measures should be based on universal design. In practice, this translates to an instrument that is given in individual or preferably group format to most students in an entire classroom, grade, school, or district with sources of bias eliminated. Psychometric properties of an adequate universal screening measure require reliability of .70 or higher, sensitivity to changes in student performance, distinction between the proficiency levels of students, and reflection of essential components of the curriculum (Ikeda et al., 2008).

The present study builds on previous research utilizing MBSP-C probes (Fuchs, Hamlett, & Fuchs, 1999) as a universal screening measure. In a 2007 literature review, MBSP-C (Fuchs

et al., 1999) was the most frequently used measurement tool for elementary mathematics and the only tool used in studies focused on the use of CBM data to improve student achievement (Foegan et al., 2007). The state mandated assessment is explored with the predictive validity with a previously established criterion measure. Given inconsistent findings regarding the role of sex and socio-economic status on students' mathematical proficiency, these factors are further investigated.

CHAPTER III

METHODS AND PROCEDURES

Introduction

The predictive validity of a timed, mixed computation math instrument, Monitoring Basic Skills Progress, Computation (MBSP-C) with the Pennsylvania System of School Assessment (PSSA) was examined in this study to determine its predictive strength as a universal screening instrument. The amount of variance in student performance that could be attributed to sex and socio-economic status (SES) or resource availability was also explored. This chapter provides a detailed explanation of the methods used to answer this research question, including the sample, research site, measures used, and procedures employed.

Design

This is a correlational, longitudinal research design. Anonymous archival data were used to determine the predictive strength of a brief computation probe administered in the fall, winter, and spring of first, second, and third grade with PSSA-M performance administered in the spring of third grade. The influence of sex and resource availability on the relationship between the MBSP-C and PSSA-M was also considered given the equivocal results regarding their role in mathematics achievement.

The universal screener utilized in this study assessed computation of mathematical facts with increasing complexity for each grade. The National Resource Council (2002) defines computation or computing as "carrying out mathematical procedures, such as adding, subtracting, multiplying, and dividing numbers flexibly, accurately, and appropriately" (p. 11). Computing supports understanding of math, and it was hypothesized that a computation probe would be a strong measure of performance on a state-administered academic achievement test.

High achievement on the PSSA-M requires an understanding of all five mathematical strands: conceptual understanding, procedural fluency, strategic competence, adaptive reasoning, and productive disposition. Student performance was assessed/measured from first through third grade for multiple reasons. The study began with students in first grade because, even though students begin to develop mathematical understanding prior to entering school, formal instruction of basic math facts is not introduced until the middle or end of kindergarten. The sooner intervention is made available to at-risk students, the better the learning outcome. The opportunity to provide early intervention to first grade students who may be at-risk for low math achievement in third grade was another reason first grade student data were included in this study. Secondly, the validity of computation measures improves dramatically when administered to first grade students as opposed to kindergarten students (Gersten et al., 2012).

To be successful in math, students need to develop all five mathematics strands while in elementary and middle school. There is a substantial amount of literature that supports the development of the strands before the end of third or fourth grades, with established arithmetic skills by the end of third grade (Martin et al., 2012; Morgan, Farkas, & Wu, 2009). Based on this assumption, students should demonstrate adequate computation skills prior to the end of third grade.

The Every Student Succeeds Act (ESSA) mandates standardized academic assessment in third grade through eighth grade. Therefore, third grade is the first opportunity educational systems have to gather criterion data on a large scale without imposing additional assessments. For practical application purposes, student scores are examined while in third grade to increase the opportunity for math intervention before the fourth grade year.

Population

The results of this study were intended for generalization to students in first through third grades in Pennsylvania. The population was based on students in school districts who were in need of a universal screening instrument for the early identification of students who are at-risk for mathematical deficits to provide academic intervention. Given the inclusion of sex and free and reduced meal status, which represents SES as variables in this study, the results should generalize to all students who are of similar SES backgrounds to the sample.

Study Site

The data were collected from a rural school district in Pennsylvania. The school district is not named to safeguard the confidentiality of study participants and the study site. The district consists of three elementary schools serving kindergarten through fourth grade, one intermediate school which houses grades 5 and 6, one middle school for grades 7 and 8, and a high school (grades 9 through 12). Approximately 3,900 students were enrolled in this school district during the 2010 through 2014 school years. Of those students, 11.4% to 12.5% were identified as receiving special education services, which was below the state average of 15.1% to 15.4% (Pennsylvania Department of Education, 2012, 2013, 2015). Less than 1% of students were identified as English Language Learners (ELL; National Center for Educational Statistics, 2015). Approximately 26% of the school population received free or reduced meals. The median household income across the district ranged from $65,169 to $72,422 (U.S. Census, 2014). Complete demographic data for each year archival data were gathered is summarized in Table 1.

Table 1

District Demographic Data for the 2010-2011 Through 2013-2014 School Years

	2010 – 2011	2011 – 2012	2012 – 2013	2013 – 2014
Total Enrollment	3,941	3,941	3,919	3,929
Total Special Education Enrollment	12.5%	12.1%	11.8%	11.4%
Free Lunch	16%	21%	21%	25%
Reduced Lunch	10%	10%	7%	7%
American Indian/Alaska Native	0.6%	0.6%	---	---
Asian	0.7%	0.7%	0.8%	0.6%
Black or African American	3.9%	3.9%	3.2%	3.2%
Hispanic	3.3%	3.3%	4.1%	4.7%
Multiracial	---	---	1.6%	2.2%
Native Hawaiian or Other Pacific Islander	---	---	---	---
White	91.1%	91.1%	90.0%	89.1%

Note. --- n = 10 or less

Sample

Anonymous and archival data from the 2010-2011 through 2013-2014 school years were examined in this study. The same math curriculum and instructional materials were adopted at all three participating elementary schools. As a result, all students received the same instructional content while enrolled in the study site for the aforementioned years. All students contained within the sample received instruction from teachers and school professionals who held the appropriate Pennsylvania teaching certification during the period of archival data collection.

Inclusion Criteria

All archival and anonymous data from students in grades 1-3 enrolled at participating schools during the 2010-2011 through 2013-2014 school years were examined. All students who completed at least one of the three administrations, fall, winter, and spring, of MBSP-C during their first, second, or third school year and took the PSSA-M in the spring of third grade were included in the data set.

Exclusionary Criteria

Exclusionary criteria were based on the availability of the anonymous and archival data. Students without any MBSP-C, all demographic data, or PSSA-M scores were excluded from this study. Data were included for analysis as long as all demographic data (i.e., sex, Individualized Education Program [IEP] status, and free and reduced meal status), PSSA-M performance in third grade, and at least one MBSP-C probe were available. Rates of attrition ranged from 4% within the third-grade comparison to 24% from the first-to-third-grade comparisons. Attrition rates included students with missing data and those who moved out of the study site. Analyses of PSSA-M Total and Geometry were not performed from the 2013-2014 academic year because Geometry was not a reported domain on that year's test (Data Recognition Corporation, 2014). Table 2 and Table 3 display the demographic composition of the sample with and without 2013-2014 data, respectively.

Table 2

Demographics of Sample With 2013-2014 PSSA Data

	First Grade		Second Grade		Third Grade	
	n	%	*n*	%	*n*	%
Total Sample Size	506		815		1205	
Male	274	54%	432	53%	631	52%
Female	232	46%	383	47%	574	48%
Free and Reduced Meal	163	32%	271	33%	422	35%
IEP Status						
Autism Spectrum Disorder	---	---	---	---	---	---
Emotional Disturbance	---	---	10	1%	14	1%
Hearing Impairment	---	---	---	---	---	---
Other Health Impairment	---	---	---	---	---	---
Specific Learning Disability	47	9%	86	11%	120	10%
Speech and Language Disability	18	4%	28	3%	36	3%
Gifted IEP	26	5%	43	5%	74	6%
504 Plan	16	3%	25	3%	36	3%
Ethnicity						
Asian	---	---	---	---	10	0.8%
Black/African American	19	4%	23	3%	40	3%
Hispanic	25	5%	37	5%	54	4%
Multi-racial	---	---	---	---	16	1%
Native American Indian	---	---	---	---	---	---
White	449	89%	738	91%	1082	90%

Note. --- *n* = 10 or less; PSSA-M = Pennsylvania System of School Assessment; IEP = Individualized Education Program.

Table 3

Demographics of Sample Without 2013-2014 PSSA Data

	First Grade		Second Grade		Third Grade	
	N	%	*n*	%	*N*	%
Total Sample Size	272		558		929	
Male	138	51%	290	52%	477	51%
Female	134	49%	268	48%	452	49%
Free and Reduced Meal	79	29%	180	32%	317	34%
IEP Status						
Autism Spectrum Disorder	---	---	17	3%	---	---
Emotional Disturbance	---	---	---	---	11	1%
Hearing Impairment	---	---	---	---	---	---
Other Health Impairment	---	---	---	---	---	---
Specific Learning Disability	21	8%	57	10%	86	9%
Speech and Language Disability	---	---	17	3%	25	3%
Gifted IEP	16	6%	33	6%	64	7%
504 Plan	---	---	16	3%	27	3%
Ethnicity						
Asian	---	---	---	---	---	---
Black/African American	---	---	13	2%	25	3%
Hispanic	---	---	28	5%	40	4%
Multi-racial	---	---	---	---	13	1%
American Indian	---	---	---	---	---	---
White	254	93%	507	91%	844	91%

Note. --- *n* = 10 or less; PSSA-M = Pennsylvania System of School Assessment; IEP = Individualized Education Program.

Assignment

This study used a convenience sample of anonymous, archival data. Students were assigned to data groups based on grade level, sex, and free and reduced meal status. The PSSA technical manuals from 2010 through 2014 were reviewed for any significant changes that would prevent aggregation of the data from multiple years into one data set. The PSSA remained relatively unchanged for third grade students from 2010 through 2012 (Data Recognition Corporation, 2011, 2012). During the 2012-2013 school year, Pennsylvania officially adopted the Pennsylvania Core Standards (PCS), which is reflected in the format of the PSSA, but not content (Data Recognition Corporation, 2013). The introduction of an English Language Arts

assessment to replace PSSA Reading and PSSA writing was the most significant change for third grade students in 2012-2013 (Data Recognition Corporation, 2014). There were no significant changes to the 2012-2013 PSSA-M with the adoption of PCS. Given that the PSSA-M did not substantively change from 2010 to 2013, outcome data from multiple years were combined into one data set. Please refer to Table 4, which depicts cohort data for each year and grade MBSP-C was collected and the year each cohort took the PSSA.

The 2014 PSSA technical manual indicated Geometry as a reported domain, but no Geometry scores were reported for PSSA-M administered to third grade students in 2014. According to the PSSA technical manual this was due to the transition from the previously used standards to the newly adopted PCS. The technical manual provided the following explanation, "However, the scores for 2014 had to align to both the current set of standards and the next set of standards. As such, there were some strands that ended up with zero items eligible for use in mathematics and therefore no items for those strands were selected for the 2014 cores" (Data Recognition Corporation, 2014, p. 266). Consequently, PSSA data from the 2013-2014 year were excluded from the data set for analysis of PSSA-M Composite score and Geometry data. Table 5 provides a visual summary of what data were collected for each school year included in this study. All other scores on the 2014 PSSA-M were comparable to their counterpart scores in the previous PSSA-M administrations included in this study.

Table 4

Student Cohort Data for Years and Grade of MBSP-C and Year of PSSA Administration

MBSP-C	Grade	Year of PSSA administration (3rd grade)
2010 – 2011	1	2013
2011 – 2012	1	2014
2010 – 2011	2	2012
2011 – 2012	2	2013
2012 – 2013	2	2014
2010 – 2011	3	2011
2011 – 2012	3	2012
2012 – 2013	3	2013
2013 – 2014	3	2014

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics

Table 5

Summary of Data Collected by Year and Grade

Year/ Grade	MBSP-C Fall	MBSP-C Winter	MBSP-C Spring	PSSA-M	Numbers and Operations			Measurement		Geometry		Algebraic Concepts		Data Analysis and Probability
					A.1	A.2	A.3	B.1	B.2	C.1	C.2	D.1	D.2	E.1
10-11 1st Grade	X	X	X	X	X	X	X	X	X	X	X	X	X	X
11-12 1st Grade	X	X	X	X	X	X	X	X	X	No Data	No Data	X	X	X
10-11 2nd Grade	X	X	X	X	X	X	X	X	X	X	X	X	X	X
11-12 2nd Grade	X	X	X	X	X	X	X	X	X	X	X	X	X	X
12-13 2nd Grade	X	X	X	X	X	X	X	X	X	No Data	No Data	X	X	X
10-11 3rd Grade	X	X	X	X	X	X	X	X	X	X	X	X	X	X
11-12 3rd Grade	X	X	X	X	X	X	X	X	X	X	X	X	X	X
12-13 3rd Grade	X	X	X	X	X	X	X	X	X	X	X	X	X	X
13-14 3rd Grade	X	X	X	X	X	X	X	X	X	No Data	No Data	X	X	X

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics; A.1, A.2., A.3, B.1, B.2., C.1, C.2, D.1, D.2, and E.1 denote standard anchors. X indicates data were available and collected. No Data indicates data were not available because it was not a reported by the PSSA.

Measurement

Dependent variable

Mathematics achievement was measured by the math composite score on the PSSA-M instrument in addition to scores on the five subtests that are aggregated into the composite score. Mathematics achievement was divided into five subtests established by the PSSA: Numbers and Operations, Measurement, Geometry, Algebraic Concepts, and Data Analysis and Probability (State Board of Education, 2015). The dependent or criterion variables of the current study include the PSSA-M Numbers and Operations achievement score, the PSSA-M Measurement achievement score, the PSSA-M Geometry achievement score, the PSSA Algebraic Concepts achievement score, the PSSA-M Data Analysis and Probability achievement score, and the PSSA-M Composite score. Data were obtained from archival school records.

Numbers and Operations is a subtest of the PSSA-M in which students demonstrate an understanding of numbers, ways of representing numbers, understanding of relationships among numbers and number systems, comprehension of meanings of operations, use of operations and understanding how they relate to each other, and the ability to make estimates. Students are also expected to solve computation problems with accuracy and fluency. The Measurement subtest assesses student understanding of measureable qualities of objects and units of measure. Students are expected to demonstrate the ability to measure objects using appropriate tools and techniques. This subtest includes calculation of time and elapsed time, length, area, volume and weight of objects, and use of a ruler. Geometry is a subtest of the PSSA-M which measures a student's ability to evaluate the defining features of two- and three- dimensional geometric shapes. Students are also expected to exhibit knowledge of geometric shapes, the relationships between geometric shapes, and concepts of symmetry and transformations. Algebraic Concepts

is a subtest of PSSA-M which measures a student's ability to demonstrate an understanding of patterns and relationships between numbers and their functions. Students are also expected to use numbers, symbols, words, tables, and graphs to investigate mathematical situations. The Data Analysis and Probability subtest of PSSA-M requires students to organize, display, interpret or analyze data in order to answer mathematical questions (Data Recognition Corporation, 2014).

Test protocols are returned to Pennsylvania Department of Education (PDE) to be scored. Individual student raw scores are converted to scaled scores for final reporting. Raw scores are converted to scaled scores in a two-step process. First, raw scores are converted to Rasch abilities. The Rasch model is a form of item response theory which takes into account the difficulty of each response item. Rasch logits are not reported because the use of negative numbers and decimals makes them more difficult for a layperson to understand. Therefore, Rasch scores are then converted to scaled scores using linear transformation techniques (Data Recognition Corporation, 2014).

Scores on the PSSA are divided into four performance level descriptors, Below Basic, Basic, Proficient, and Advanced. A score falling within the Below Basic level indicates partial and selective understanding of the skills represented on the third grade mathematics portion of the PSSA. A Basic score suggests a student is able to solve basic and routine problems through application of the skills covered in the third grade of the PSSA-M (State Board of Education, 2015). Students at this performance level demonstrate the ability to use place-value to round, order, and add and subtract whole numbers without regrouping. Further, students performing at the Basic level can typically solve simple computation problems, identify fractions, and match mathematical equations with real world situations. Students at this performance level are able to

read and interpret data represented in visual displays, tell time with an analog clock, measure lengths, and count money.

A score in the Proficient range suggests a student demonstrates problem solving of practical and real-world problems. Skills characteristically demonstrated by students who perform within the Proficient range include use of place value to add, subtract, and multiply whole numbers; application of computation skills to solve word problems; understanding and identification of fractions; calculation of elapsed time; capacity to round monetary amounts; ability to measure and estimate mass, length, and liquid volume; and ability to organize, display, and translate visually represented data to solve problems. A scaled score within the Advanced range on the third grade PSSA indicates a student is able to demonstrate complex problem solving and an in-depth understanding of the skills, concepts, and procedures encompassed in the five reporting categories that compose the PSSA-M. Typically students performing at the Advanced level are able to can use addition, subtraction, multiplication, and division to solve multistep word problems; represent fractions multiple ways; explain arithmetic patterns; use symbols to represent unknown quantities; solve for missing values; apply order of operations when problem solving; and solve for area. Students also demonstrate the skills required to calculate change, elapsed time, and use units of measure to display data and problem solve (State Board of Education, 2015).

As previously noted, raw scores are converted into scaled scores using the two-step process. Performance levels are then determined based on cut scores generated from the range of possible scaled scores. Table 6 depicts the cut-offs for scaled scores in each descriptive category.

Table 6

PSSA Descriptive Category Cut-off Scores

	2010-2011	2011-2012	2012-2013	2013-2014
Advanced	1370	1370	1370	1370
Proficient	1180	1180	1180	1180
Basic	1044	1044	1044	1044
Score Range	750-1832	750-1843	750-1859	750-1914

Note. Score Range represents the range of possible scores; PSSA = Pennsylvania System of School Assessment.

Reliability of PSSA. Overall consistency of the PSSA is moderate to strong with a reliability coefficient of .76. Decision consistency for classification scores on third grade mathematics is strong with reliability coefficients that ranged from .84 to .98. Decision consistency likelihood that a student will be classified as proficient or not on another version of the PSSA or test measuring the same skills. The inter-rater reliability of third grade PSSA-M is strong with reliability coefficients that ranged from .91 to .96 (Data Recognition Corporation, 2011, 2012, 2013, 2014). These reliability coefficients indicate the third grade PSSA-M is a reliable measure.

Validity of PSSA. The validity of the PSSA-M ranged from moderate to strong when correlated with the Stanford Achievement Test (SAT) correlation coefficients ranged from .70 to .90. The internal consistency of the PSSA-M is moderate to strong (.55 to .95). The content of the PSSA-M was also found to be strongly linked to eligible content, the Pennsylvania Common Core Standards, which indicated strong content validity. A case can also be made for moderate to strong consequential validity, as the percentage of students who scored within the proficient or advanced range has improved consistently since 2007 (Data Recognition Corporation, 2014).

Independent Variables

There are no manipulated independent variables in this study. However, the MBSP-C data, sex, free and reduced meal status, and grade will act as independent predictor variables

when performing statistical analysis. All data are archival, and there is no treatment. Given the nature of the study, the MBSP-C data at any given assessment period across first through third grade are predictor variables.

Monitoring Basic Skills Progress. Monitoring Basic Skills Progress (Fuchs, Hamlett, & Fuchs, 1998, 1999) are a series of parallel form progress monitoring probes available for computation (MBSP-C) and Concepts and Application (MBSP-CA). MBSP-C was used in this study. A review of the existing literature indicates MBSP-C demonstrates favorable qualities of a universal screening instrument. However, due to concerns with small sample size and exclusion of students with disabilities from previous studies, more research is needed to confirm favorable characteristics. MBSP-C offers 30 parallel forms for use in grades 1 through 6. Fuchs et al. (1998, 1999) developed the measures by selecting a sampling of computation problems represented within Tennessee's state standards for each grade level. Probes can be group administered with standardized directions. Administration time is 2 minutes for grades 1 and 2, 3 minutes for grades 3 and 4, 5 minutes for grade 5, and 6 minutes for grade 6.

The skills represented on each MBSP-C remain consistent for each parallel form and sample computation skills students are expected to master throughout that grade level. The first grade computation probes consist of nine basic addition problems (i.e., 3 + 2 =), two addition problems with three addends (i.e., 1 + 3 + 4 =), two addition without regrouping problems (i.e., 33 + 4 =), 10 basic subtraction problems (i.e., 9 – 8 =), and two subtraction without regrouping problems (i.e., 44 – 3 =). The second grade probes consist of seven basic addition fact problems with two or more addends, three addition without regrouping problems, two addition with regrouping problems (i.e., 45 + 38 =), seven basic subtraction problems, three subtraction without regrouping problems (i.e., 45 – 4 =), or three addition with regrouping problems (i.e., 34

– 8 =). Third grade computation probes consist of three addition with regrouping problems, three subtraction with regrouping problems, two subtraction with regrouping using 0 (i.e., 407 – 298 =), nine basic multiplication fact problems (i.e., 4 x 4 =), two multiplication with regrouping problems (i.e., 56 x 4 =), and six problems of basic division facts (i.e., 42 ÷ 6 =). While all alternative forms of MBSP-C probes contain the same type and amount of each problem for each grade level to ensure a consistent level of difficulty, the order they are presented is randomized (Fuchs et al., 1998, 1999).

MBSP-C can be scored for problem correct or digits correct. This study focuses on digits correct because it is more sensitive to growth and change over time. Fuchs et al. (1998, 1999) published normative digits correct scores for the MBSP-C probe. According to these data, average scores (25^{th} to 75^{th} percentile) in the fall of first grade range from five to 12 digits correct, eight to 20 digits in the winter, and 11 to 25 digits correct in the spring. In the fall of second grade, average digits correct range from seven to 14, average digits correct in the winter ranges from 13 to 24, and 16 to 31 digits correct in the spring. Average digits correct in third grade ranges from nine to 19 in the fall, 13 to 26 in the winter, and 22 to 37 in the spring. Normative digits correct scores for MBSP-C are summarized in Table 7.

Table 7

MBSP-C Normative Digits Correct Scores for First Through Third Grade

Grade	Percentile	Fall	Winter	Spring
1	25^{th}	5	8	11
	50^{th}	8	14	17
	75^{th}	12	20	25
2	25^{th}	7	13	16
	50^{th}	11	19	23
	75^{th}	14	24	31
3	25^{th}	9	13	22
	50^{th}	13	21	29
	75^{th}	19	26	37

Foegen et al. (2007) identified MBSP-C as a curriculum sampling measure because it was developed from a set of state standards (Fuchs et al., 1999; Jiban et al., 2007). However, MBSP-C demonstrates characteristics of a robust indicator or general outcome measure. Therefore, it can generalize to more educational settings than just those adopting the state curriculum from which it was developed. General outcome measures in mathematics are an area that requires more research. There is a consensus in existing research that basic computation and arithmetic skills are bottleneck skills which demonstrate moderate to strong correlations with future math achievement outcomes (Geary, 2004; Geary, et al., 2012; Mazzocco, Devlin, & McKenney, 2008; Mazzocco & Thompson, 2005).

Reliability of MBSP-C. MBSP-C has consistently demonstrated strong test-retest reliability in students with (.73 to .92) and without disabilities (.73 to .88), indicating universal design. Correlations for aggregated odd/even scores are also strong in students with (.91 to .97) and without disabilities (.81 to .88). It is important to note a small sample size (n = 79) of students with disabilities was used to establish reliability. Of the 75 students, 54 had been identified as having a specific learning disability and 25 students had been diagnosed as having a behavior disorder. A larger sample size was used to determine reliability in students not identified as having a disability (n = 1,145). Both studies included students in first through sixth grade. A literature review indicated MBSP-C consistently demonstrated strong reliability coefficients for internal consistency (.94 to .98) and alternate form (.73 to .93; Foegen et al., 2007).

Validity of MBSP-C. Five major types of validity should be considered when selecting a universal screening measure: content validity, validity based on response processes, internal structure validity, validity based on relations to other variables, and consequential validity

(Albers & Kettler, 2014). The validity of MBSP-C was assessed by its authors through content and criterion validity analyses (Fuchs et al., 1999). The content validity was reviewed by regular education teachers, special education teachers, and curriculum supervisors from four school districts. There was a recommendation from one school district that one problem be deleted from two grade levels. Based on this feedback, the authors determined the content validity of MBSP-C to be adequate.

To determine criterion validity, the authors correlated the results of MBSP-C probe with three previously established math measures, *Math Computation Test* (MCT) (Fuchs, Fuchs, Hamlett, & Stecker, 1991) and two subtests of the *Stanford Achievement Test* (SAT). This analysis was conducted using the data from 65 students who had been identified as having mild to moderate disabilities, 50 with a specific learning disability, and 15 with an emotional disturbance (Fuchs et al., 1999). The sample consisted of students in second through fifth grade. The results, summarized in Table 8, indicate a moderate to strong correlation between student performance on MBSP-C and previously established measures of mathematical achievement. Due to the small sample size, additional study of the criterion validity of this instrument is warranted; hence the purpose of the current investigation.

Table 8

Validity of MBSP-C

Group	*N*	MCT-PROB	MCT-DIG	SAT-NC	SAT-MC
Grade 2	10	**.91**	**.84**	**.88**	**.93**
Grade 3	19	**.81**	**.87**	**.67**	**.55**
Grade 4	24	**.89**	**.84**	.49	**.60**
Grade 5	12	**.66**	**.77**	.59	.59
Total	65	**.82**	**.88**	**.66**	**.67**

Note. MCT = *Math Computation Test*; PROB = problems correct; DIG = digits correct; SAT = Stanford Achievement Test; NC = Concepts of Number subtest; MC = Math Computation subtest; numbers in boldface are statistically significant.

The validity of MBSP-C was further explored by Shapiro et al. (2006). MBSP-C administered in third grade demonstrated moderate correlations with the third grade state assessment (.41 to .53) and strong positive predictive power (.68 to .88). Given the nature of universal screenings, a validity coefficient alpha of .70 or higher is considered acceptable, so these findings are considered to support the use of computation measures as universal screening instruments (Albers & Kettler, 2014).

Sex and free or reduced meal status. All independent variables are continuous with the exception of sex and free or reduced meal status, which are nominal. Sex is unchanging and free or reduced meal status, which represents resource availability, is not easily changed, but both are likely to have a significant effect on MBSP-C and PSSA-M performance. An analysis of the U.S National Assessment of Educational Progress (NEAP) from 1990 to 2003, found that sex gaps within math achievement continue to exist, with males performing slightly better than females, especially in the upper end of score distributions. Performance sex gaps were largest in the areas of measurement, number and operations, and geometry. The same analysis found significant disparities between achievement gaps in different socio-economic groups (McGraw, Lubuenski, & Strutchens, 2006).

Previous research indicates that math performance is significantly related to sex and free or reduced meal status which represents SES or resource availability. Consequently, these variables were included with the MSBP-C to fully appraise the relationship between relevant dependent variables and PSSA-M performance.

Procedure

Archival anonymous data were examined for this study. Student data, including MBSP-C scores for the fall, winter, and spring across first through third grades, third grade PSSA-M

scores, sex, and free or reduced meal status for the 2010-2011 through 2013-2014 school years were gathered by the school district's data secretary and provided to the primary researcher with all identifying information removed. Student data sets were assigned generic numerical codes, i.e., 1, 2, 3, etc. to ensure anonymity. At no point was the primary researcher given access to identifying information.

Data Collection

Universal screening data collection occurred three times per year: fall, winter, and spring of every year for all students. Within each testing period, all students were assessed within a two-week testing window. Students who were absent during group administration of a MBSP-C probe completed the screening instrument with the intervention support teacher in a small group or one-on-one setting within that two-week testing window. Administration at all three elementary schools was performed by the same teacher who used the scripted directions provided in Appendix B. The PSSA was administered in the spring of third grade, during the testing window provided by PDE. The testing windows of the archival PSSA data were: March 14th-25th for the 2010-2011 school year, March 12th-23rd for the 2011-2012 school year, April 8th-18th for the 2012-2013 school year and March 17th-28th for the 2013-2014 school year.

Beginning with the 2012-2013 school year, all teachers who proctored the PSSA were required to complete an online training program developed and monitored by PDE to maintain the integrity of standardized administration. The online training consisted of three interactive modules followed by a quiz. The three modules were Preparing to Administer the Assessment; Administering the Assessment; and After the Assessment. Modules were completed in sequential order. At the conclusion of the training modules, participants were prompted to take a quiz. After completion of the quiz, a certificate verifying participation and achievement was

generated. This certificate was printed and submitted to the school assessment coordinator. Prior to the 2012-2013 school year, the same content was provided to teachers by building administrators in the forum of a facility meeting using a PowerPoint presentation developed by PDE. The building administrators were required to attend a PSSA administration training prior to reviewing the PowerPoint with staff. PDE required building administrators to document that the information in the PDE-developed PowerPoint presentation was disseminated to staff proctoring the PSSA. This building administrator facilitator training is still provided in addition to the online training program. These requirements by PDE, to which the district complied, increase the likelihood that the PSSA was administered in a standardized manner and strengthen the integrity of data used for this investigation.

Data Analyses

Correlations were generated for the PSSA-M composite score, five subtest scores, and MBSP-C to determine the strength of the relationship between all scores. This analysis provided more specific information regarding MBSP-C predictive power.

Regression analysis was used to determine the regression equation. This is then used to predict a score on the bases of one or more other score. Multiple linear regression (MLR) analysis designates a linear relationship between one or more predictor variable and the criterion variable. A linear relationship means a straight line can be drawn through the data representing the relationship between the predictor and criterion variable.

Research Question

The broad research question under investigation in the current study is: To what extent does a universal mathematics screening, MBSP-C, in first, second, and third grade, sex, and free or reduced meal status predict math achievement as reported on the five components of the

PSSA-M in third grade? It is hypothesized that MBSP-C scores in first, second, and third grade will predict math achievement as measured by the five components of the PSSA-M in third grade. It is hypothesized that student performance in the fall of first grade will have the weakest correlation with PSSA-M performance and student performance in the spring of third grade will have the strongest correlation with third grade PSSA-M achievement due to time proximity between MBSP-C and PSSA-M administration. When validating a number sense screening tool for use in kindergarten and first grade, Jordan et al. (2010) found a significant increase in the main effect over the course of six administrations as students demonstrated age-appropriate changes in achievement.

The variables considered within this research question included MBSP-C – Fall, Winter, and Spring in first, second, and third grade; PSSA-M Numbers and Operations score; PSSA-M Measurement score; PSSA-M Geometry score; PSSA-M Algebraic Concepts score; PSSA-M Data Analysis and Probability score; and PSSA-M Composite score. It is hypothesized that MBSP-C will have the strongest correlation with the numbers and operations portion of the PSSA-M. The numbers and operations section of the PSSA-M asks students to demonstrate an understanding of numbers, ways of representing numbers, relationships among numbers and number systems, meanings of operations, use of operations and understanding how they relate to each other, the ability to compute accurately and fluently, and the capacity to make reasonable estimates. These skills closely resemble those assessed on the MBSP-C probes. Therefore, the relationship between MBSP-C and the Numbers and Operations subtest of the PSSA-M will be the strongest. It is further hypothesized that sex and free or reduced meal status will have a moderate association with math achievement, based on highlights from the 2007 Trends in International Mathematics and Science Study (Gonzales et al., 2009).

Assumptions

Several statistical assumptions were confirmed to appropriately utilize multiple regression analysis (Huck, 2008). It is assumed the dependent variable in this study is measured on a continuous scale. It is assumed the data from independent variables are continuous or categorical in nature. It is assumed that data were independently observed and collected, that the relationship between the variables is linear, and that multicollinearity of the data is not present. It is further assumed that data demonstrated homoscedasticity with no significant outliers or high leverage points, and residuals were normally distributed.

The following steps were taken to determine that these assumptions were met: The data were reviewed for outliers, descriptive statistics were reviewed, histograms were inspected for normality, pairwise comparisons within a scattergram were examined, the correlation matrix was examined for multicollinearity, and the Durbin-Watson statistic was calculated.

Stepwise multiple linear regression analyses were run to examine the predictive relationship of the MBSP-C probe and PSSA-M performance in the spring of third grade. MBSP-C, free and reduced meal status, and sex functioned as the independent or predictor variables. PSSA-M is the dependent or criterion variable. Predictor variables are entered into a stepwise MLR based on their ability to account for specific variance in the outcome variable or the variance that is not already predicted by predictor variables that are already entered in the equation (Leary, 2001). The predictor variable with the strongest correlation to PSSA-M was entered first during the stepwise MLR. The predictor variable entered in step 2 is the variable which accounts for the most variance beyond what has already been accounted for by the first predictor variable. The third predictor variable added accounts for the variance beyond the

predictor variables entered in step 1 and step 2. The beta weights will be examined to determine which variable is a stronger predictor once the assumptions have been confirmed.

Summary

This chapter provides the methods and procedures used to answer the research question examining the predictive power of a computation probe, MBSP-C, sex, and free and reduced meal status in first, second, and third grade with the state assessment administered in the spring of third grade. Descriptions of participant selection, demographics of the population and sample, research site, measures used, and procedure are provided. The study design and data analysis are depicted, as well as possible study limitations.

Table 9

Research Question, Hypothesis, and Variables

Research Question	Hypothesis	Variables
To what extent does a universal mathematics screening, Monitoring Basic Skills Progress – Computation (MBSP-C), in first, second, and third grade, sex, and socio-economic status predict math achievement as reported on the five math components of the Pennsylvania Standard of State Assessments (PSSA-M) in third grade?	It is hypothesized that MBSP-C scores in first, second, and third grade will predict math achievement as measured by the five components of the PSSA-M in third grade. It is hypothesized that student performance in the fall of first grade will have the weakest correlation with PSSA-M performance and student performance in the spring of third grade will have the strongest correlation with third grade PSSA-M achievement. It is also hypothesized that MBSP-C will have the strongest correlation with the numbers and operations portion of the PSSA-M. It is hypothesized that sex, resource availability will be associated with math achievement.	• MBSP-C digits correct from Fall of first, second, and third grade • MBSP-C digits correct from Winter of first, second, and third grade • MBSP-C digits correct from Spring of first, second, and third grade • Third grade PSSA-M Composite Score • Third grade PSSA-M Numbers and Operations subtest score • Third grade PSSA-M Measurement subtest score • Third grade PSSA-M Geometry subtest score • Third grade PSSA-M Algebraic Concepts subtest score • Third grade PSSA-M Data Analysis and Probability subtest score • Sex • Free or reduced lunch status

CHAPTER IV

DATA AND ANALYSIS

The strength of the predictive relationship between Monitoring Basic Skills Progress, Computation probe (MBSP-C), a math computation probe, administered in the fall, winter, and spring of first, second, and third grade with the Pennsylvania System of School Assessment in Mathematics (PSSA-M) administered in the spring of third grade was examined in this study. The predictive relationship of sex and socio-economic status (SES) with mathematical performance was also studied. The predictive validity of MBSP-C administered in the fall, winter, and spring of first, second, and third grade, sex, and SES was analyzed to determine the predictive validity of MBSP-C and capacity to function as a universal screening instrument in mathematics. The results of this research question are presented in this chapter.

Results of Statistical Analysis

Complications

There are two minor complications that should be acknowledged. The first, second, and third grade cohorts are composed primarily of the same group of students. However, they are not identical. This means that the third grade PSSA-M data associated with each grade cohort is different. In order to reflect these differences, separate analyses were completed for each grade level predicting third grade PSSA-M. Therefore, the appropriate corresponding PSSA-M data is reported at each grade level. Differences in student cohorts from grade to grade can be explained by student movement in and out of the school district. Second, geometry was not a reported category on the 2014 third grade PSSA-M. As a result, PSSA-M composite and geometry data from the 2013-2014 school year were excluded from analysis.

Test of Assumptions for Statistical Procedures

Data were analyzed with multiple linear regression (MLR). Assumptions include the dependent variable in this study is measured on a continuous scale. It is assumed the data from independent variables are continuous or categorical in nature. It is assumed that data were independently observed and collected, that the relationship between the variables is linear, and that multicollinearity of the data is not present. It is further assumed that data demonstrated homoscedasticity with no significant outliers or high leverage points, and residuals were normally distributed.

The following steps were taken to determine that these assumptions were met: The data were reviewed for outliers, descriptive statistics were reviewed, histograms were inspected for normality, pairwise comparisons within scatterplots were examined, histograms of residuals and Normal P-P plots were examined for multicollinearity, and the Durbin-Watson statistic was calculated.

Performance on PSSA is reported as interval data thus meeting the assumption that the dependent variable is continuous. All independent variables are continuous in nature, with the exception of sex and free and reduced lunch status which are nominal data. Descriptive statistics are reported by grade of MBSP-C administration and corresponding third grade PSSA-M data. The data were examined visually with the use of frequency distributions and histograms. It is important to note, the PSSA is a criterion referenced assessment. Therefore, a negatively skewed distribution is expected and even desired since it suggests students successfully obtained the content being taught (Brown, 1997; Osborne, 2013). It should also be noted, much higher skewness and kurtosis values are commonly accepted in research relating to social sciences (Byrne, 2010).

First grade descriptive statistics. Please refer to Figures 3 through 11 for histograms of MBSP-C data in the fall, winter, and spring of first grade in addition to the corresponding third grade PSSA-M composite data and PSSA subtest math data.

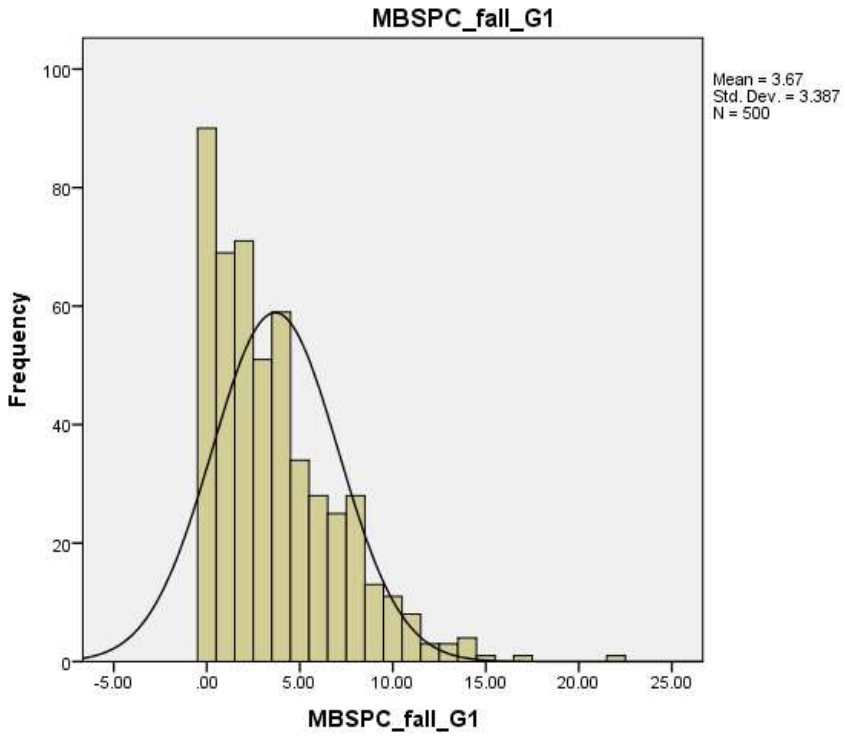


Figure 3. Histogram of first grade Monitoring Basic Skills Progress-Computation fall data.

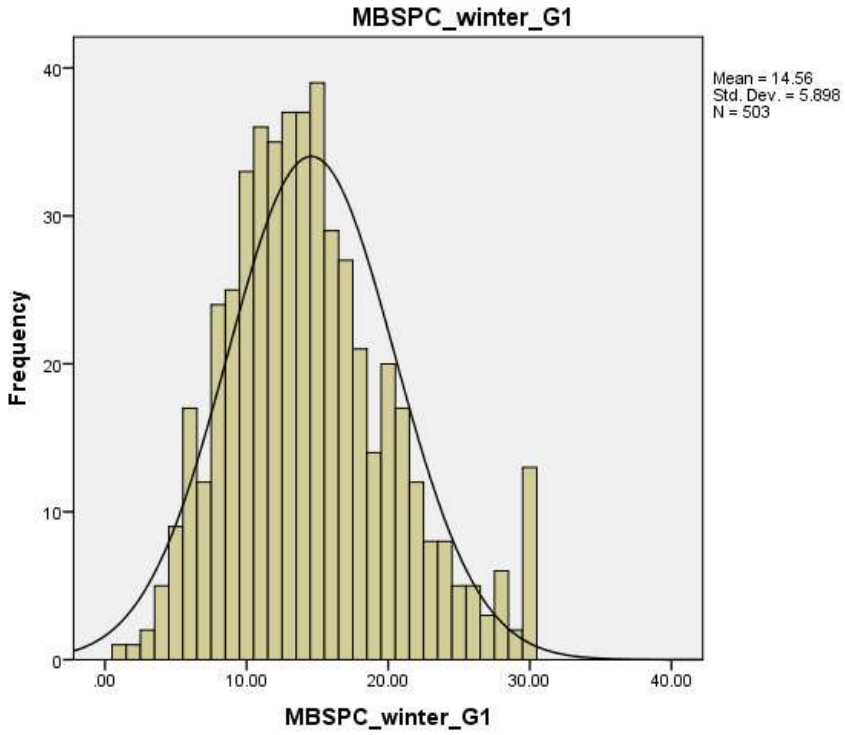


Figure 4. Histogram of first grade Monitoring Basic Skills Progress-Computation winter data.

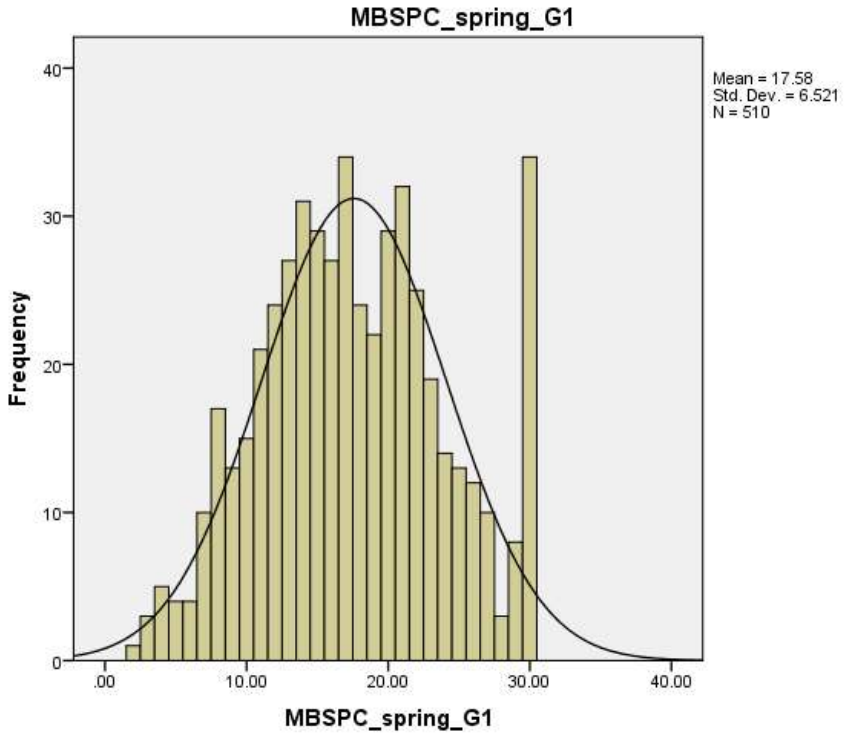


Figure 5. Histogram of first grade Monitoring Basic Skills Progress-Computation spring data.

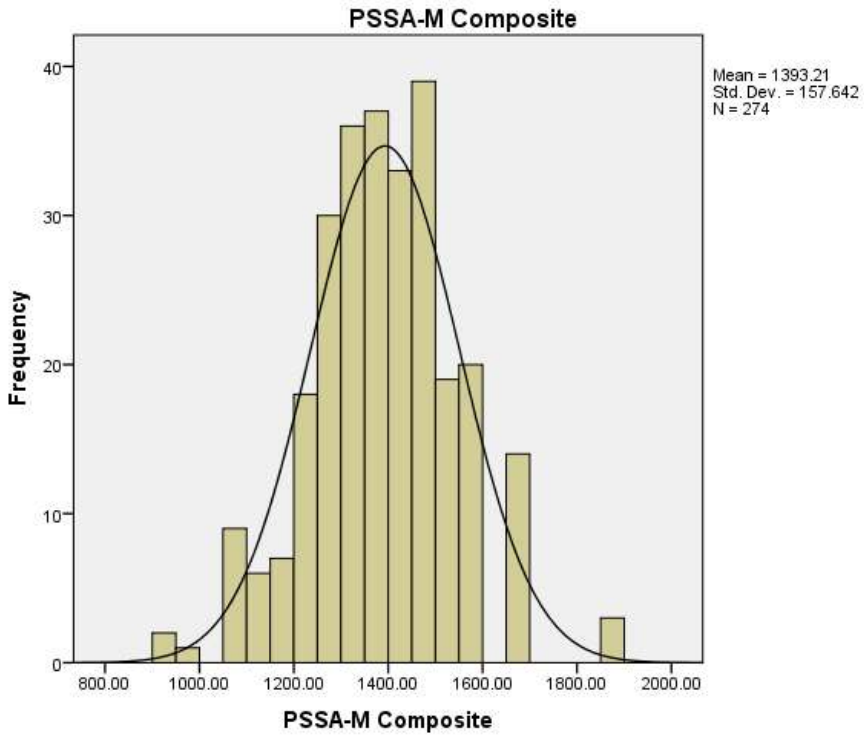


Figure 6. Histogram of third grade 2013 Pennsylvania System of School Assessment, Mathematics composite data from students in first grade during the 2010 – 2011 school year.

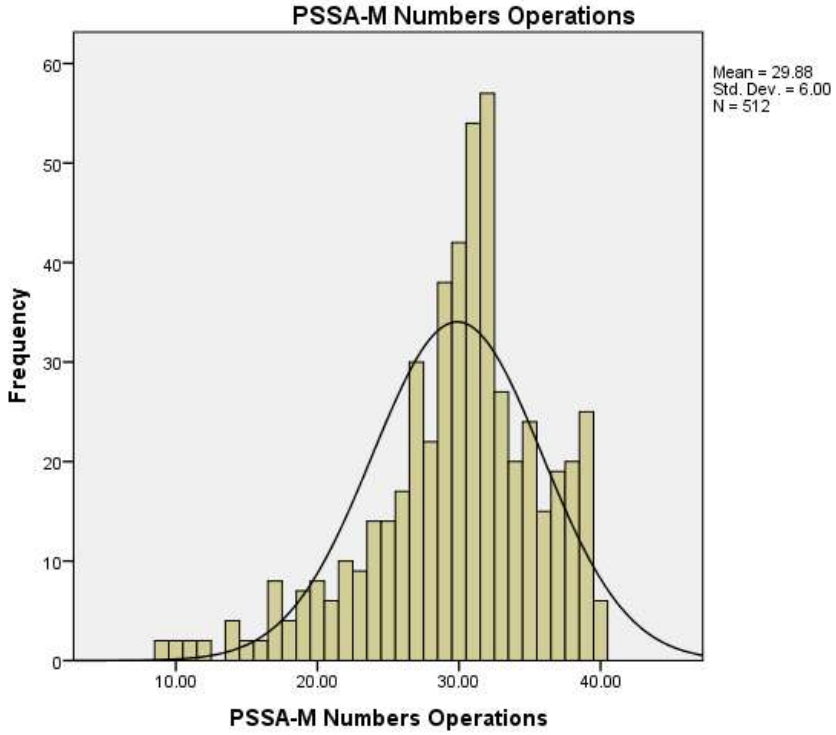


Figure 7. Histogram of third grade 2013 and 2014 Pennsylvania System of School Assessment, Mathematics Numbers and Operations data from students in first grade during the 2010 – 2011 and 2011 – 2012 school years.

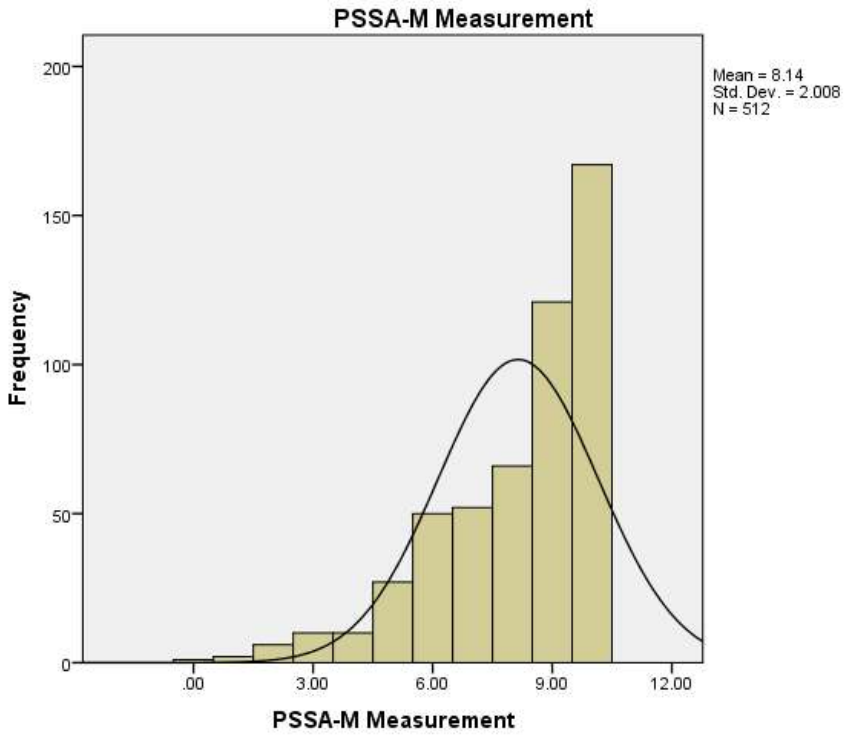


Figure 8. Histogram of third grade 2013 and 2014 Pennsylvania System of School Assessment, Mathematics Measurement data from students in first grade during the 2010 – 2011 and 2011 – 2012 school years.

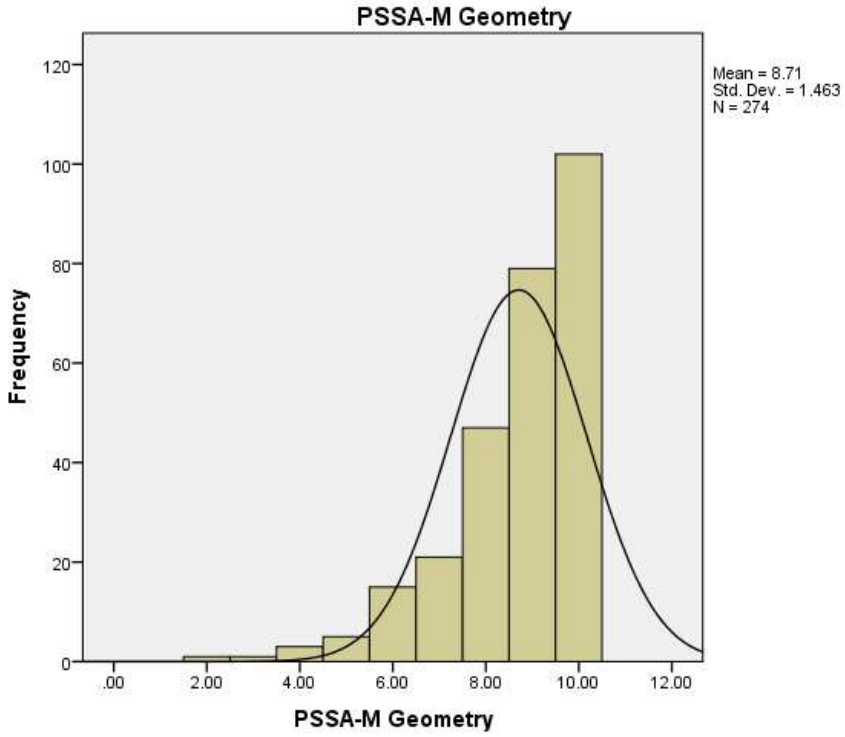


Figure 9. Histogram of third grade 2013 Pennsylvania System of School Assessment, Mathematics Geometry data from students in first grade during the 2010 – 2011 school year.

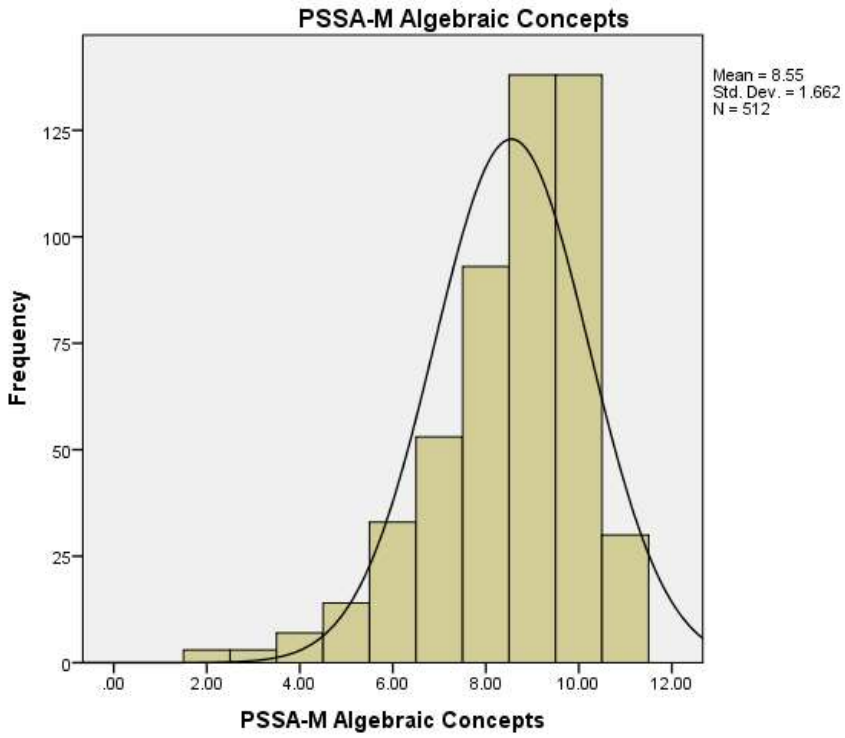


Figure 10. Histogram of third grade 2013 and 2014 Pennsylvania System of School Assessment, Mathematics Numbers and Operations data from students in first grade during the 2010 – 2011 and 2011 – 2012 school years.

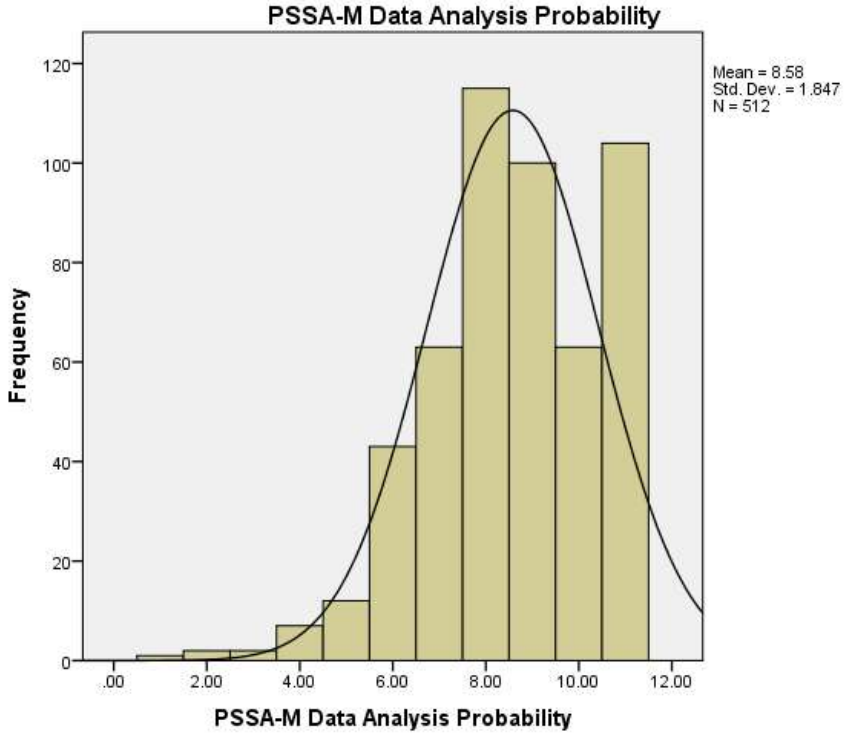


Figure 11. Histogram of third grade 2013 and 2014 Pennsylvania System of School Assessment, Mathematics Data Analysis and Probability data from students in first grade during the 2010 – 2011 and 2011 – 2012 school years.

Based on visual examination of histograms, first grade MBSP-C winter data, first grade MBSP-C spring data, PSSA-M composite data, Numbers and Operations subtest data, and Algebraic Concepts subtest data appear to be normally distributed. Visual inspections of histograms indicated negatively skewed distributions on the Measurement subtest, Geometry subtest, and Data Analysis and Probability subtest of the third grade PSSA-M. A negatively skewed distribution is expected on a criterion referenced assessment such as the PSSA and does not significantly violate the assumption of normality to conduct MLR. A negatively skewed distribution on a criterion referenced assessment indicates students have acquired the skills that are being taught (Brown, 1997; Osborne, 2013). First grade MBSP-C fall data had a positively skewed distribution. This may be due to a floor effect, given the limited computation instruction provided to first grade students in the beginning of the school year. Skewness and kurtosis values were generated to further explore the normality of the distributions. The acceptable or normal range for skewness values has been defined as -1.00 to 1.00 (Huck, 2014). Although the distributions appeared atypical, skewness coefficients fell within the acceptable range for all variables with the exception of first grade MBSP-C fall data (skewness = 1.23), Measurement data (skewness = -1.21), Geometry data (skewness = -1.48), and Algebraic Concepts data (skewness = -1.08). Kurtosis values ranged from -.58 to 2.51, all of which fell within the accepted range of -3.00 to 3.00 (Byrne, 2010; Huck, 2014). It is determined that first grade data did not significantly violate the assumption of normality given PSSA is a criterion referenced assessment. Please refer to Table 10 for sample size, mean, standard deviation, range, skewness, and kurtosis statistics.

First grade data were examined using boxplots to identify significant outliers. Significant outliers were defined as data points which exceeded three times the interquartile range. Two

data sets were identified as outliers and excluded from further analysis. The boxplots of first grade data are available for review in Appendix D.

Table 10

Descriptive Statistics for First Grade MBSP-C and PSSA Data

Variable	*N*	*M*	*SD*	Range	Skewness	Kurtosis
MBSP-C Fall	500	3.67	3.39	0-22	1.23	2.05
MBSP-C Winter	503	14.55	5.90	1-30	.59	.13
MBSP-C Spring	510	17.58	6.52	2-30	.11	-.58
PSSA-M Composite	274	1393.21	157.64	931-1859	.06	.461
PSSA-M Numbers and Operations	512	29.88	6.00	9-40	-.85	.95
PSSA-M Measurement	512	8.14	2.01	0-10	-1.21	1.11
PSSA-M Geometry	274	8.70	1.46	2-10	-1.47	2.51
PSSA-M Algebraic Concepts	512	8.55	1.66	2-11	-1.08	1.40
PSSA-M Data Analysis and Probability	512	8.58	1.85	1-11	-.597	.41

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics.

Second grade descriptive statistics. Please refer to Figures 12 through 20 for histograms of MBSP-C data in the fall, winter, and spring of second grade, in addition to the corresponding third grade PSSA-M Composite data and PSSA-M subtest data for a visual inspection of normality.

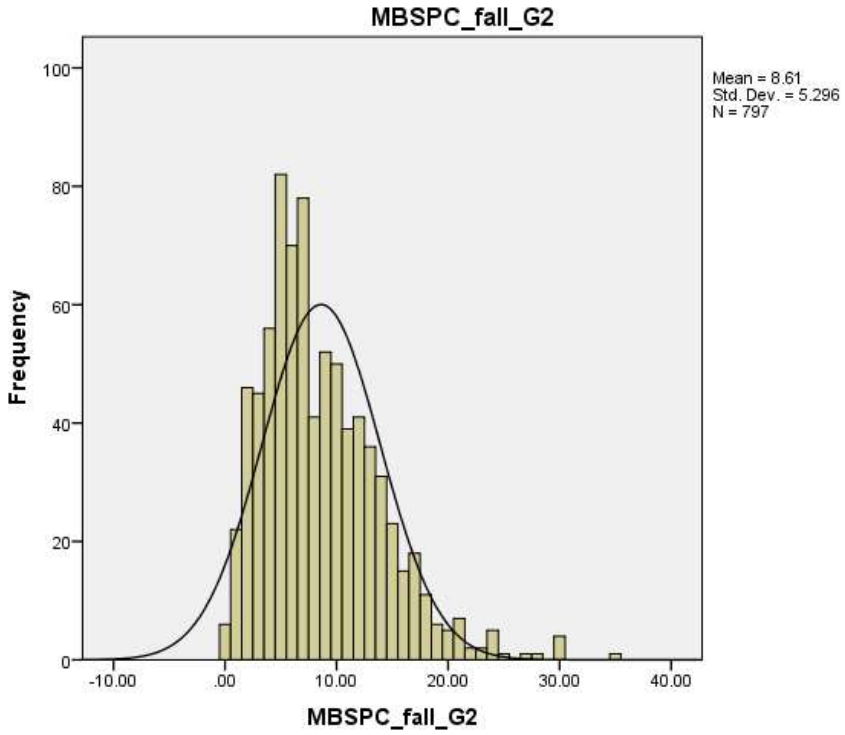


Figure 12. Histogram of second grade Monitoring Basic Skills Progress-Computation fall data.

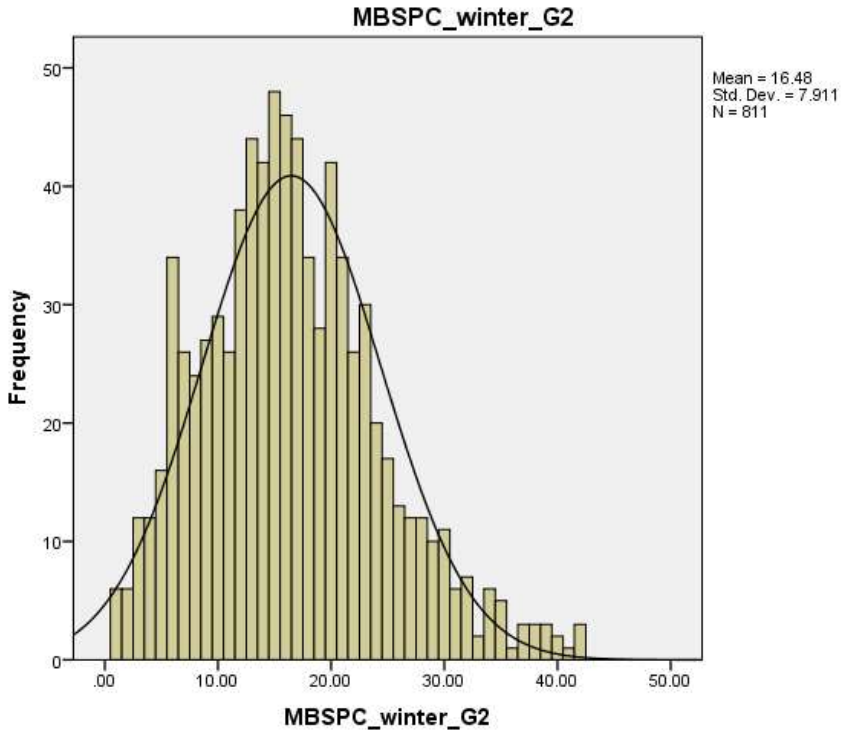


Figure 13. Histogram of second grade Monitoring Basic Skills Progress-Computation winter data.

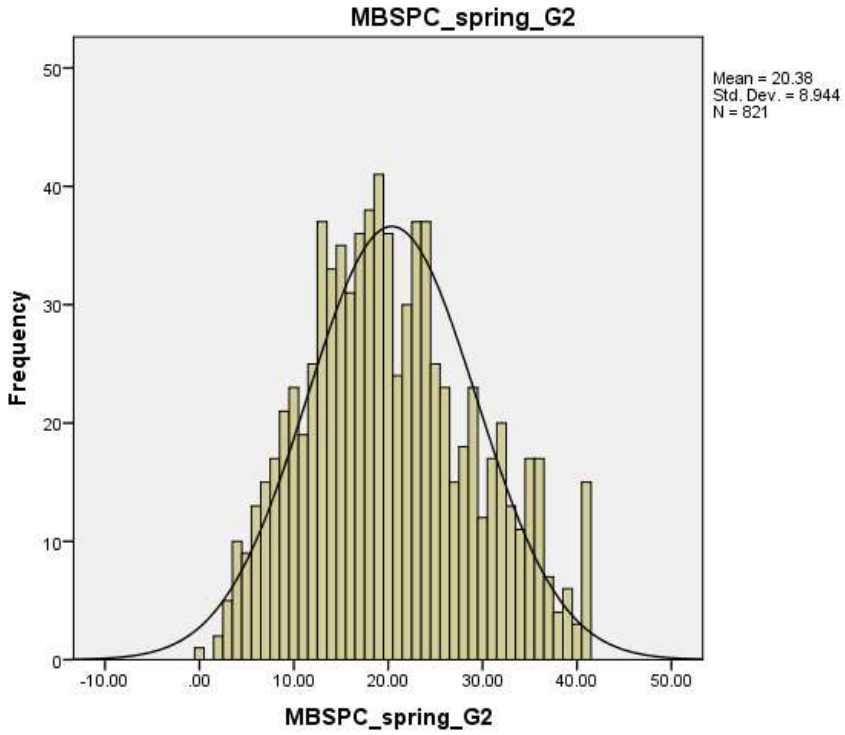


Figure 14. Histogram of second grade Monitoring Basic Skills Progress-Computation spring data.

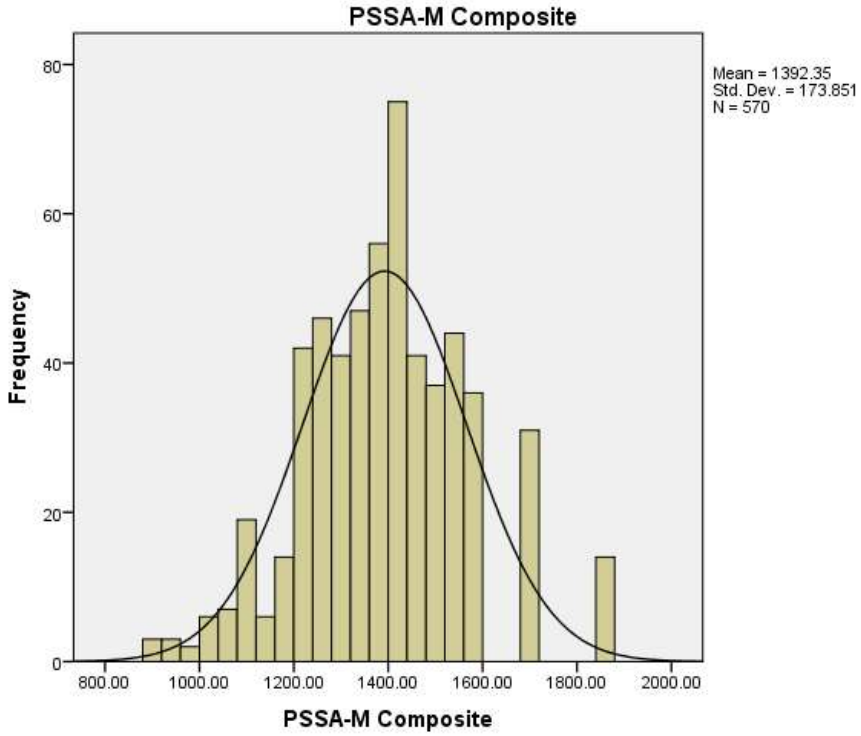


Figure 15. Histogram of third grade 2012 and 2013 Pennsylvania System of School Assessment, Mathematics composite data from students in second grade during the 2010 – 2011 and 2011 – 2012 school years.

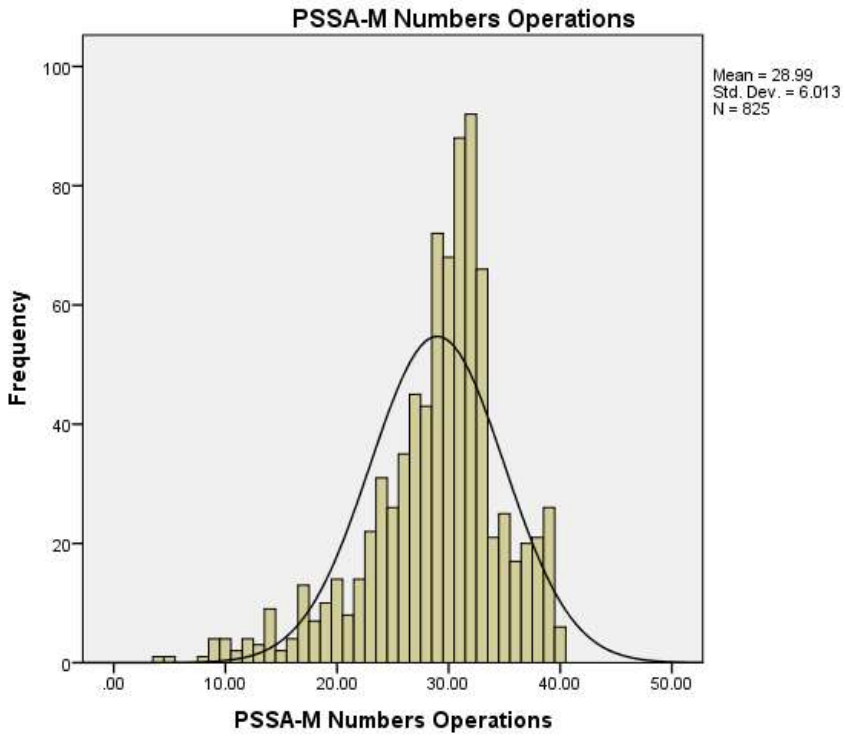


Figure 16. Histogram of third grade 2012, 2013, and 2014 Pennsylvania System of School Assessment, Mathematics Numbers and Operations data from students in second grade during the 2010 – 2011, 2011 – 2012, and 2012 – 2013 school years.

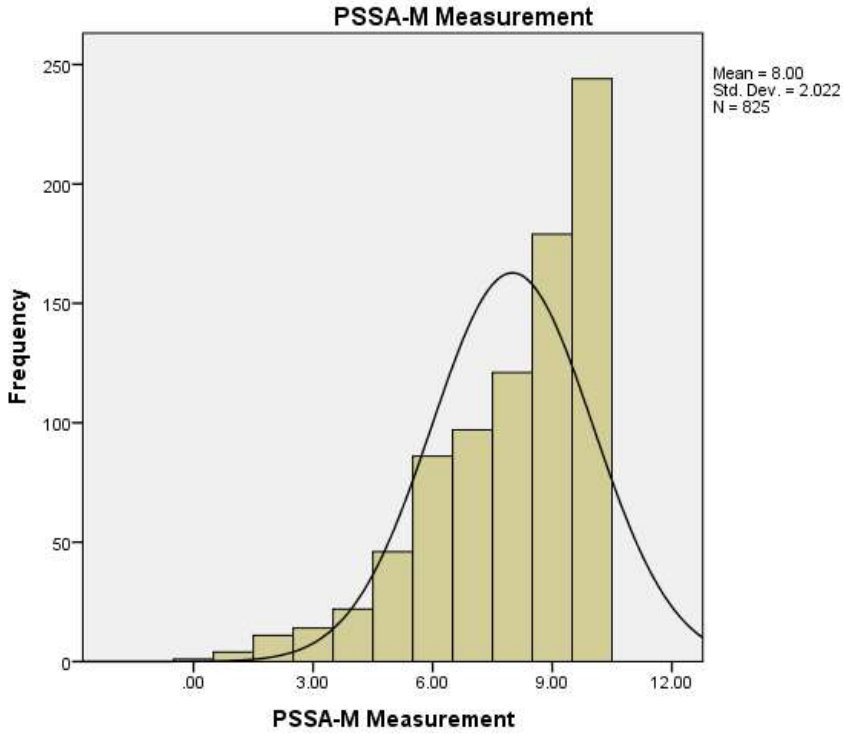


Figure 17. Histogram of third grade 2012, 2013, and 2014 Pennsylvania System of School Assessment, Mathematics Measurement data from students in second grade during the 2010 – 2011, 2011 – 2012, and 2012 – 2013 school years.

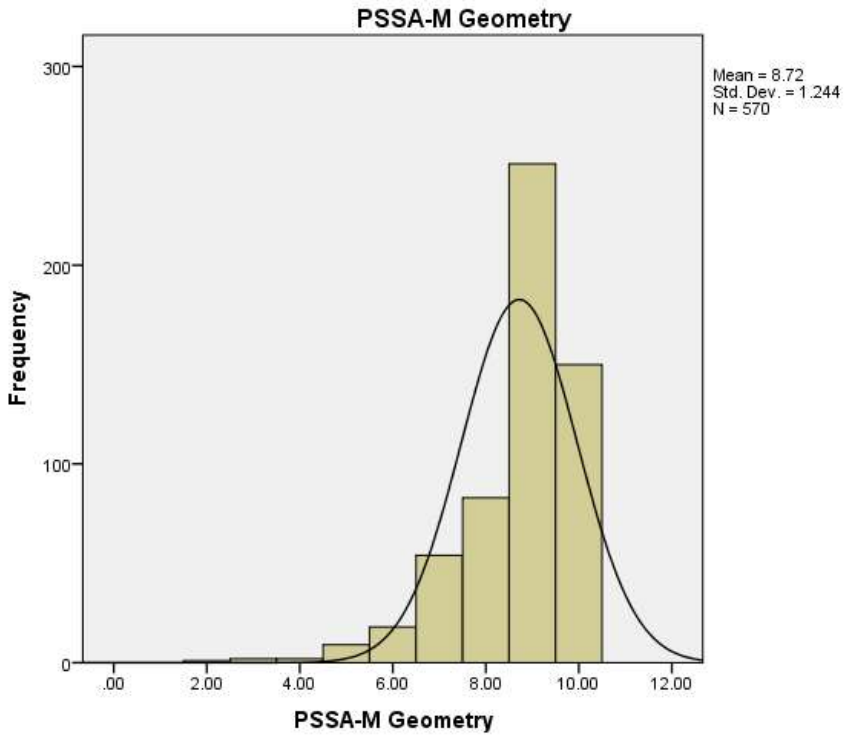


Figure 18. Histogram of third grade 2012 and 2013 Pennsylvania System of School Assessment, Mathematics Geometry data from students in second grade during the 2010 – 2011 and 2011 – 2012 school years.

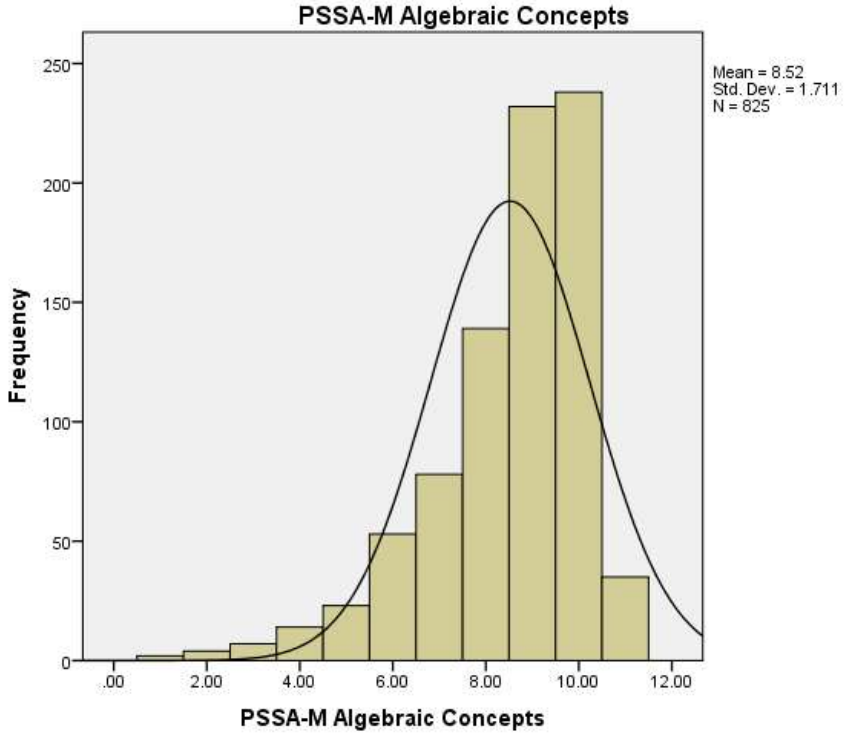


Figure 19. Histogram of third grade 2012, 2013, and 2014 Pennsylvania System of School Assessment, Mathematics Algebraic Concepts data from students in second grade during the 2010 – 2011, 2011 – 2012, and 2012 – 2013 school years.

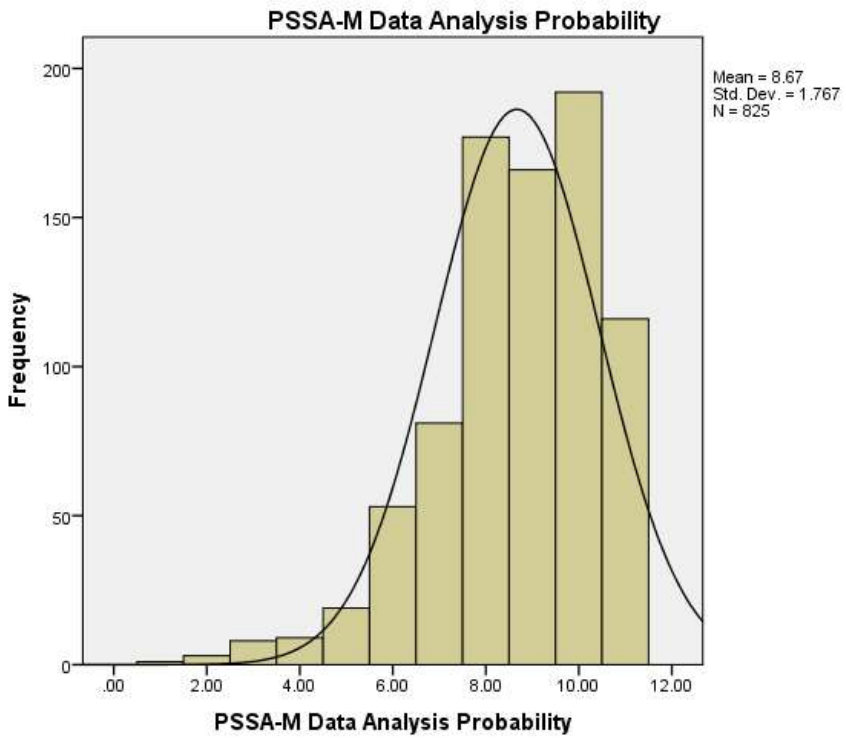


Figure 20. Histogram of third grade 2012, 2013, and 2014 Pennsylvania System of School Assessment, Mathematics Data Analysis and Probability data from students in second grade during the 2010 – 2011, 2011 – 2012, and 2012 – 2013 school years.

Data for second grade MBSP-C winter probe, second grade MBSP-C spring probe, PSSA-M Composite, and the Numbers and Operations subtest of the PSSA-M appear to be normally distributed based on visual inspection of histograms. The histogram for second grade MBSP-C fall data indicated a positively skewed distribution. The distribution of Measurement, Geometry, Algebraic Concepts, and Data Analysis and Probability data all appear negatively skewed based on a visual inspection of histograms. This negatively skewed distribution is expected on a criterion referenced assessment such as the PSSA and does not violate the assumptions of MLR. It indicates students acquired the skills that are being taught (Brown, 1997; Osborne, 2013). Skewness and kurtosis statistics were generated to further explore the normality of distribution. Skewness values fell within the acceptable range (-1.00 to 1.00) for second grade MBSP-C in the winter and spring, PSSA-M composite data, Numbers and

Operations subtest data, and Data Analysis and Probability subtest data. Negatively skewed distributions were confirmed by skewness values falling below -1.00 on the Measurement, Geometry, and Algebraic Concepts subtests of the third grade PSSA-M. A skewness statistic of 1.09 confirmed a positively skewed distribution of second grade MBSP-C fall data. Kurtosis values fell within normal limits for all distributions, with the exception Geometry with a kurtosis value of 3.49. It is determined second grade data did not significantly violate the assumption of normality given PSSA is a criterion referenced assessment. Please refer to Table 11 for sample size, mean scores, standard deviation, range, skewness, and kurtosis statistics.

Second grade data were examined using boxplots to identify significant outliers. Significant outliers were defined as data points which exceed three times the interquartile range. Five data sets were identified as outliers and excluded from further analysis. Boxplots of second grade data are provided for review in Appendix E.

Table 11

Descriptive Statistics for Second Grade MBSP-C and PSSA-M Data

Variable	*N*	*M*	*SD*	Range	Skewness	Kurtosis
MBSP-C Fall	797	8.61	5.30	0-35	1.09	1.79
MBSP-C Winter	811	16.48	7.91	1-42	.53	.21
MBSP-C Spring	821	20.38	8.94	0-41	.29	-.54
PSSA-M Composite	570	1392.35	173.85	894-1859	.08	.35
PSSA-M Numbers and Operations	825	28.99	6.01	4-40	-.99	1.50
PSSA-M Measurement	825	8.00	2.02	0-10	-1.08	.77
PSSA-M Geometry	570	8.72	1.24	2-10	-1.55	3.49
PSSA-M Algebraic Concepts	825	8.52	1.71	1-11	-1.28	1.93
PSSA-M Data Analysis and Probability	825	8.67	1.77	1-11	-.92	1.10

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics.

Third grade descriptive statistics. Please refer to Figures 21 through 29 for histograms of MBSP-C data in the fall, winter, and spring of third grade, in addition to the corresponding

third grade PSSA-M Composite data and PSSA-M subtest data for a visual inspection of normality.

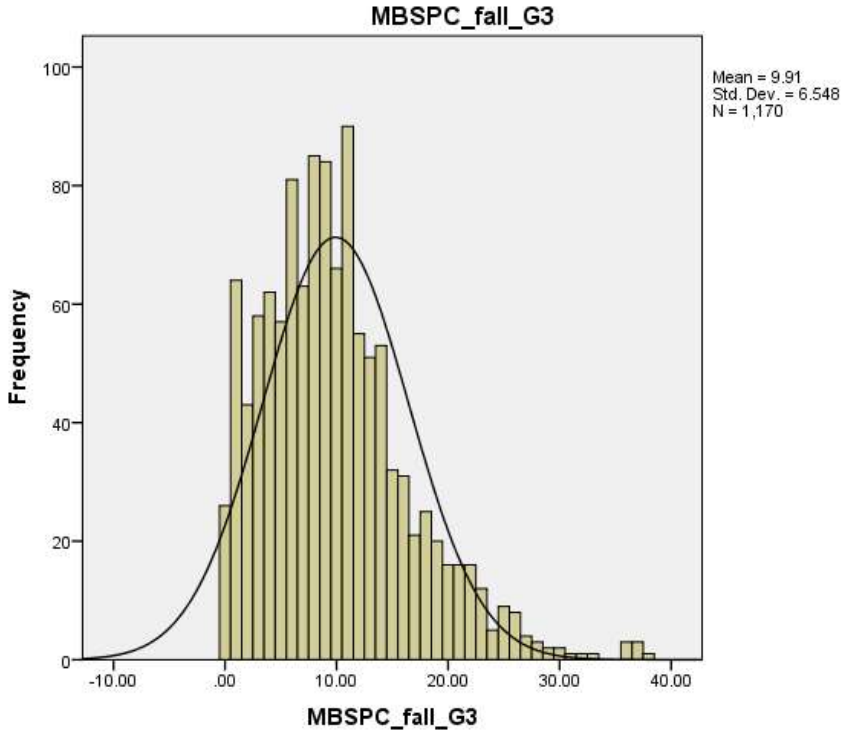


Figure 21. Histogram of third grade Monitoring Basic Skills Progress-Computation fall data.

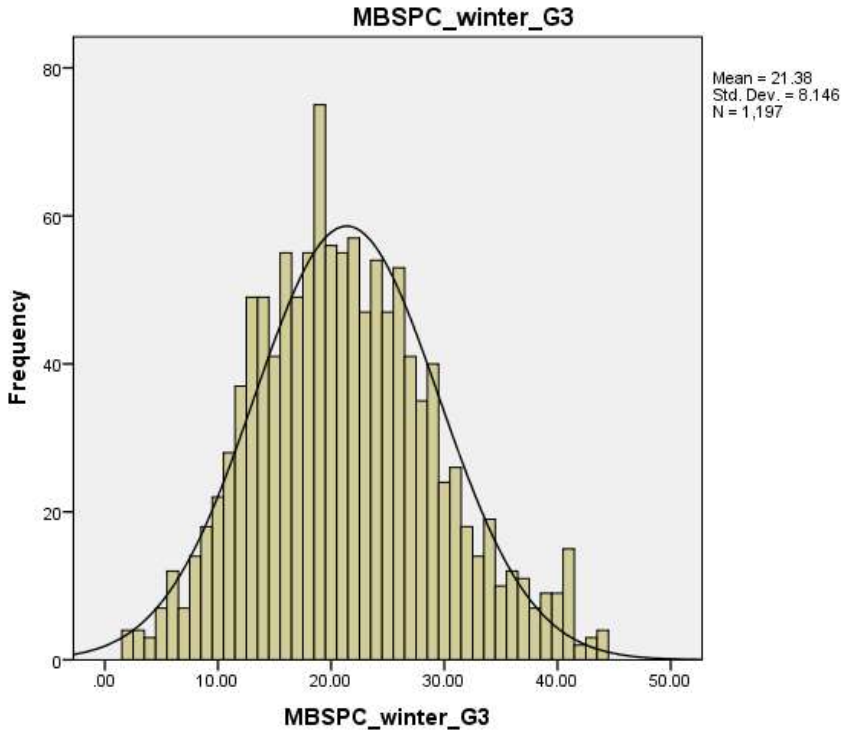


Figure 22. Histogram of third grade Monitoring Basic Skills Progress-Computation winter data.

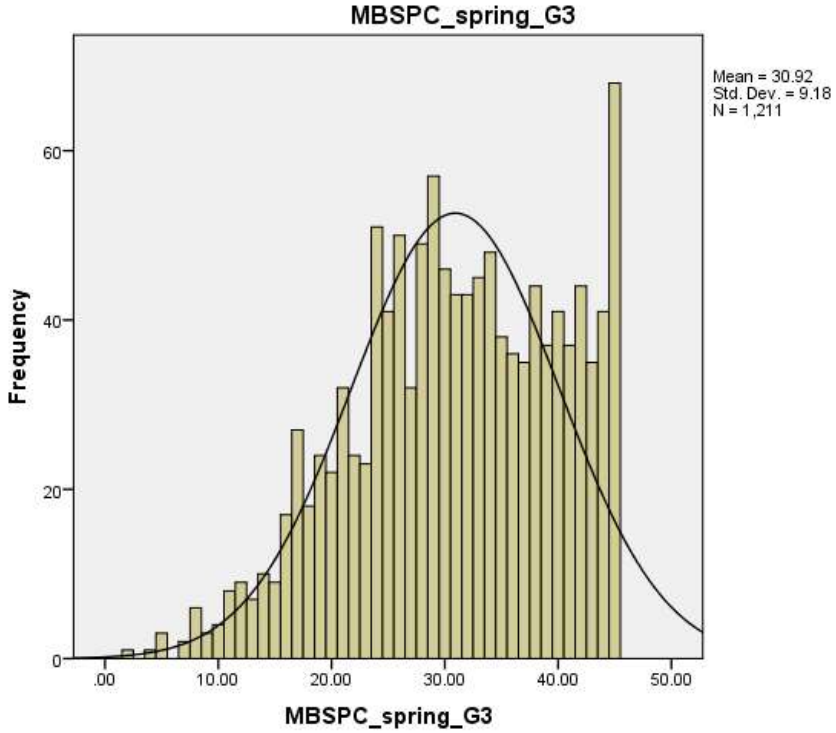


Figure 23. Histogram of third grade Monitoring Basic Skills Progress-Computation spring data.

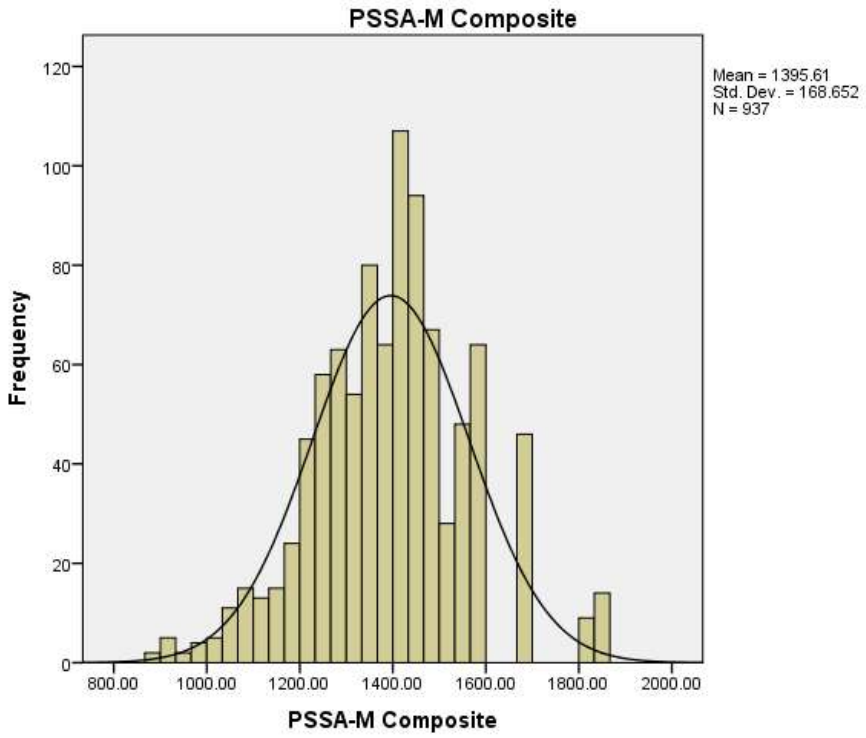


Figure 24. Histogram of third grade 2011, 2012, and 2013 Pennsylvania System of School Assessment, Mathematics composite data from students in third grade during the 2010 – 2011, 2011 – 2012, and 2012 – 2013 school years.

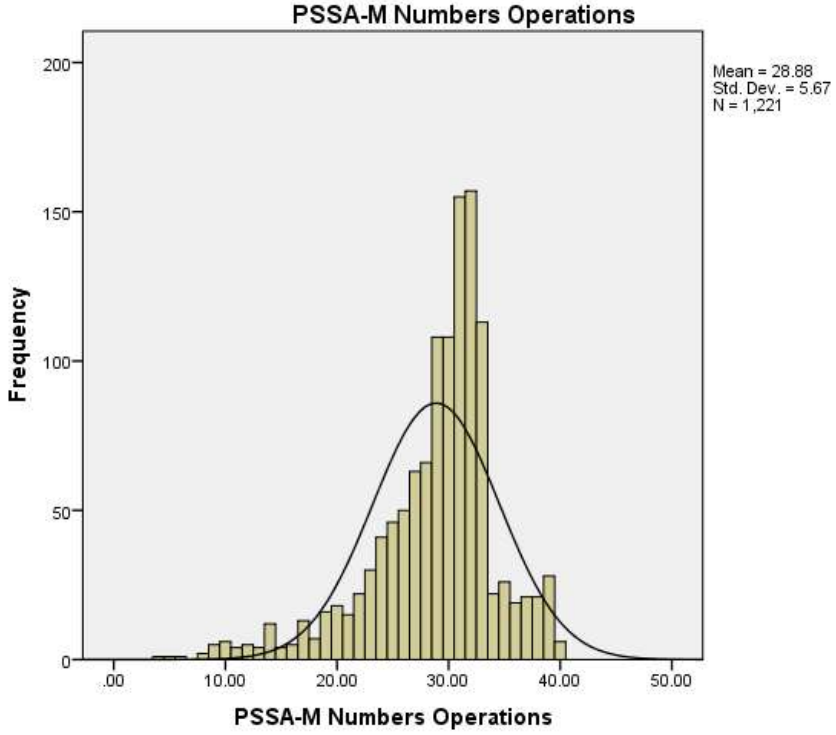


Figure 25. Histogram of third grade 2011, 2012, 2013, and 2014 Pennsylvania System of School Assessment, Mathematics Numbers and Operations data from students in third grade during the 2010 – 2011, 2011 – 2012, 2012 – 2013, and 2013 – 2014 school years.

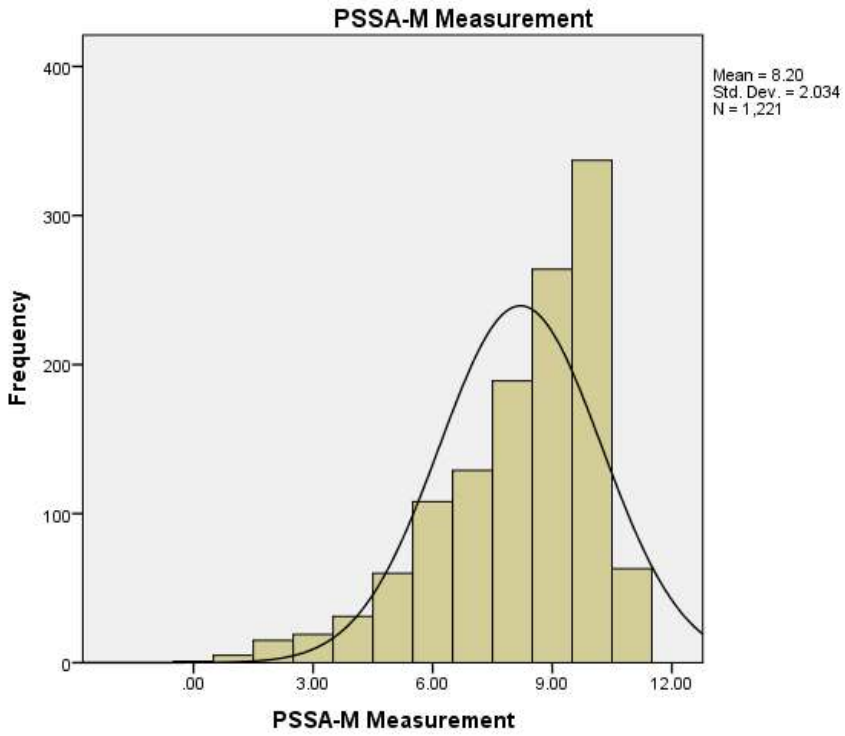


Figure 26. Histogram of third grade 2011, 2012, 2013, and 2014 Pennsylvania System of School Assessment, Mathematics Measurement data from students in third grade during the 2010 – 2011, 2011 – 2012, 2012 – 2013, and 2013 – 2014 school years.

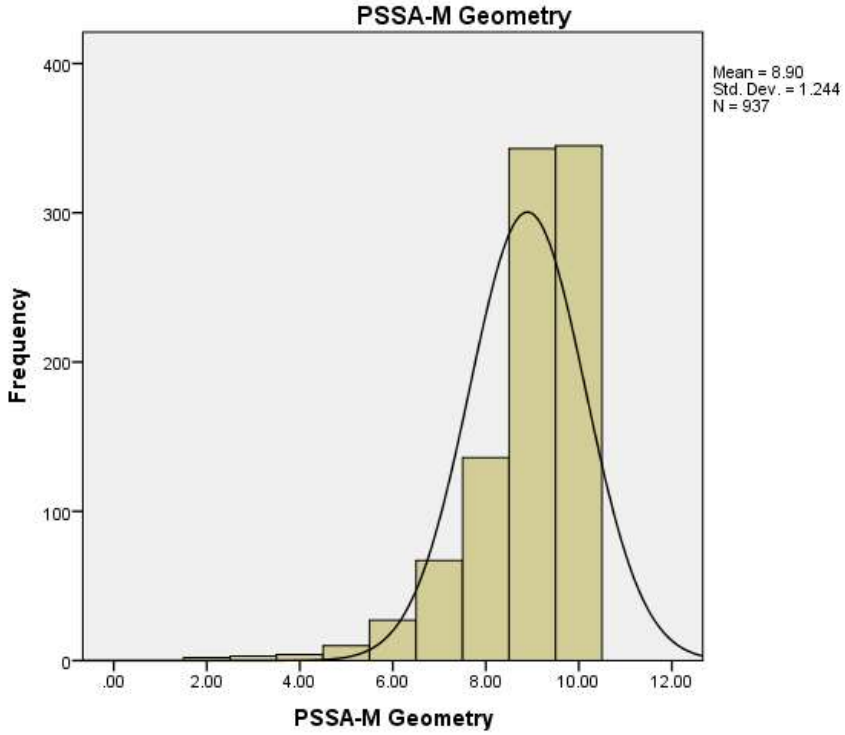


Figure 27. Histogram of third grade 2011, 2012, and 2013 Pennsylvania System of School Assessment, Mathematics Geometry data from students in third grade during the 2010 – 2011, 2011 – 2012, and 2012 – 2013 school years.

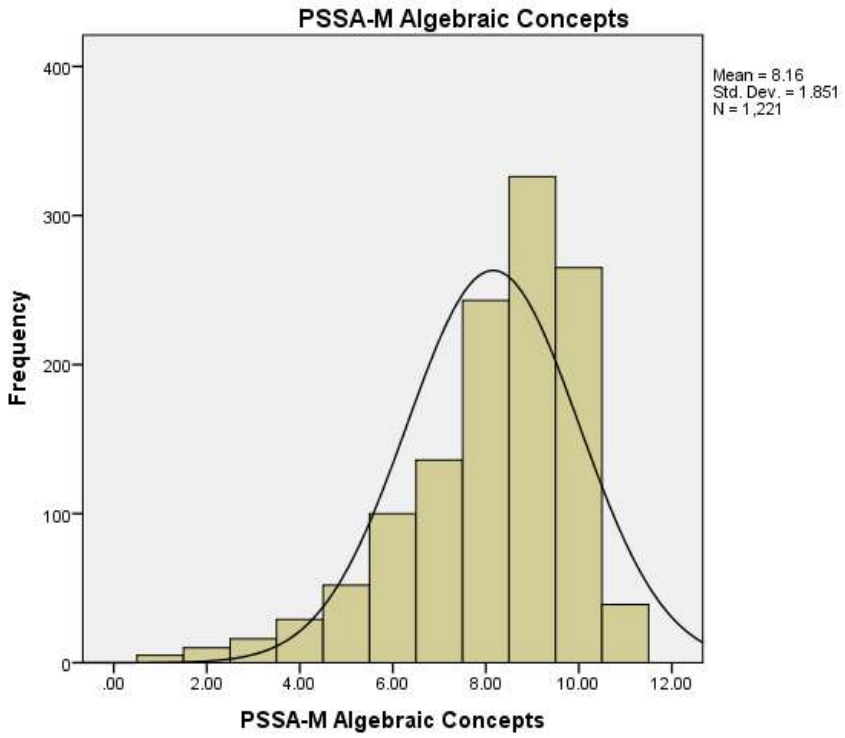


Figure 28. Histogram of third grade 2011, 2012, 2013, and 2014 Pennsylvania System of School Assessment, Mathematics Algebraic Concepts data from students who were in third grade during the 2010 – 2011, 2011 – 2012, 2012 – 2013, and 2013 – 2014 school years.

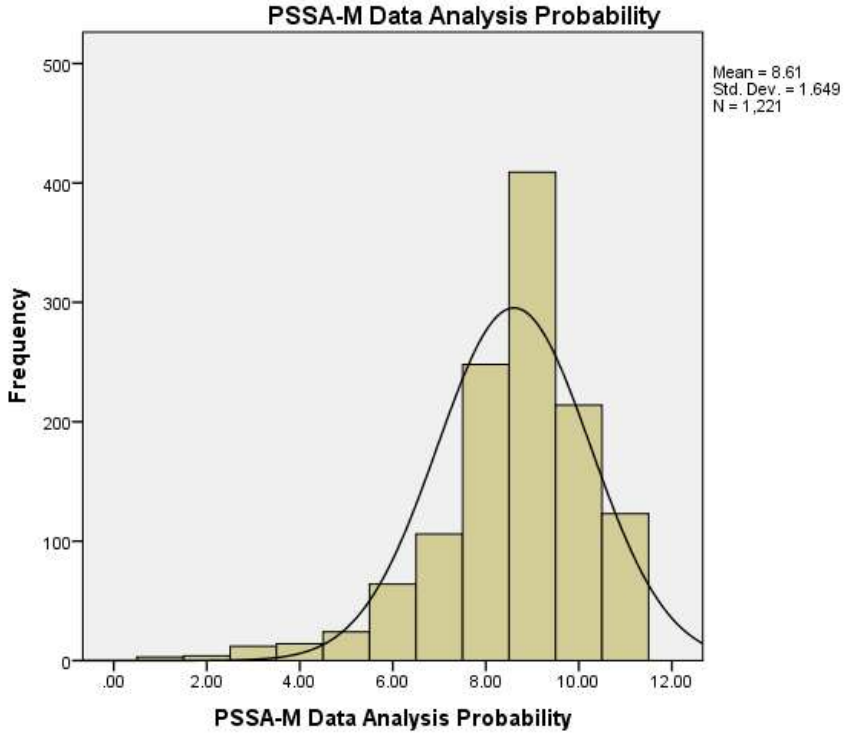


Figure 29. Histogram of third grade 2011, 2012, 2013, and 2014 Pennsylvania System of School Assessment, Mathematics Data Analysis and Probability data from students in third grade during the 2010 – 2011, 2011 – 2012, 2012 – 2013, and 2013 – 2014 school years.

A visual examination of third grade MBSP-C and PSSA-M data histograms indicated normal distribution of third grade MBSP-C in the fall, winter, spring, and PSSA-M composite data. Skewness and kurtosis values fell within the acceptable range, confirming normal distribution of MBSP-C and PSSA-M data. Histograms for the PSSA-M subtests indicated a negatively skewed distribution. This was confirmed by skewness statistics that fell outside of the acceptable range of -1.00 to 1.00. Kurtosis values for all variables fell within the acceptable range of -3.00 to 3.00 with the exception of the Geometry subtest, which was leptokurtic (kurtosis value of 4.23). Third grade data met the assumption of normality. Please refer to Table 12 for sample size, mean scores, standard deviation, range, skewness, and kurtosis statistics.

Third grade data were examined using boxplots to identify significant outliers. Significant outliers were defined as data points which fell three times above the interquartile range. Ten data sets from third grade data sets were identified as outliers and excluded from further analysis. Boxplots of third grade data are provided for review in Appendix F.

Table 12

Descriptive Statistics for Third Grade MBSP-C and PSSA-M Data

Variable	*N*	*M*	*SD*	Range	Skewness	Kurtosis
MBSP-C Fall	1170	9.91	6.55	0-38	.98	1.38
MBSP-C Winter	1197	21.38	8.14	2-44	.33	-.13
MBSP-C Spring	1211	30.92	9.18	2-45	-.34	-.54
PSSA-M	937	1395.6	168.65	872-1859	.03	.50
Numbers and Operations	1221	28.88	5.67	4-40	-1.16	2.11
Measurement	1221	8.20	2.03	0-11	-1.06	.86
Geometry	937	8.90	1.24	2-10	-1.71	4.23
Algebraic Concepts	1221	8.16	1.85	1-11	-1.10	1.25
Data Analysis and Probability	1221	8.60	1.65	1-11	-1.17	2.38

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics.

It is important to note, negatively skewed data on a criterion-referenced assessment, such as the PSSA, are typical and do not violate the assumptions of MLR (Brown, 1997; Osborne,

2013). Negatively skewed distributions, on criterion referenced assessments, indicate a majority of students were able to acquire the skills being assessed and is desired over normal or positively skewed distributions. The positively skewed distribution of MBSP-C fall data in first and second grades may reflect global deficits of skills students are expected to learn throughout that grade level and can be explained by a lack of an opportunity to learn or a floor effect. This may be magnified in the first grade MBSP-C fall data due to the novelty of a fluency-based mathematics assessments for first grade students at the beginning of the school year.

Independence of Observations

Durbin-Watson statistics were generated in order to check for independence of observations. Durbin-Watson values range from 0 to 4 and a value of 2 indicates no autocorrelation. Values of 1 to 3 are considered acceptable and imply independence of observation (Field, 2013). Durbin-Watson values for each MLR analysis at each grade level fell within the acceptable range. Therefore, it can be concluded that residuals are independent of each other and the assumption for MLR is met. Please refer to Table 13 for a summary of Durbin-Watson values.

Table 13

Durbin-Watson Values From Multiple Linear Regression of Dependent Variables

	First Grade	Second Grade	Third Grade
PSSA-M Composite	2.26	2.09	2.02
Numbers and Operations	1.89	1.78	1.86
Measurement	2.06	1.97	1.82
Geometry	2.18	1.79	1.78
Algebraic Concepts	2.18	2.03	1.85
Data Analysis and Probability	1.34	1.33	1.43

Note. PSSA-M = Pennsylvania System of School Assessment, Mathematics.

Linear Relationships

In order to properly analyze data with MLR, there must be a linear relationship between the dependent variable (e.g., PSSA) and each independent variable and the dependent variable

and the independent variables collectively. Scatterplots were generated in order to visually inspect whether or not a linear relationship existed between the dependent and independent variables. A visual examination of the scatterplots indicate positive linear relationships between the dependent and independent variables. Therefore, the assumption of linearity is met for the purpose of MLR.

Homoscedasticity

Homoscedasticity indicates an equal variance of the dependent variable for each level of an independent variable. Homoscedasticity is determined by a visual inspection of scatterplots of the standardized residuals and standardized predicted values. A lack of patterns on the scatterplot indicates homoscedasticity (Aldrich & Cunningham, 2016).

First grade homoscedasticity figures. A visual inspection of first grade standardized residuals and standardized predicted values scatterplots indicates the assumption of homoscedasticity was met for the PSSA-M Composite variable and PSSA-M Numbers and Operations dependent variable. Please refer to Figures 30 and 31. Scatterplots of PSSA-M Measurement, PSSA-M Geometry, PSSA-M Algebraic Concepts, and PSSA-M Data Analysis and Probability demonstrate some patterning which indicates heteroscedasticity. Please refer to Figures 32 through 35. Therefore, the assumption of homoscedasticity is not met for these dependent variables. Lack of homoscedasticity is not a gross violation of multi-linear regression, but does weaken the regression model (Field, 2013; Statistics Solutions, 2013).

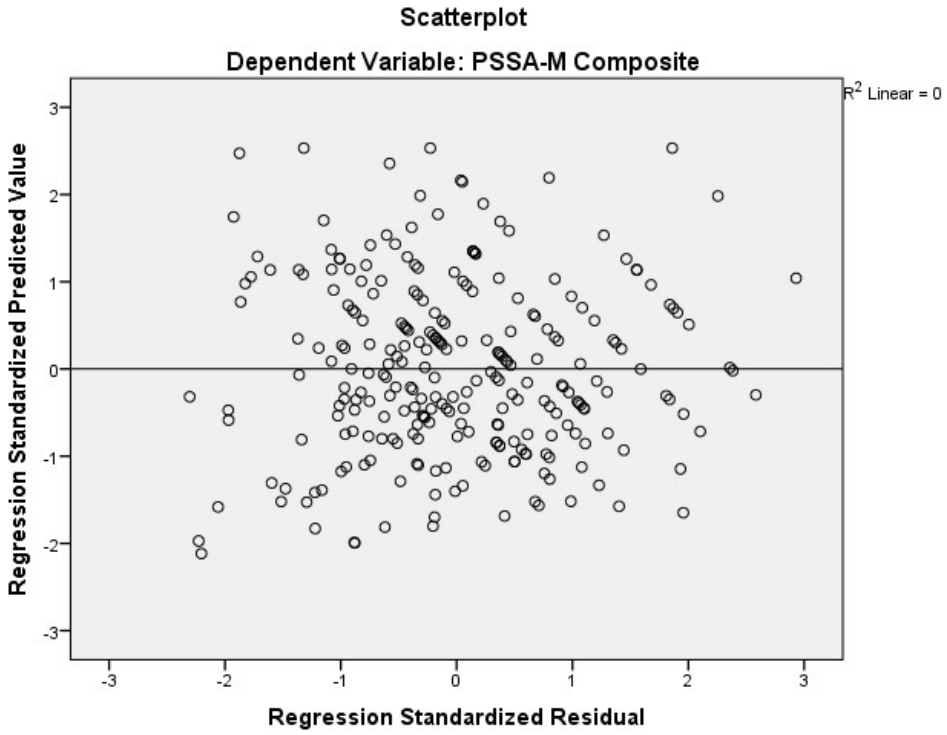


Figure 30. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Composite for first grade cohort.

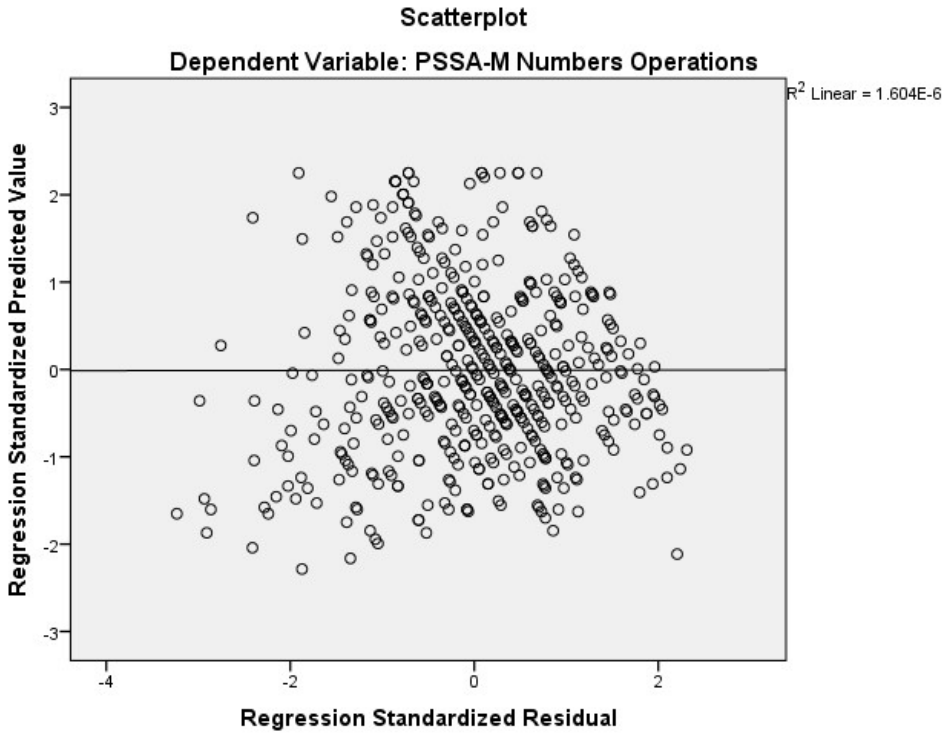


Figure 31. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Numbers and Operations subtest for first grade cohort.

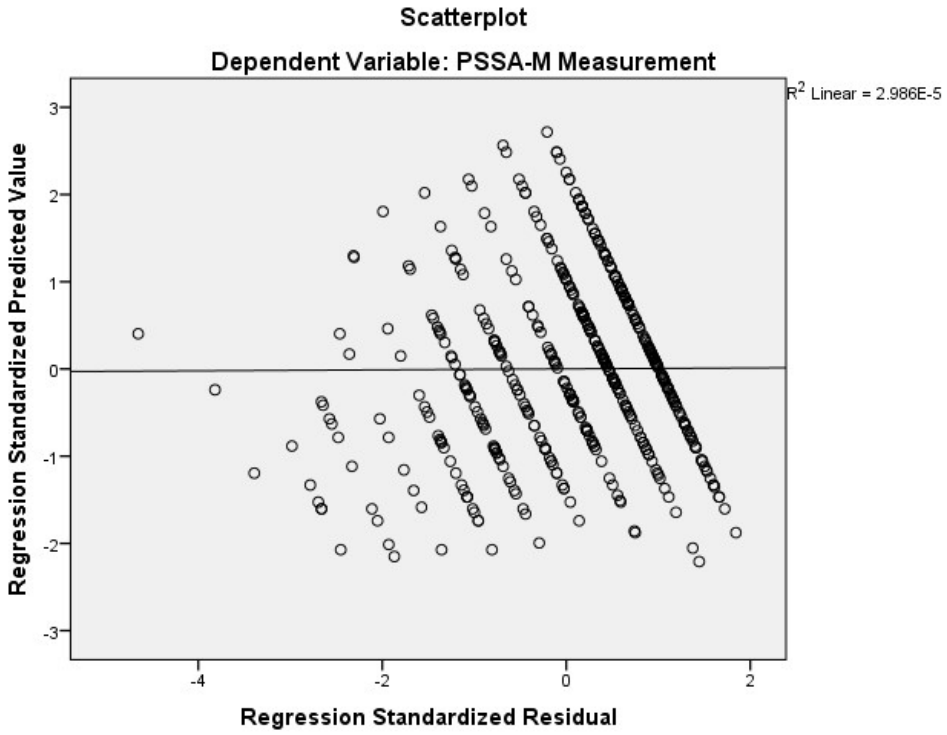


Figure 32. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Measurement subtest for first grade cohort.

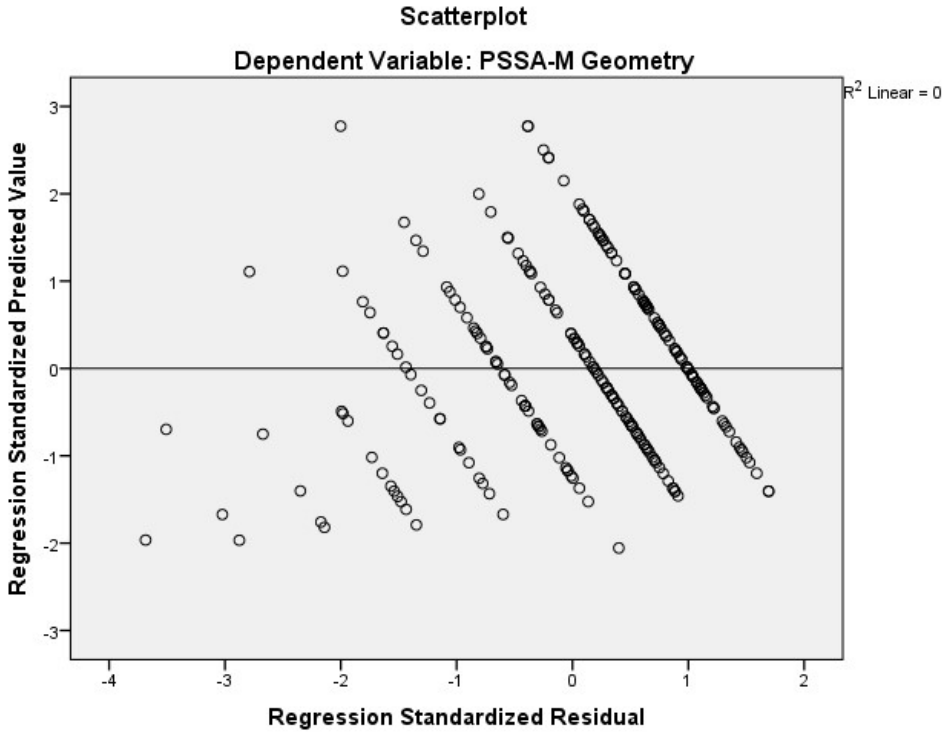


Figure 33. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Geometry subtest for first grade cohort.

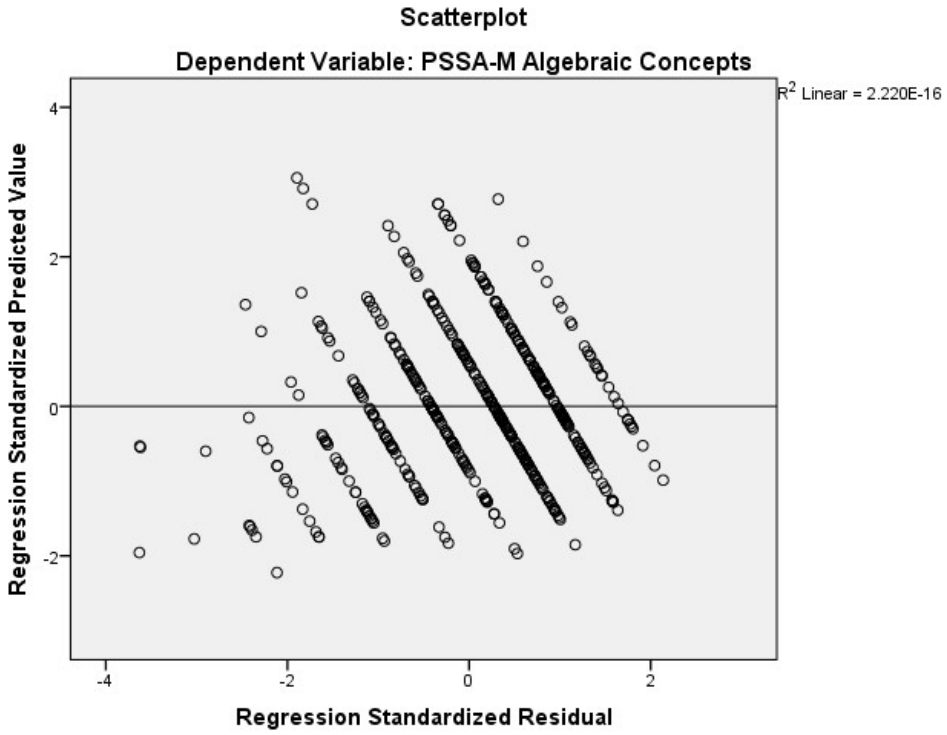


Figure 34. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Algebraic Concepts subtest for first grade cohort.

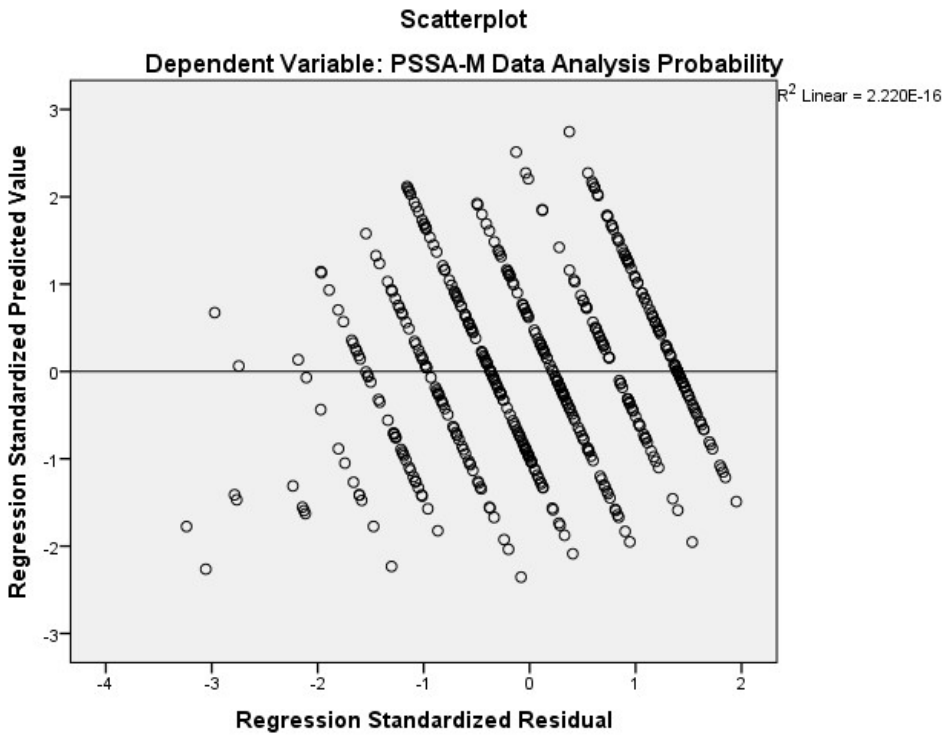


Figure 35. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Data Analysis and Probability subtest for first grade cohort.

Second grade homoscedasticity. The homoscedasticity of the second grade dependent variables was investigated by a visual examination of a scatterplot of the standardized regression residuals and standardized predicted values. Please refer to Figures 36 through 41. A visual inspection of these scatterplots indicates the assumption of homoscedasticity was met for PSSA-M Composite and PSSA-M Numbers and Operations. Homoscedasticity was not observed for the Measurement, Geometry, Algebraic Concepts, and Data Analysis and Probability PSSA-M subtests. As previously noted, a violation of the homoscedasticity assumption weakens the regression model but is not a significant violation (Statistics Solutions, 2013).

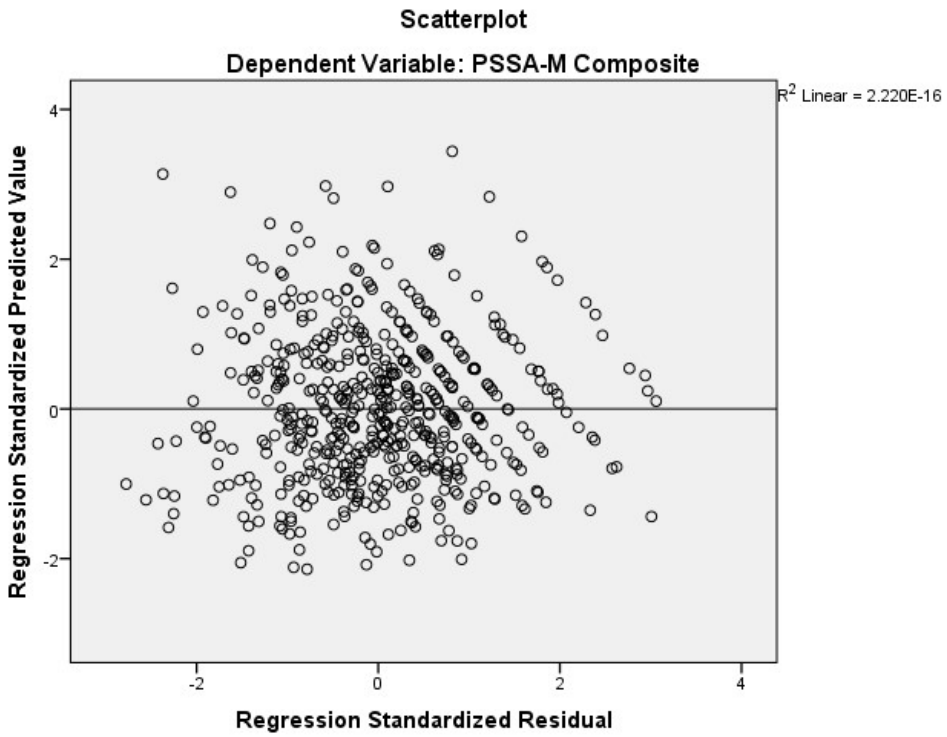


Figure 36. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Composite for the second grade cohort.

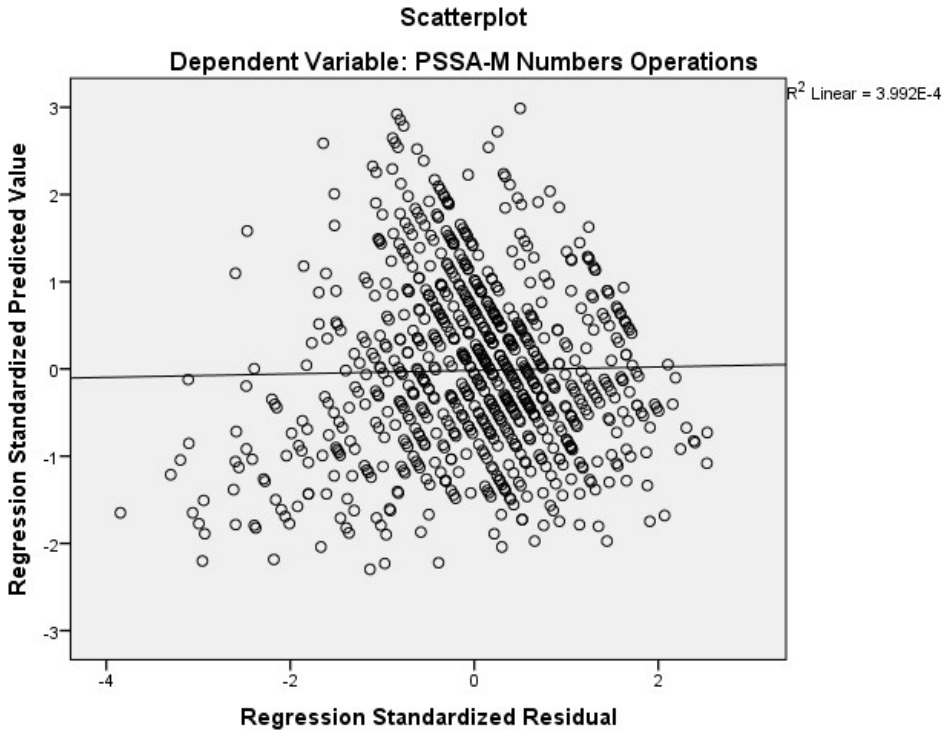


Figure 37. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Numbers and Operations subtest for the second grade cohort.

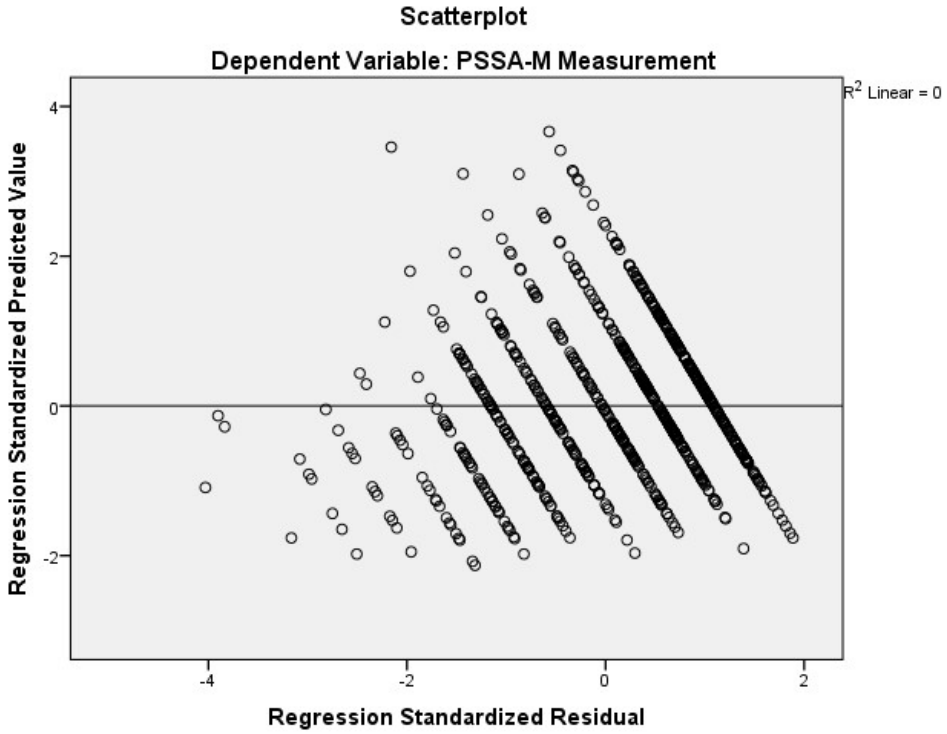


Figure 38. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Measurement subtest for the second grade cohort.

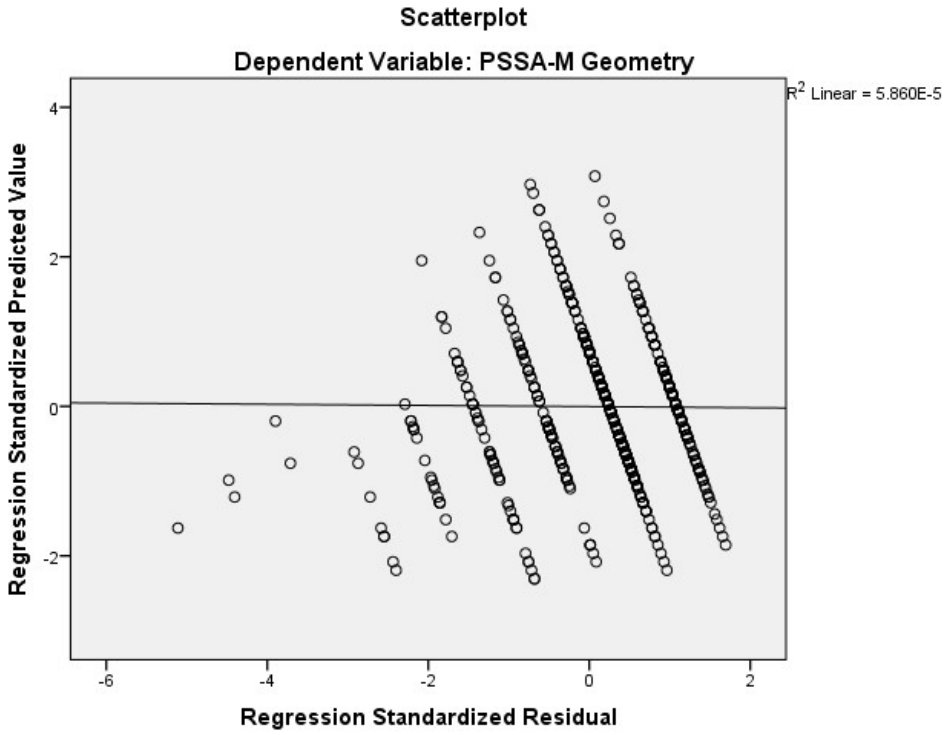


Figure 39. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Geometry subtest for the second grade cohort.

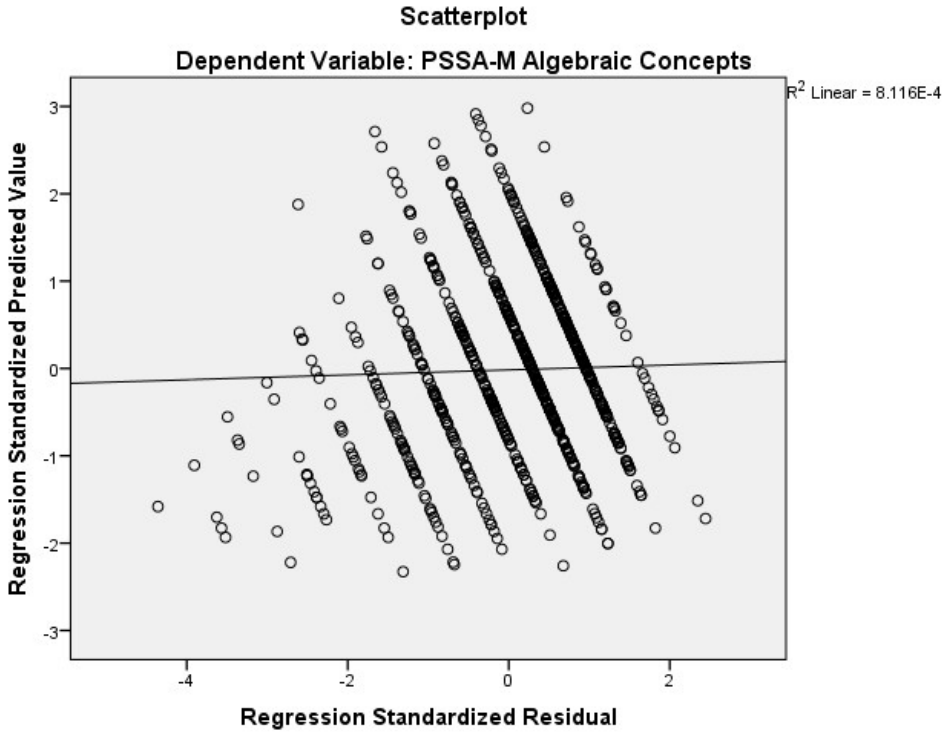


Figure 40. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Algebraic Concepts subtest for the second grade cohort.

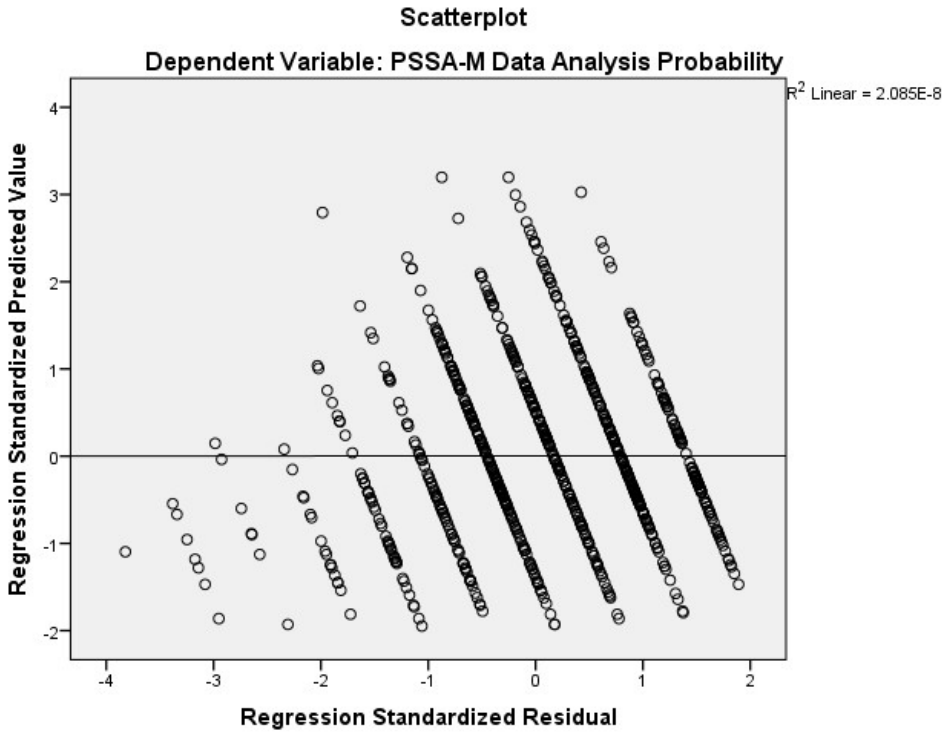


Figure 41. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Data Analysis and Probability subtest for the second grade cohort.

Third grade homoscedasticity. The homoscedasticity of the third grade dependent variables was investigated by a visual examination of a scatterplot of the standardized regression residuals and standardized predicted values. Similar to first and second grade data, the assumption of homoscedasticity was met for the PSSA-M Composite and PSSA-M Numbers and Operations dependent variables, but violated for the remaining PSSA-M subtests. Please refer to Figures 42 through 47. A lack of homoscedasticity weakens the regression model but is not a significant violation (Statistics Solutions, 2013).

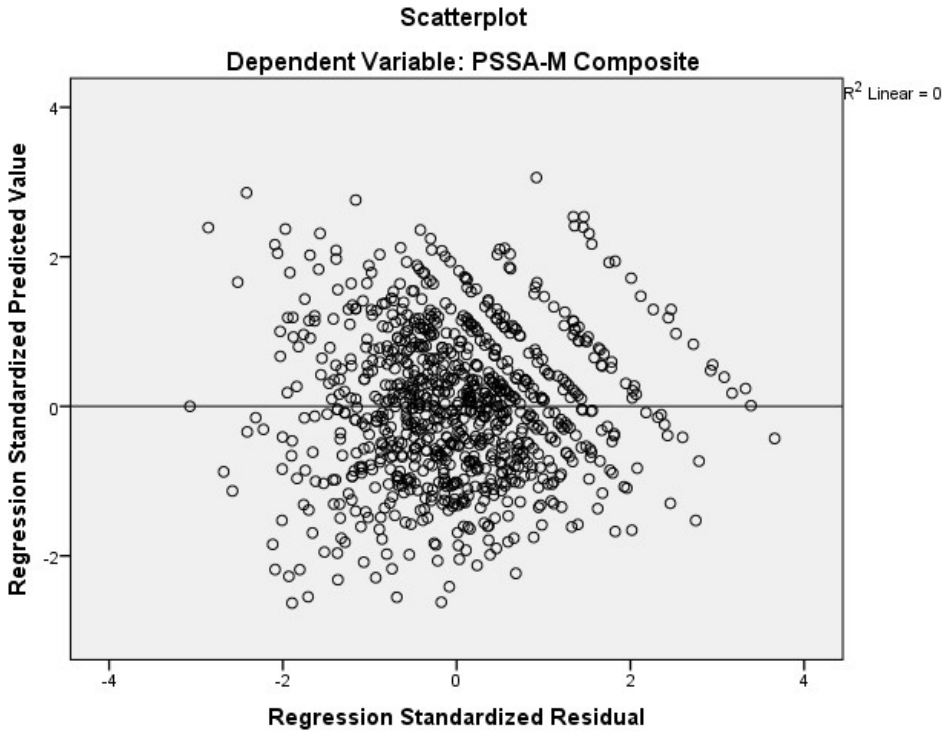


Figure 42. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Composite for the third grade cohort.

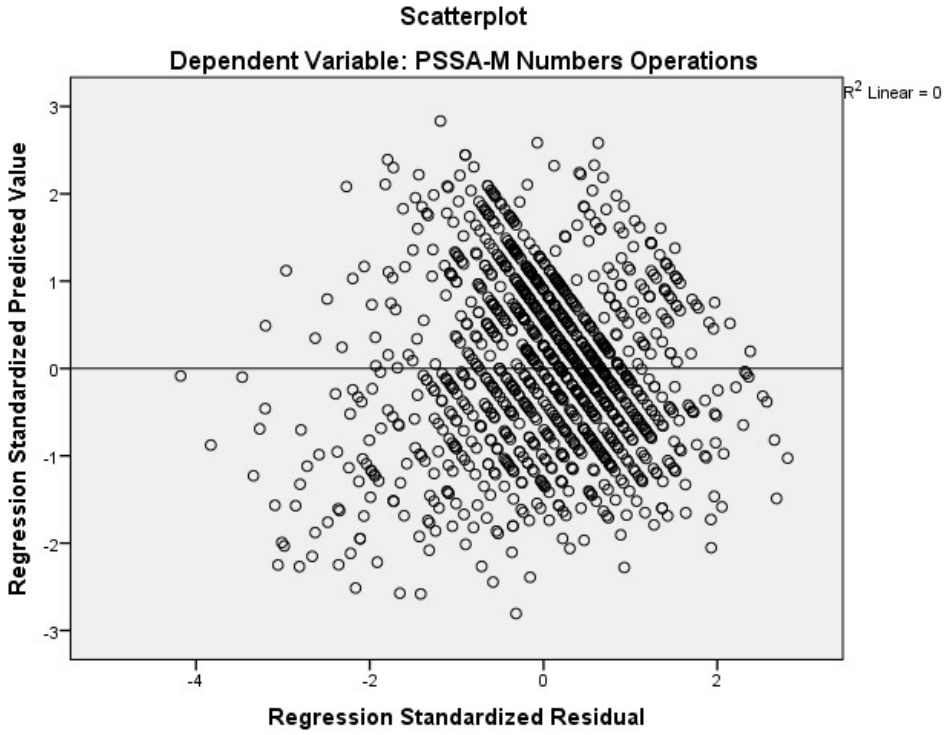


Figure 43. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Numbers and Operations subtest for the third grade cohort.

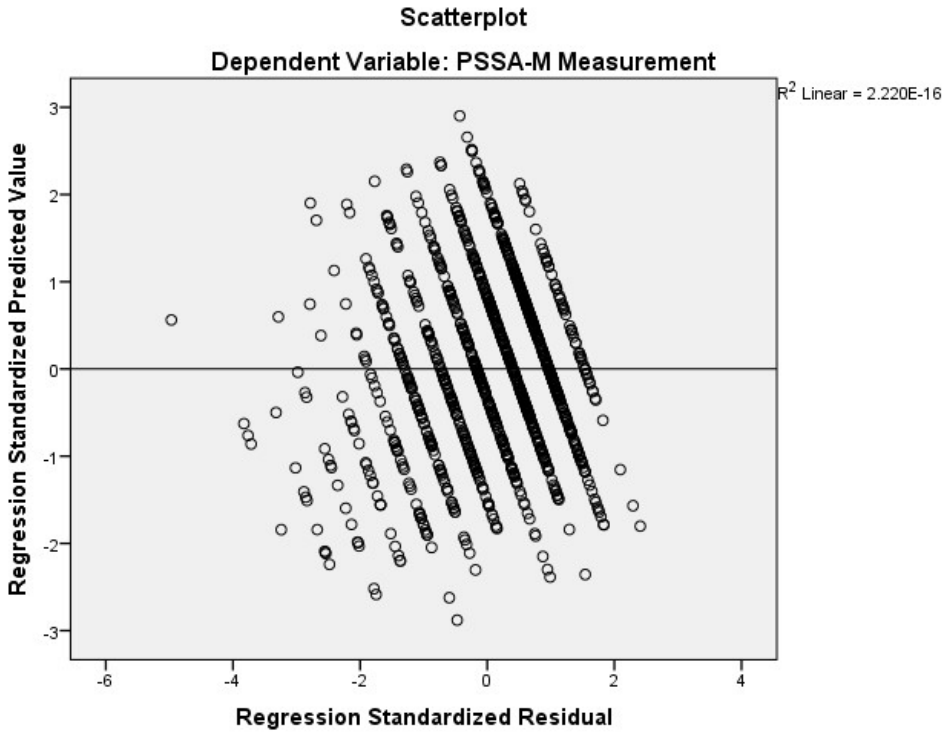


Figure 44. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Measurement subtest for the third grade cohort.

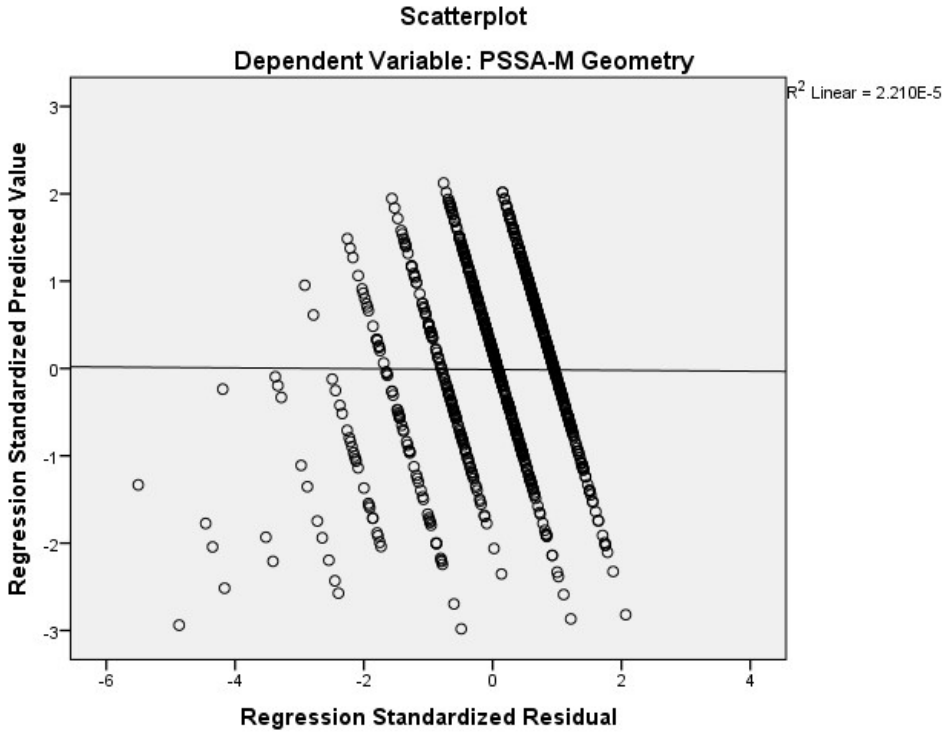


Figure 45. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Geometry subtest for the third grade cohort.

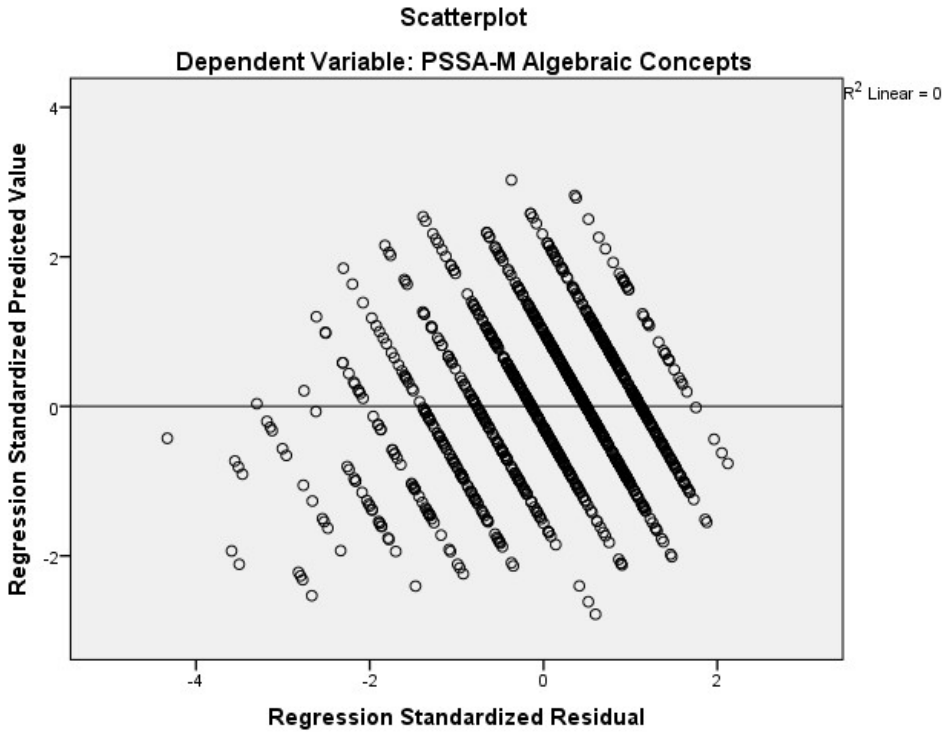


Figure 46. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Algebraic subtest for the third grade cohort.

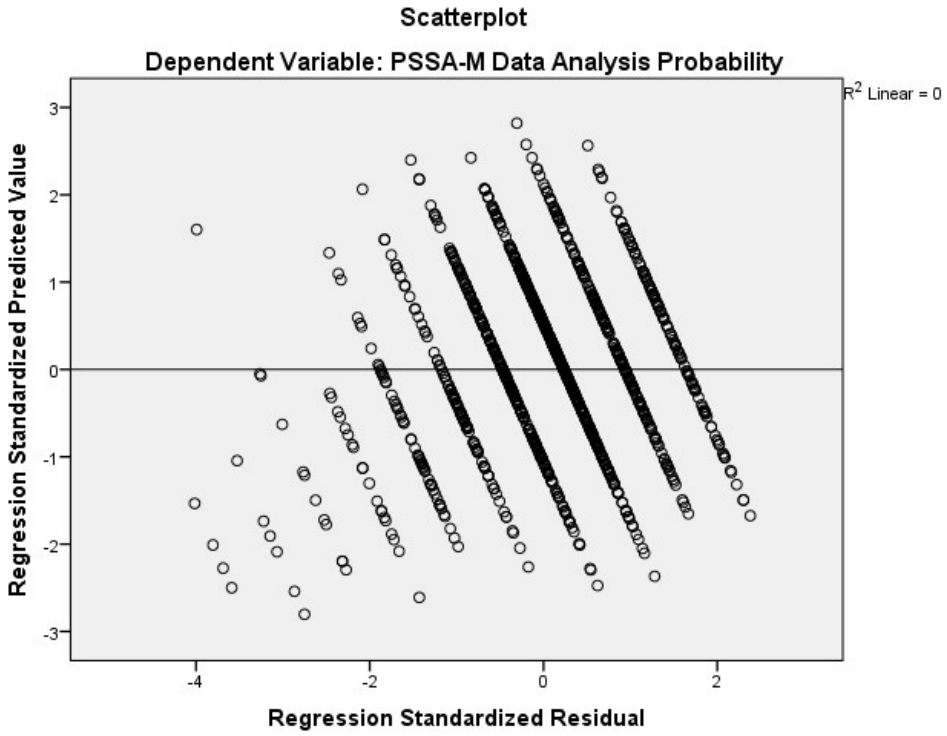


Figure 47. Scatterplot of regression standardized predicted value and regression standardized residual of Pennsylvania System of School Assessment, Mathematics Data Analysis and Probability subtest for the third grade cohort.

Multicollinearity

Multicollinearity occurs when two or more independent variables are highly correlated with each other. Multicollinearity is problematic when conducting MLR because it is difficult to determine which independent variable is contributing to the variance in the dependent variable. Therefore, data should not demonstrate multicollinearity when conducting a MLR. Multicollinearity is determined by visually inspecting histograms of standardized residuals and Normal P-P plots. Normally distributed residuals will fall along the diagonal line of the plot. Lines that do not trend along the diagonal indicate a deviation from normality (Fields, 2013). Multicollinearity is not present when the residuals are normally distributed.

First grade multicollinearity statistics. Histograms and Normal P-P plots of the standardized residuals were generated and inspected to determine whether or not multicollinearity was present with first grade data. Multicollinearity was not observed with PSSA-M Composite data and on all PSSA-M subtests, with the exception of Measurement and Geometry, as evidenced by normal distributions on both histograms and Normal P-P plots. A slight left skew was observed on the Normal P-P plots for the Geometry subtest and Measurement subtest. However, the left skewed was not significant enough to indicate a violation of the multicollinearity assumption. It is not necessary for data points to be perfectly aligned with the diagonal as MLR is robust to deviations from normality (Laerd Statistics, 2015). Please refer to Figures 48 through 59. Therefore, it is concluded the multicollinearity assumption is met.

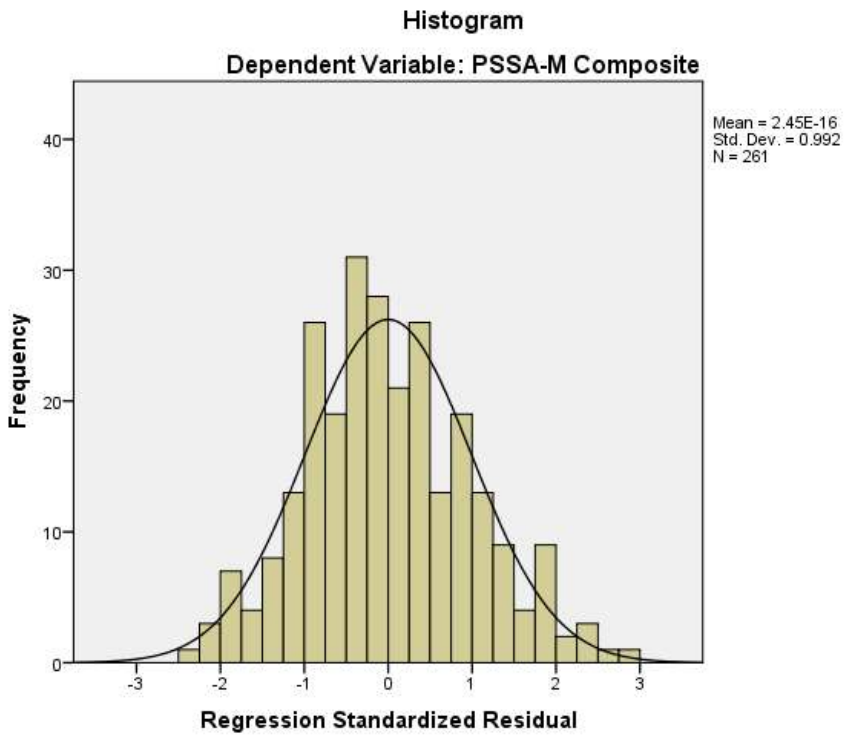


Figure 48. Histogram of Pennsylvania System of School Assessment, Mathematics Composite standardized residuals for the first grade cohort.

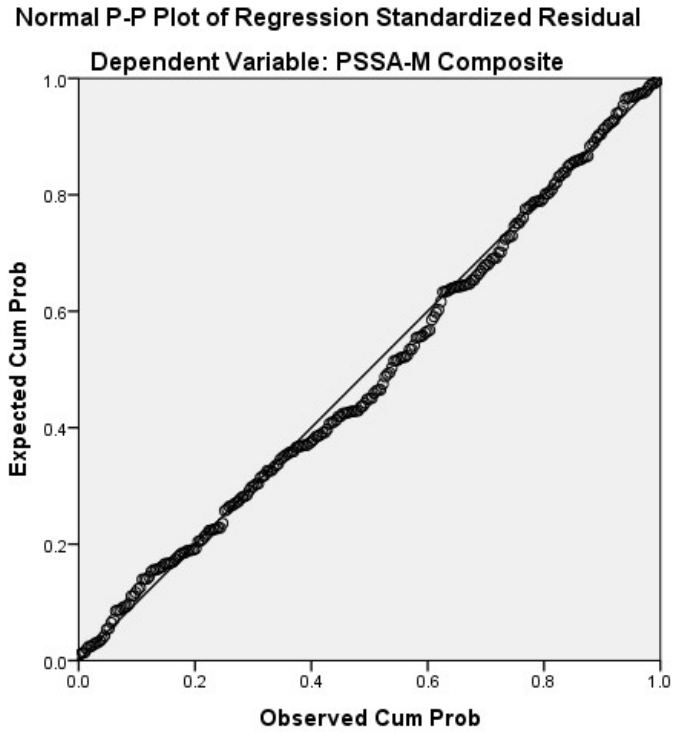


Figure 49. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Composite standardized residuals for the first grade cohort.

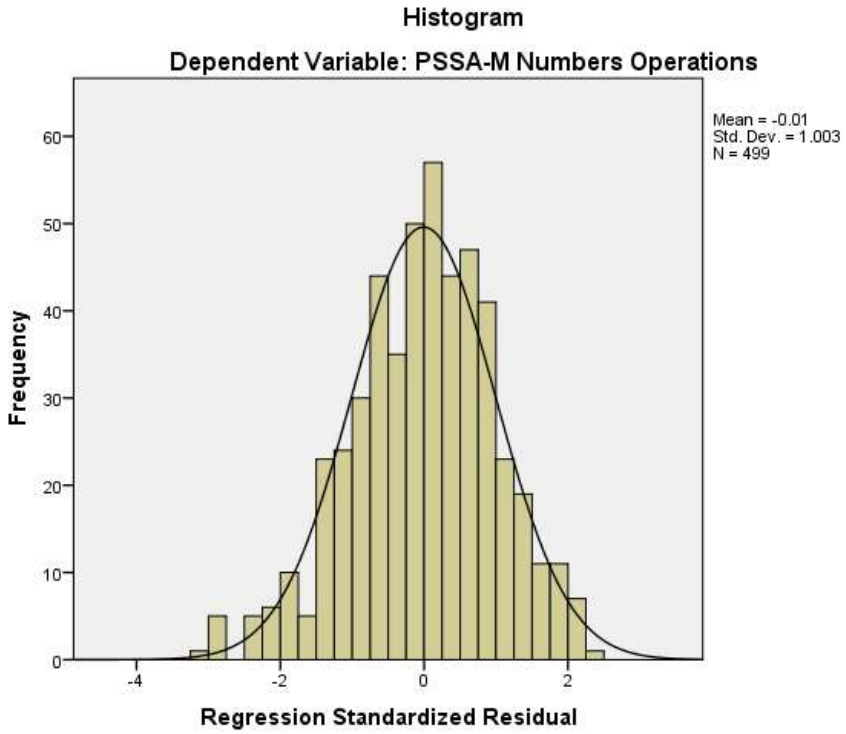


Figure 50. Histogram of Pennsylvania System of School Assessment, Mathematics Numbers and Operations standardized residuals for the first grade cohort.

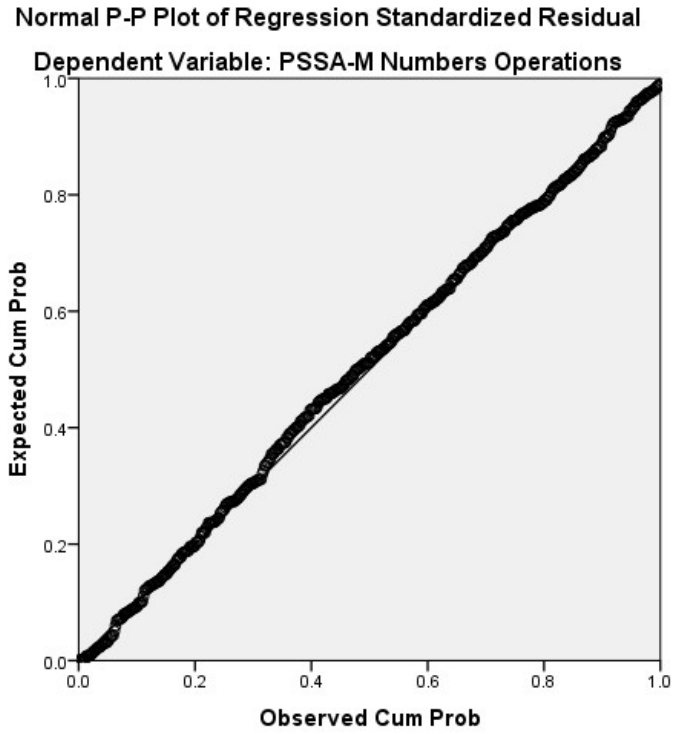


Figure 51. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Numbers and Operations standardized residuals for the first grade cohort.

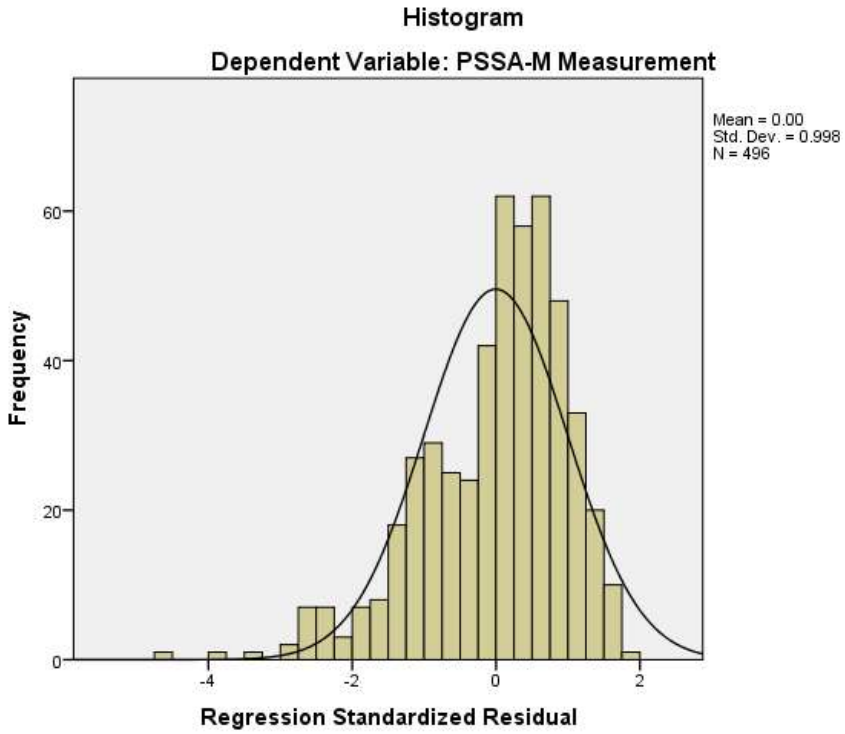


Figure 52. Histogram of Pennsylvania System of School Assessment, Mathematics Measurement standardized residuals for the first grade cohort.

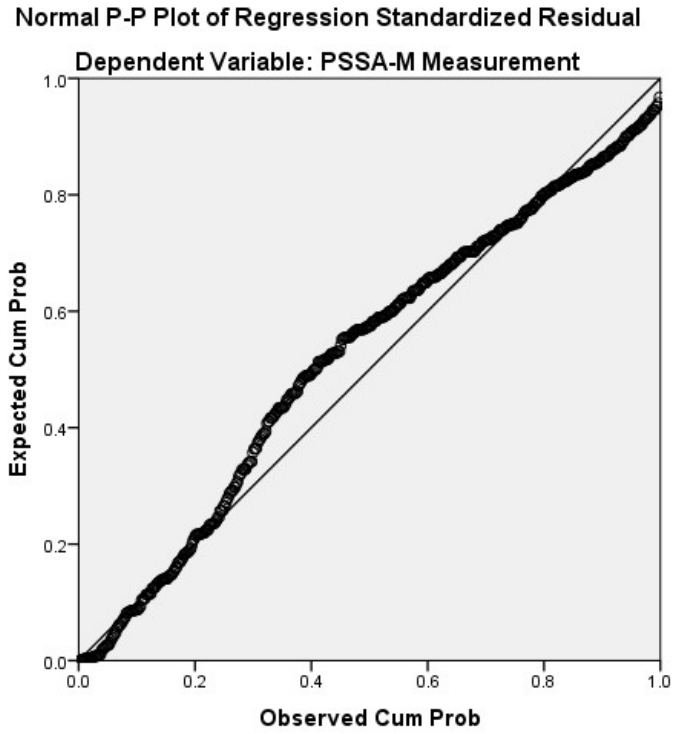


Figure 53. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Measurement standardized residuals for the first grade cohort.

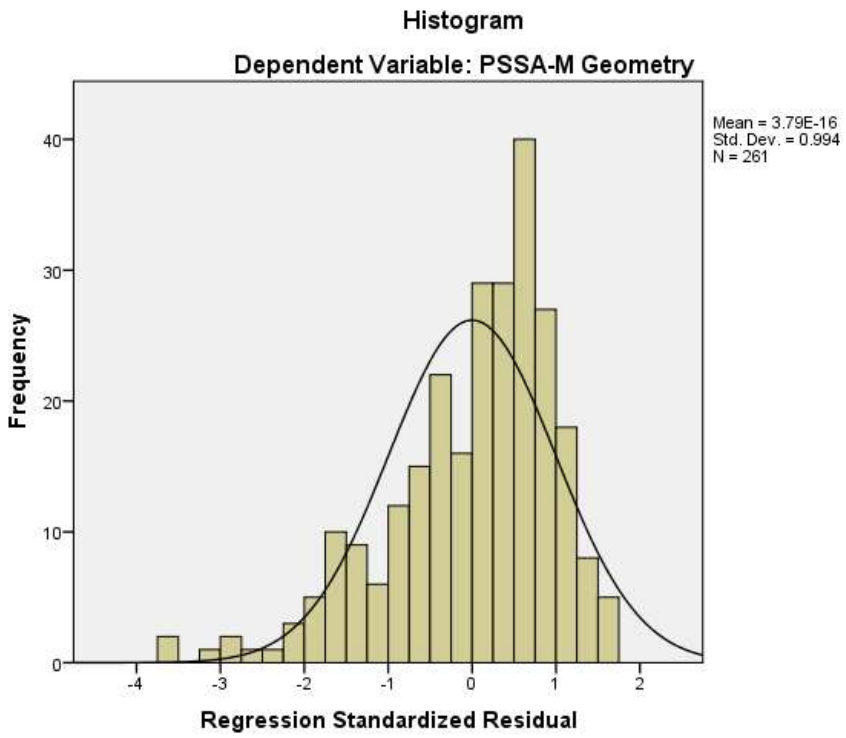


Figure 54. Histogram of Pennsylvania System of School Assessment, Mathematics Geometry standardized residuals for the first grade cohort.

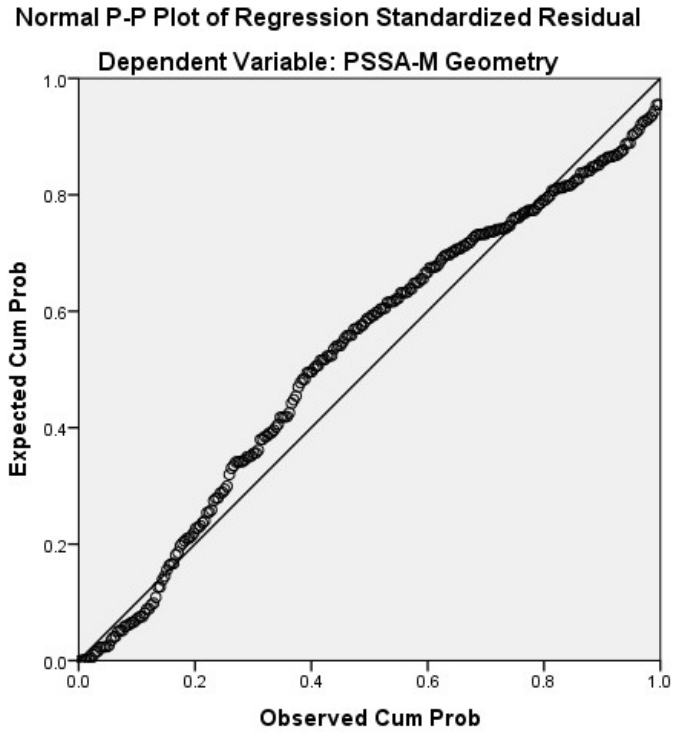


Figure 55. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Geometry standardized residuals for the first grade cohort.

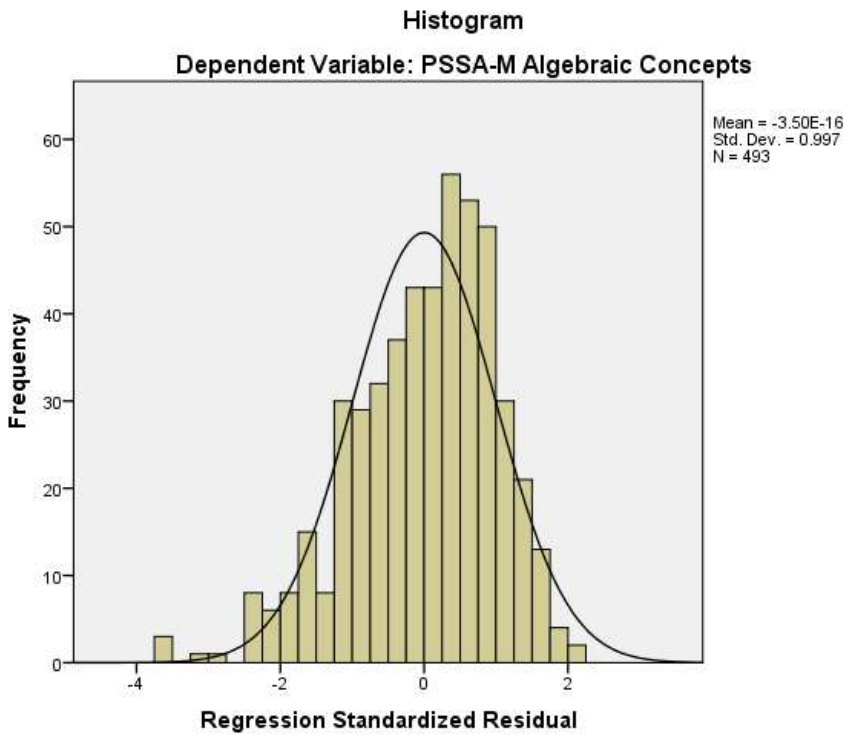


Figure 56. Histogram of Pennsylvania System of School Assessment, Mathematics Algebraic Concepts standardized residuals for the first grade cohort.

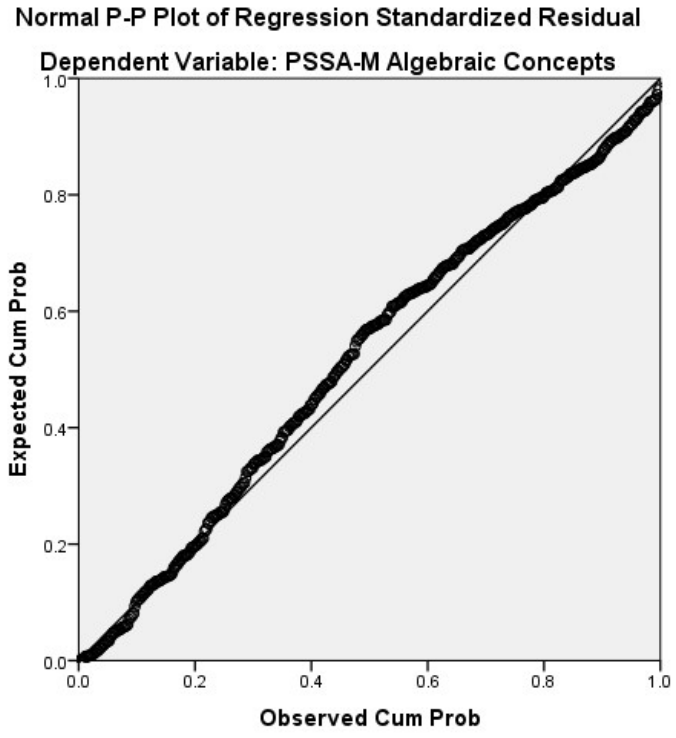


Figure 57. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Algebraic Concepts standardized residuals for the first grade cohort.

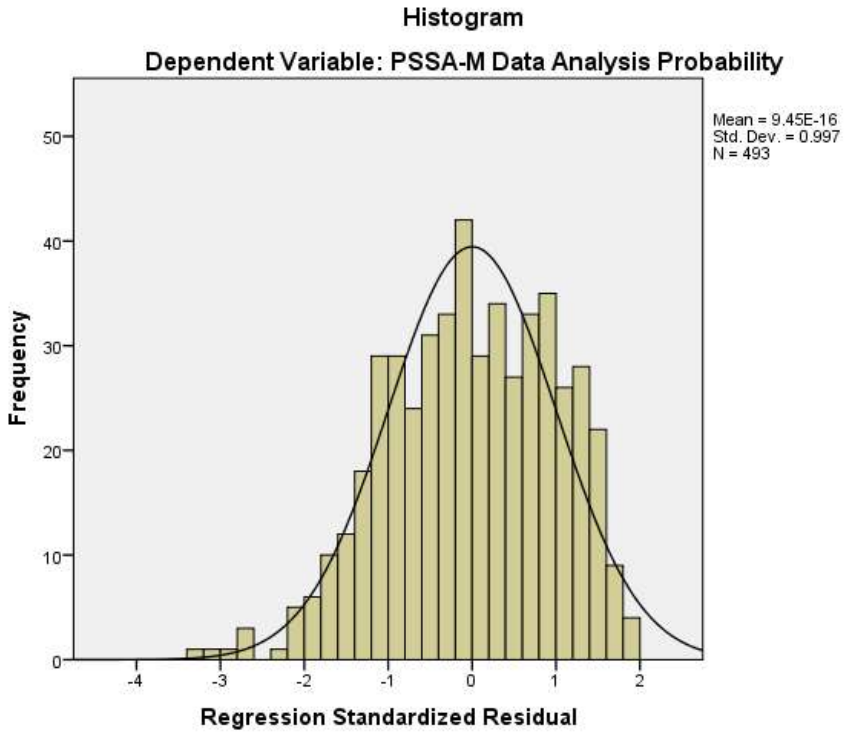


Figure 58. Histogram of Pennsylvania System of School Assessment, Mathematics Data Analysis and Probability standardized residuals for the first grade cohort.

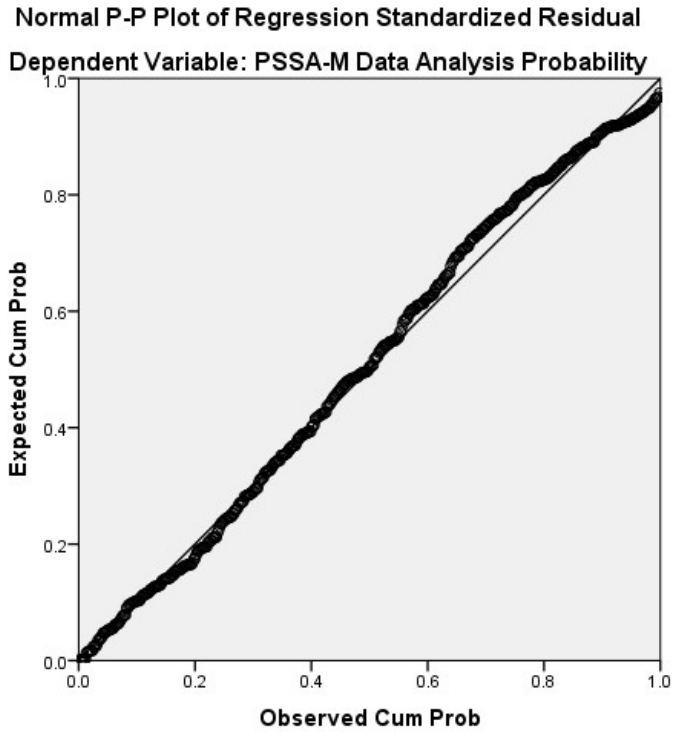


Figure 59. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Data Analysis and Probability standardized residuals for the first grade cohort.

Second grade multicollinearity statistics. Histograms and Normal P-P plots of the standardized residuals were generated and visually inspected to evaluate whether second grade data were multicollinear. Multicollinearity was not observed with PSSA-M Composite data and on all PSSA-M subtests, as evidenced by normal distributions on both histograms and Normal P-P plots. A slight left skew was observed on the Normal P-P plots for the Geometry subtest, Measurement subtest, and Algebraic Concepts subtest. However, the left skewed was not significant enough to indicate a violation of the multicollinearity assumption. It is not necessary for data points to be perfectly aligned with the diagonal as MLR is robust to deviations from normality (Laerd Statistics, 2015). Therefore, the multicollinearity assumption was met for second grade data. Please refer to Figures 60 through 71.

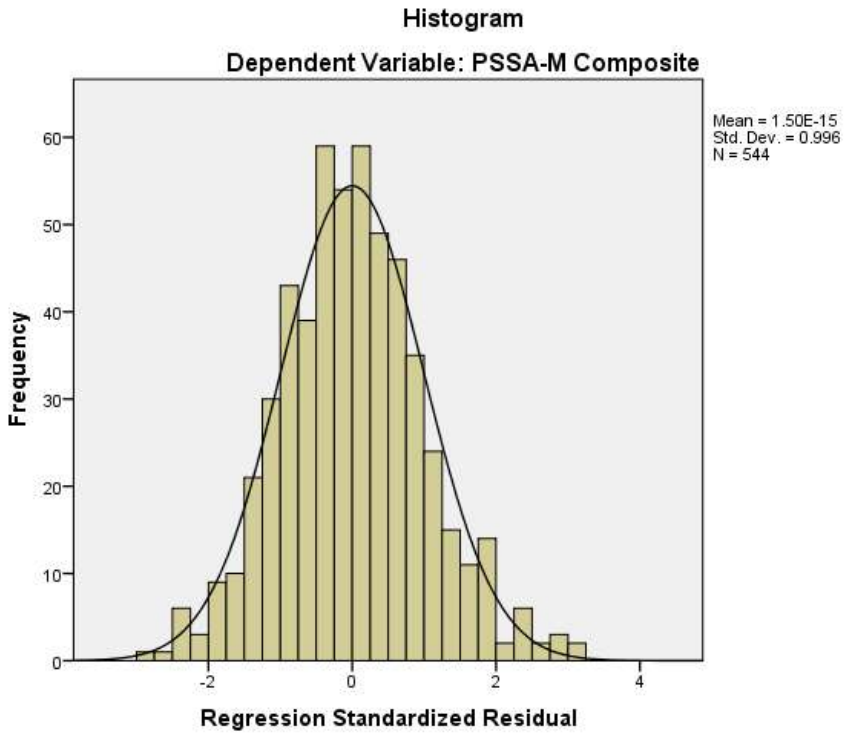


Figure 60. Histogram of Pennsylvania System of School Assessment, Mathematics Composite standardized residuals for the second grade cohort.

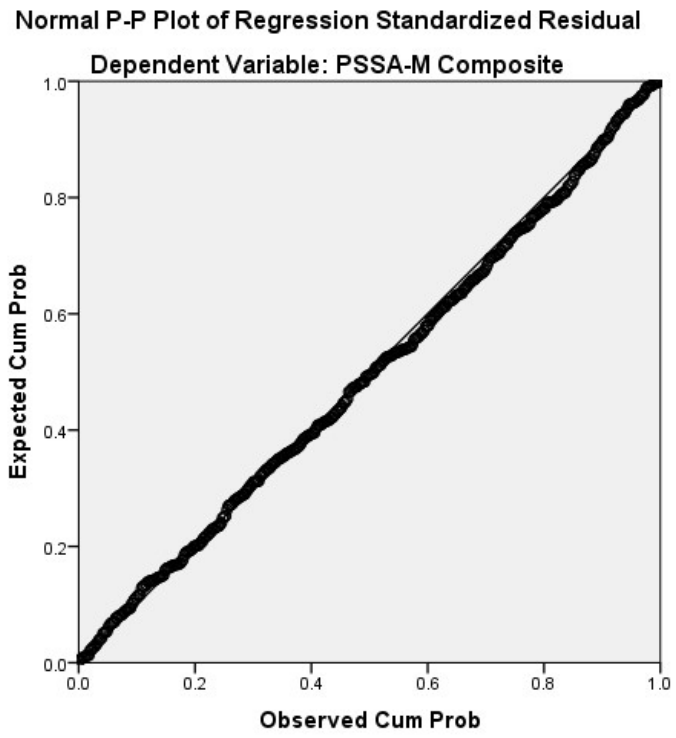


Figure 61. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Composite standardized residuals for the second grade cohort.

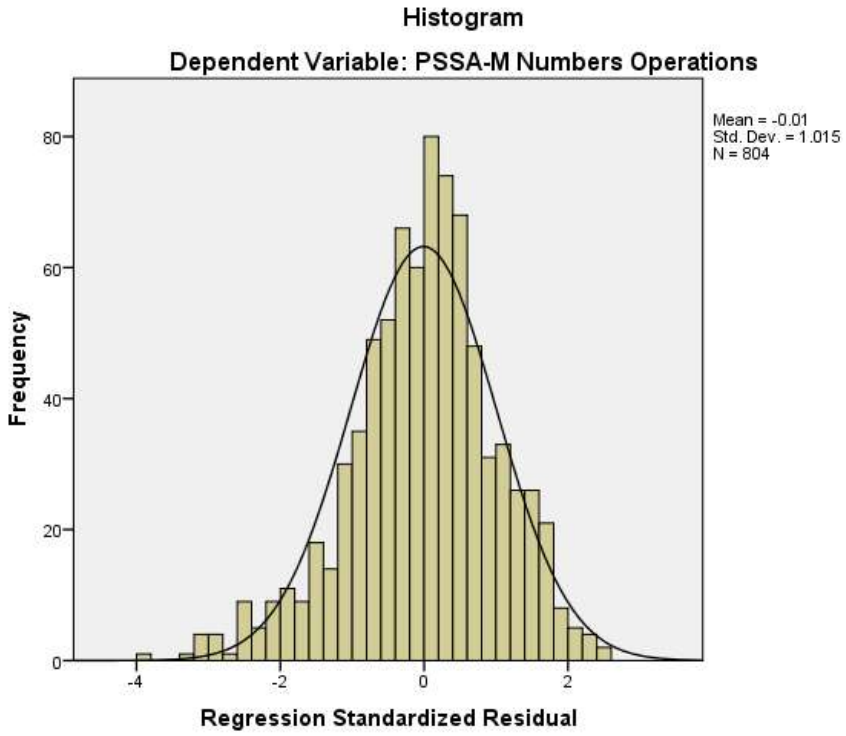


Figure 62. Histogram of Pennsylvania System of School Assessment, Mathematics Numbers and Operations standardized residuals for the second grade cohort.

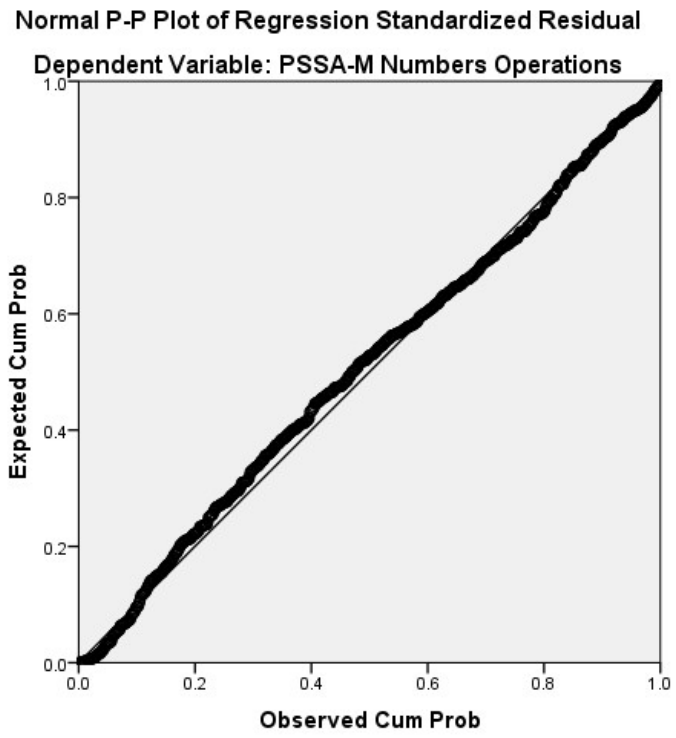


Figure 63. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Numbers and Operations standardized residuals for the second grade cohort.

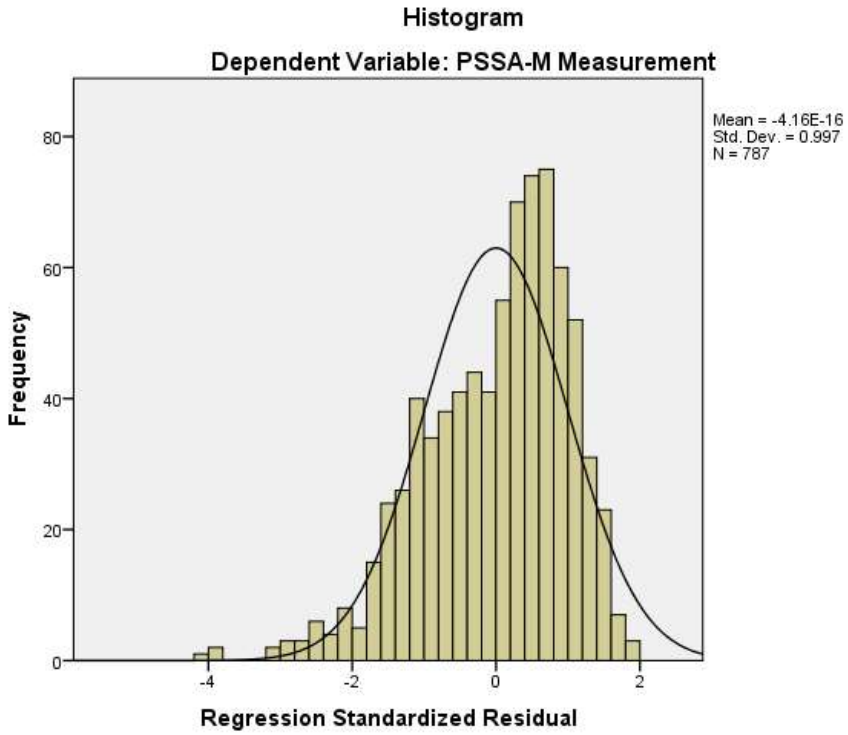


Figure 64. Histogram of Pennsylvania System of School Assessment, Mathematics Measurement standardized residuals for the second grade cohort.

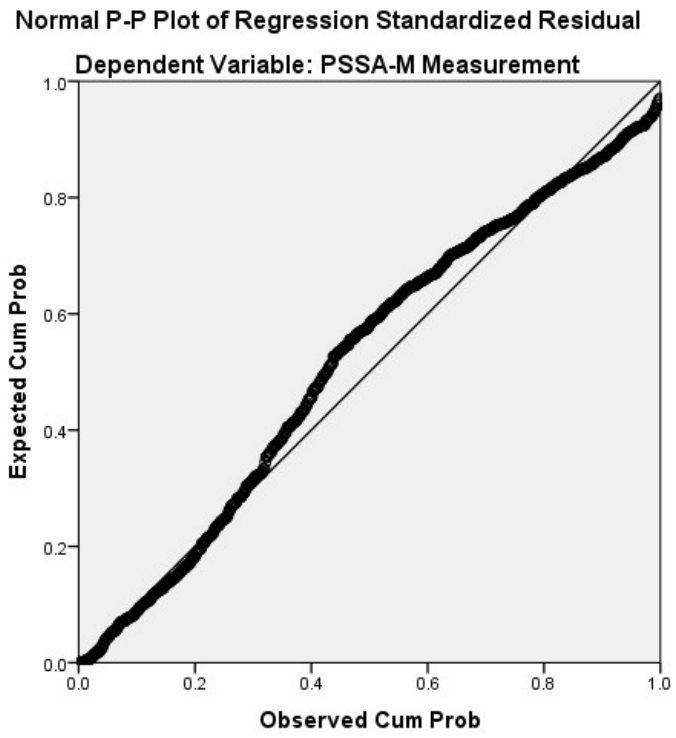


Figure 65. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Measurement standardized residuals for the second grade cohort.

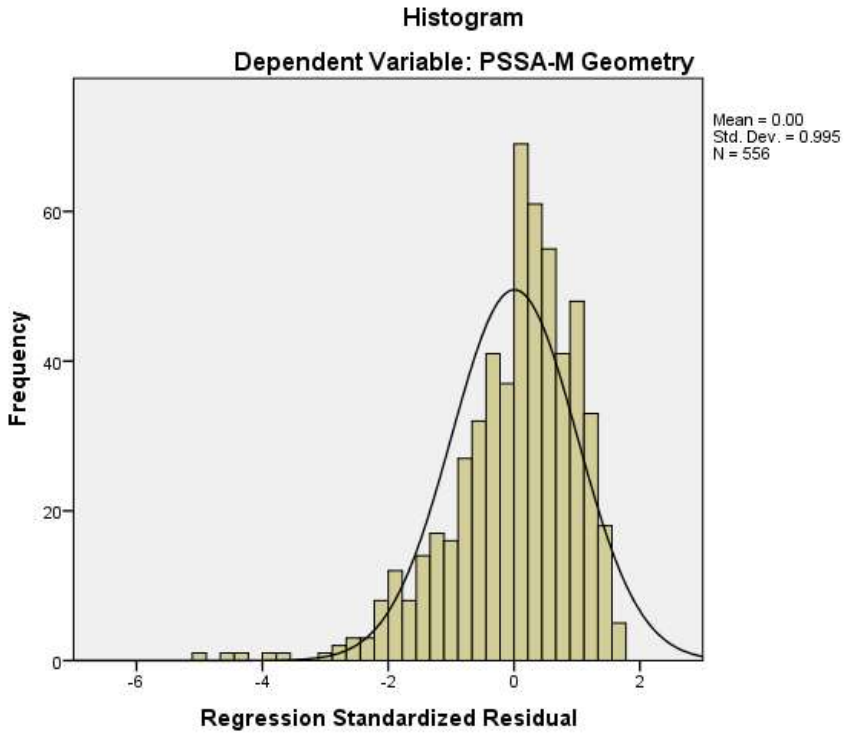


Figure 66. Histogram of Pennsylvania System of School Assessment, Mathematics Geometry standardized residuals for the second grade cohort.

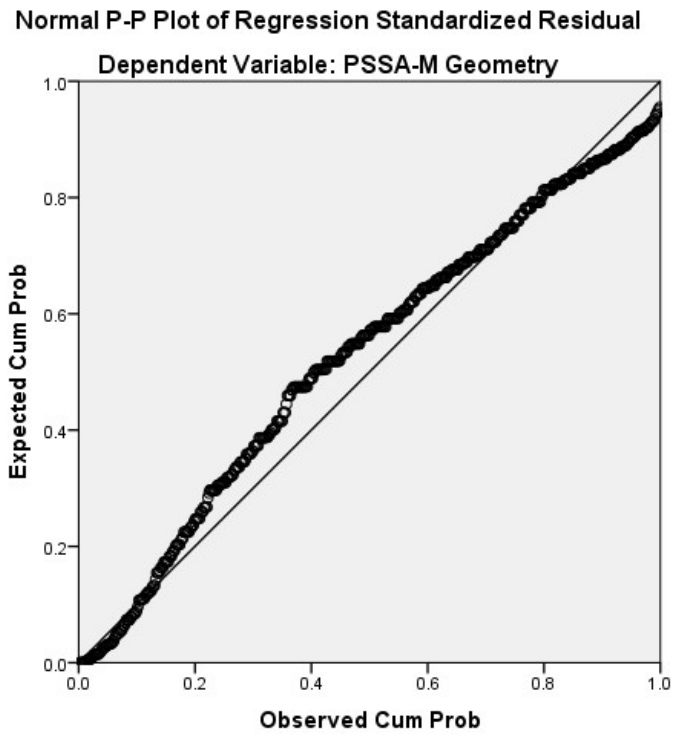


Figure 67. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Geometry standardized residuals for the second grade cohort.

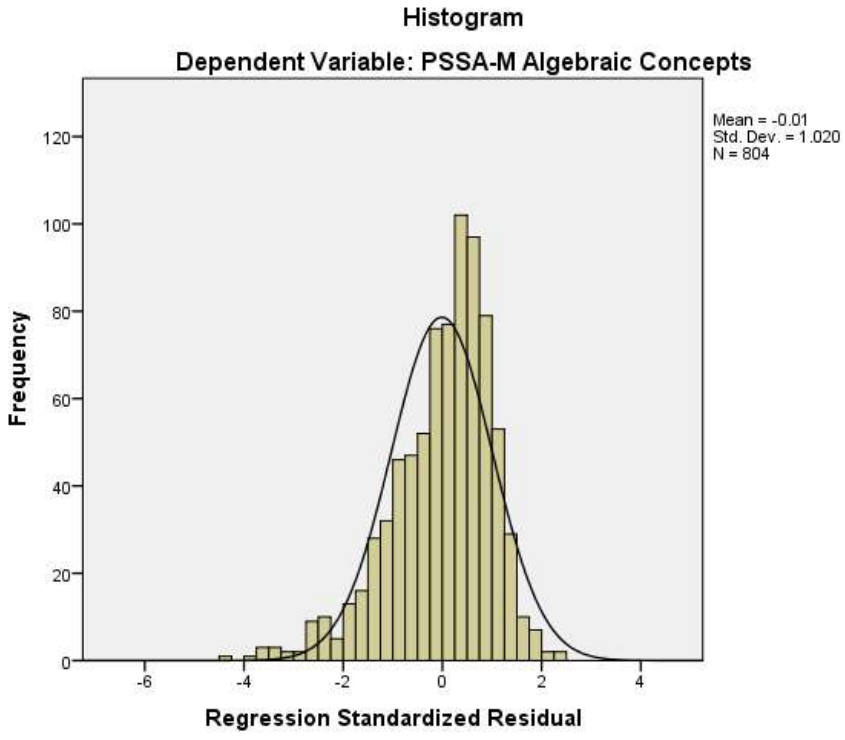


Figure 68. Histogram of Pennsylvania System of School Assessment, Mathematics Algebraic Concepts standardized residuals for the second grade cohort.

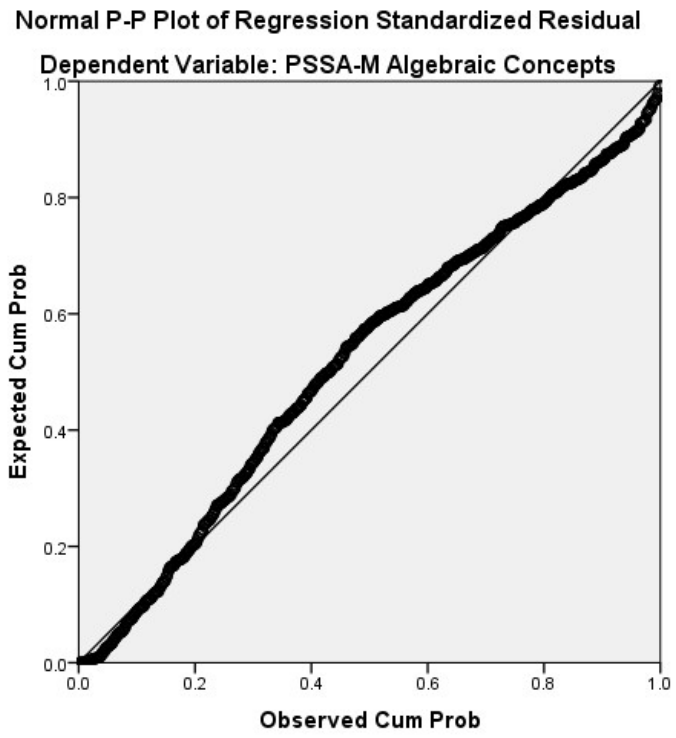


Figure 69. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Algebraic Concepts standardized residuals for the second grade cohort.

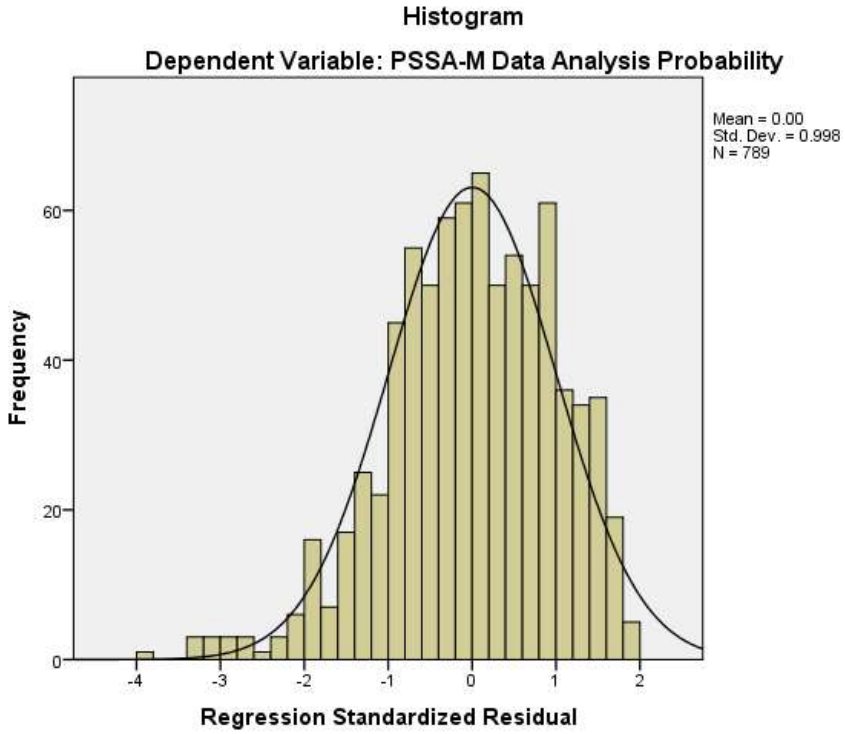


Figure 70. Histogram of Pennsylvania System of School Assessment, Mathematics Data Analysis and Probability standardized residuals for the second grade cohort.

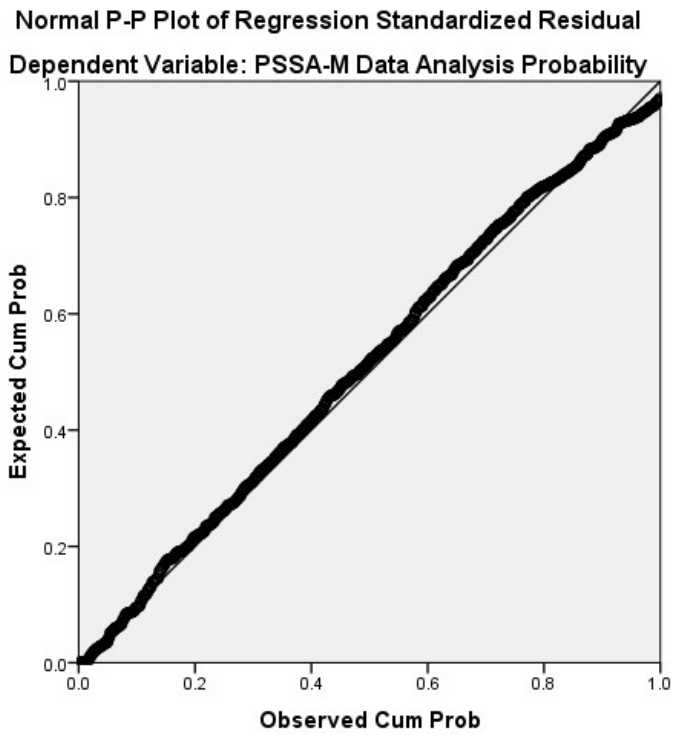


Figure 71. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Data Analysis and Probability standardized residuals for the second grade cohort.

Third grade multicollinearity statistics. Histograms and Normal P-P plots of the standardized residuals were generated and visually inspected to evaluate whether third grade data were multicollinear. Multicollinearity was not observed with PSSA-M Composite data and on all PSSA-M subtests, as evidenced by normal distributions on both histograms and Normal P-P plots. A slight left skew was observed on the Normal P-P plots for the Geometry subtest, Measurement subtest, and Algebraic Concepts subtest. However, the left skewed was not significant enough to indicate multicollinearity. It is not necessary for data points to be perfectly aligned with the diagonal as MLR is robust to deviations from normality (Laerd Statistics, 2015). Please refer to Figures 72 through 83. Therefore, the multicollinearity assumption was met for third grade data.

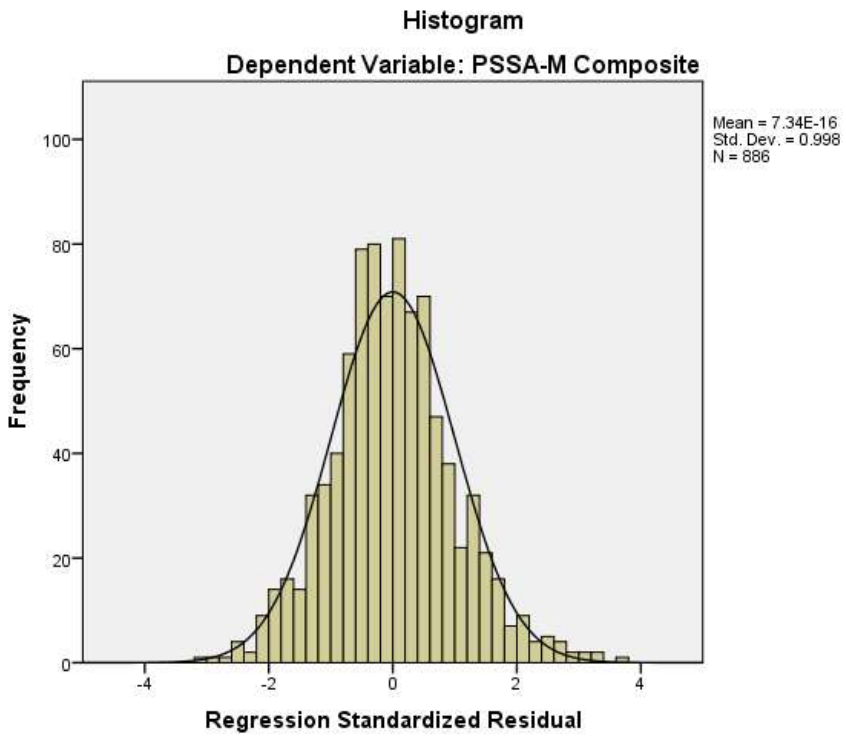


Figure 72. Histogram of Pennsylvania System of School Assessment, Mathematics Composite standardized residuals for the third grade cohort.

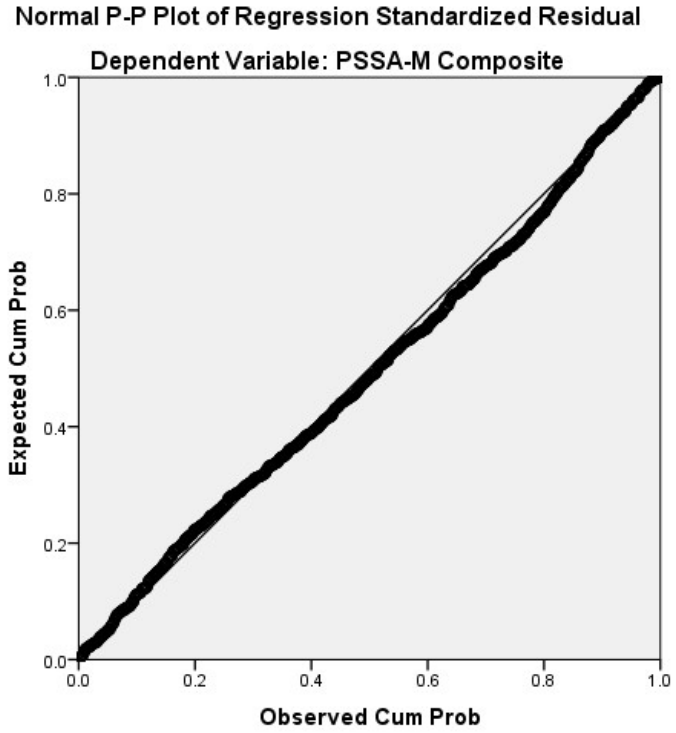


Figure 73. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Composite standardized residuals for the third grade cohort.

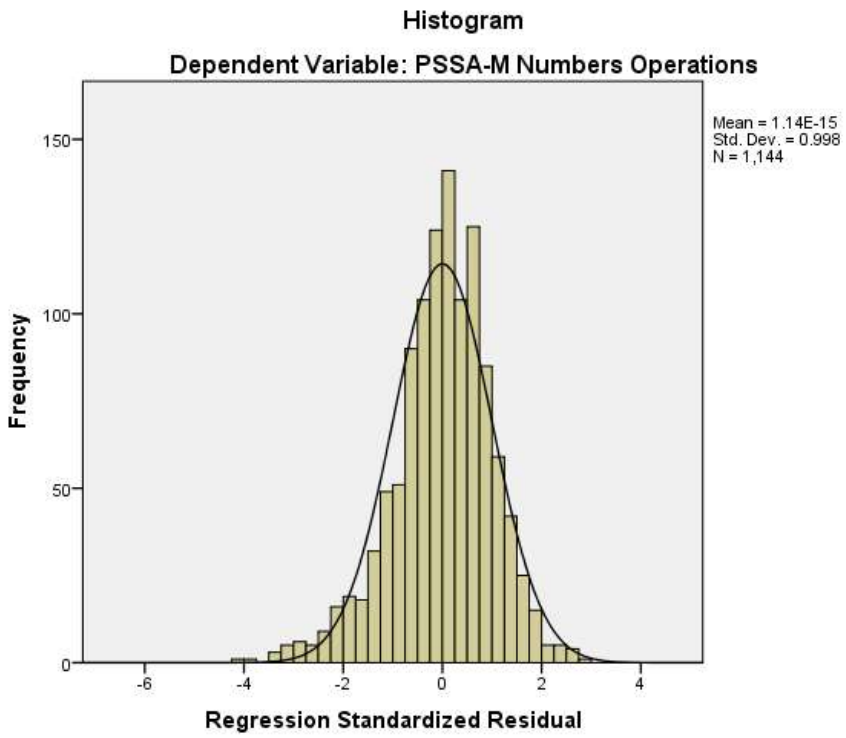


Figure 74. Histogram of Pennsylvania System of School Assessment, Mathematics Numbers and Operations standardized residuals for the third grade cohort.

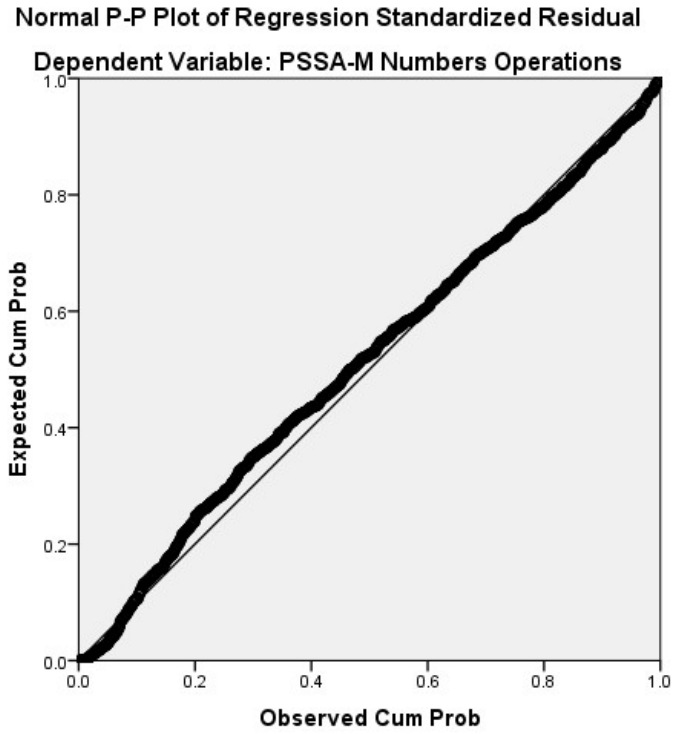


Figure 75. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Numbers and Operations standardized residuals for the third grade cohort.

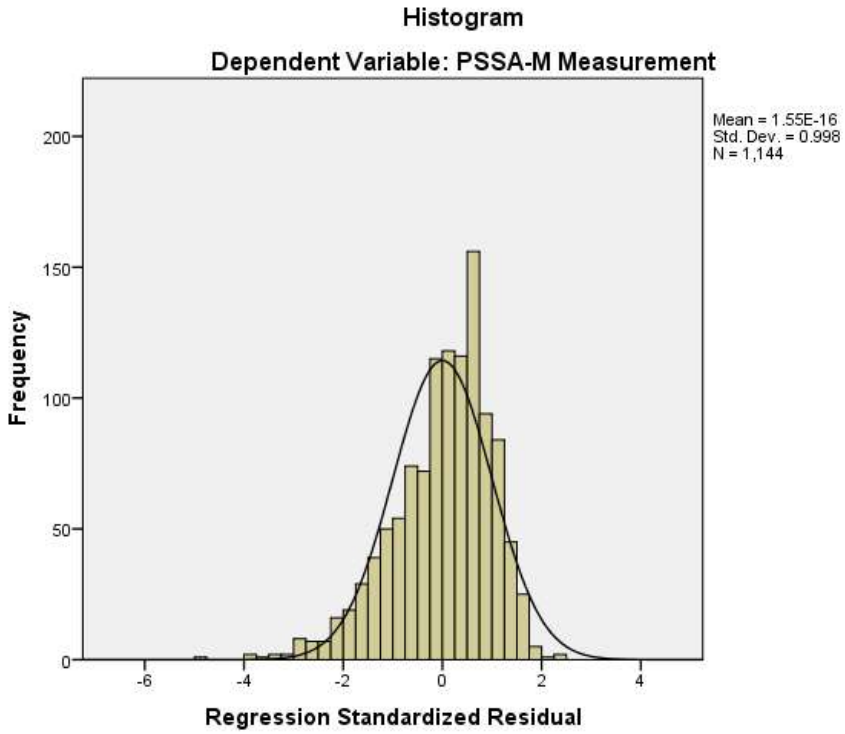


Figure 76. Histogram of Pennsylvania System of School Assessment, Mathematics Measurement standardized residuals for the third grade cohort.

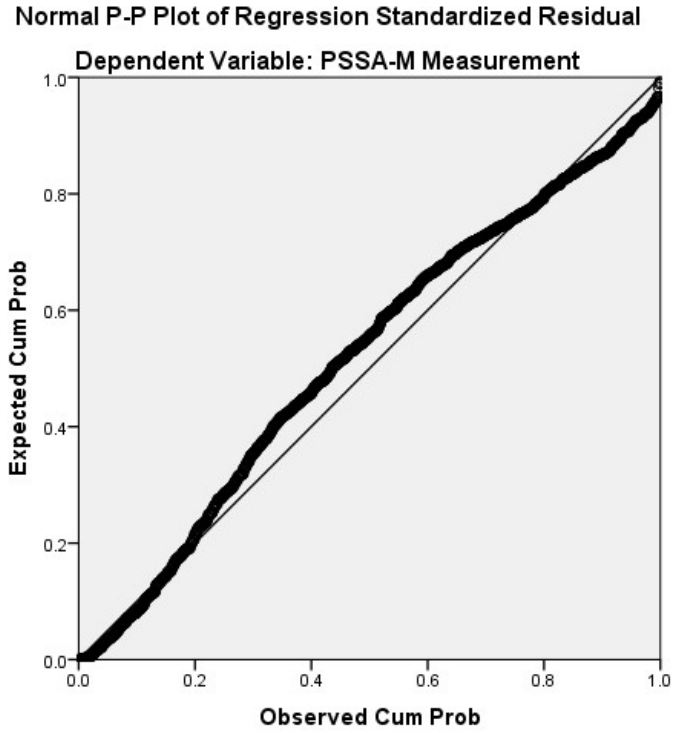


Figure 77. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Measurement standardized residuals for the third grade cohort.

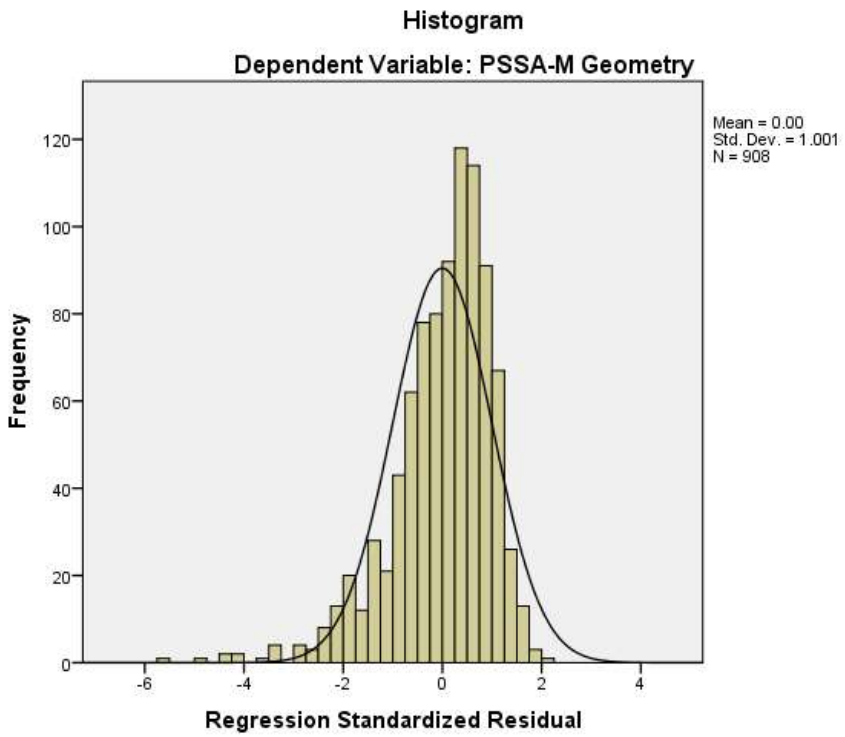


Figure 78. Histogram of Pennsylvania System of School Assessment, Mathematics Geometry standardized residuals for the third grade cohort.

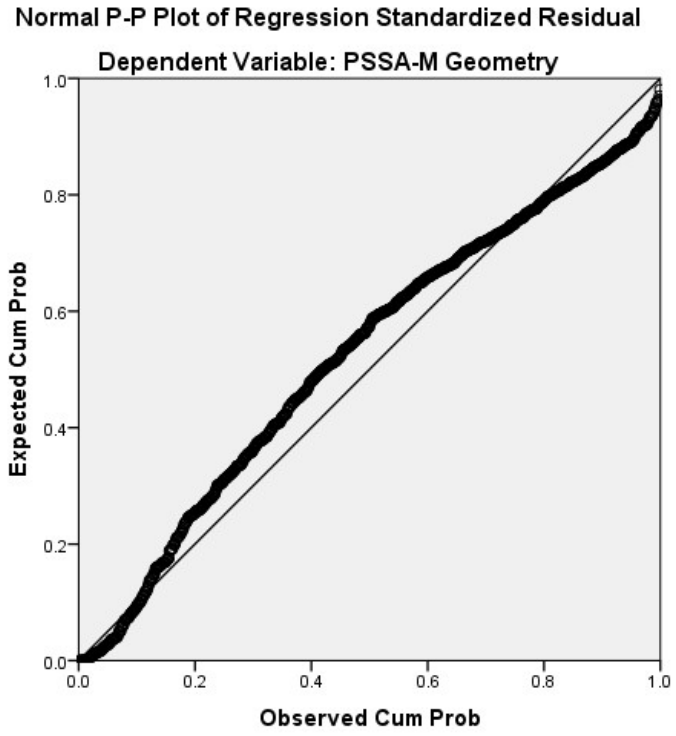


Figure 79. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Geometry standardized residuals for the third grade cohort.

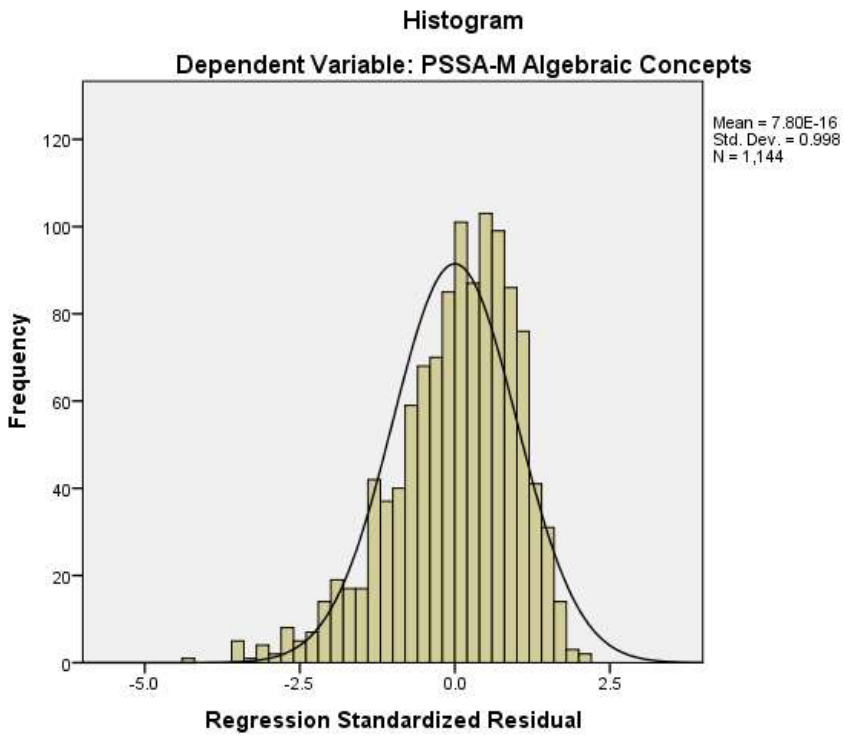


Figure 80. Histogram of Pennsylvania System of School Assessment, Mathematics Algebraic Concepts standardized residuals for the third grade cohort.

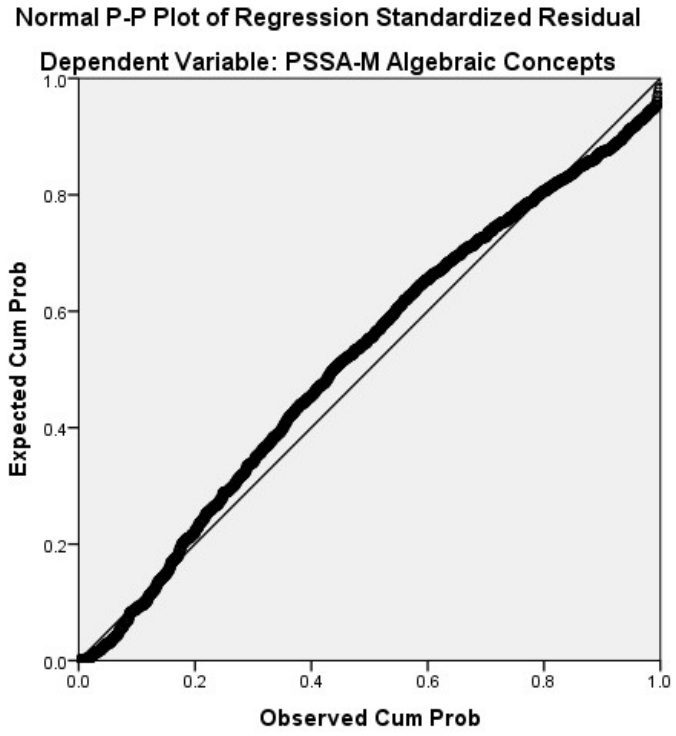


Figure 81. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Algebraic Concepts standardized residuals for the third grade cohort.

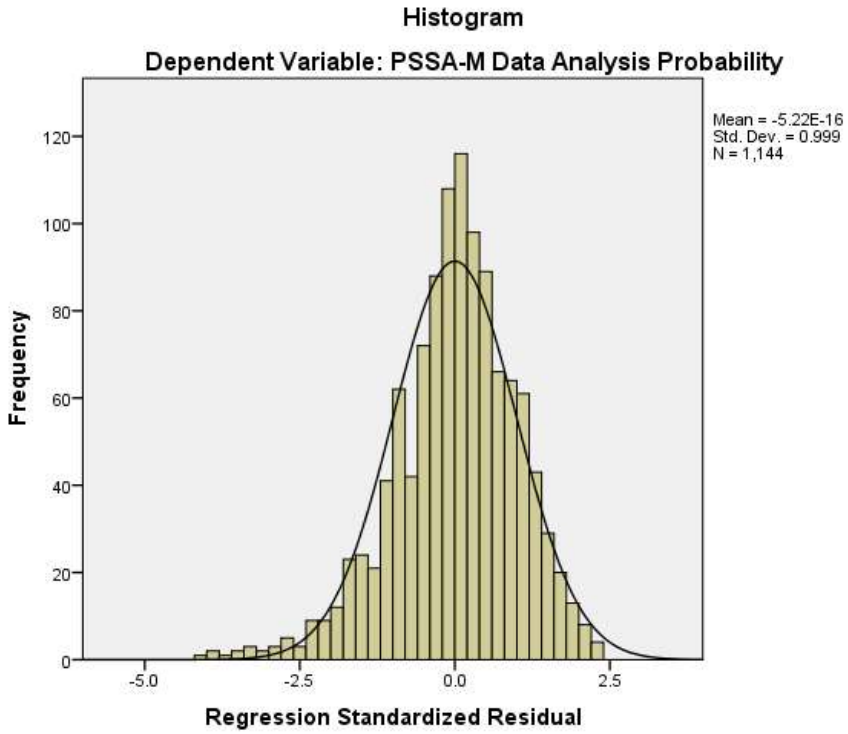


Figure 82. Histogram of Pennsylvania System of School Assessment, Mathematics Data Analysis and Probability standardized residuals for the third grade cohort.

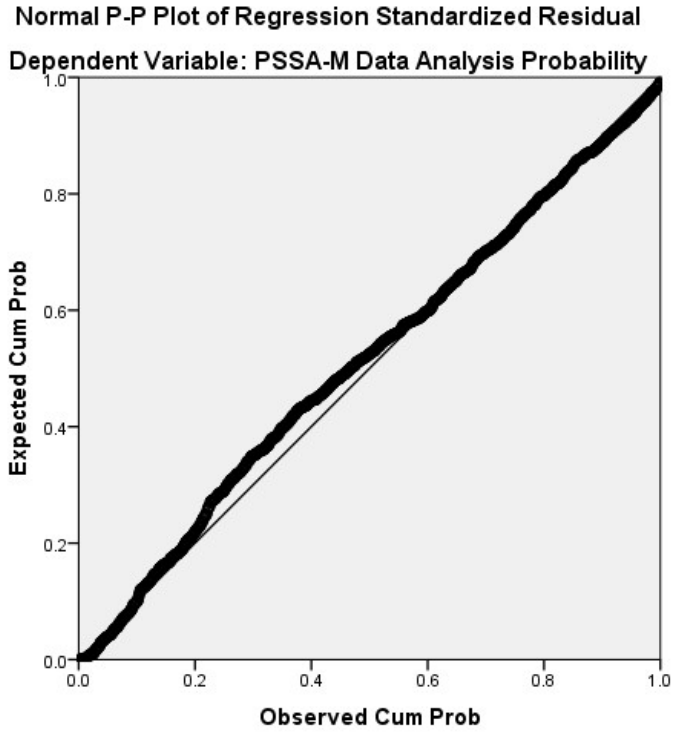


Figure 83. Normal P-P plot of Pennsylvania System of School Assessment, Mathematics Data Analysis and Probability standardized residuals for the third grade cohort.

Multiple Linear Regression

After reviewing the assumptions required for MLR, it was determined the data demonstrated the properties necessary for this statistical analysis. MLR is a statistical procedure that can be used to determine how multiple independent variables contribute to variance in a dependent variable (Aldrich & Cunningham, 2016). While there are different forms of MLR, the present study utilized stepwise MLR to analyze the relationship between MBSP-C, sex, and resource availability and PSSA-M data in order to determine which of the independent variables accounted for a significant amount of variance in the dependent variable.

In a stepwise approach, independent variables are entered into the regression equation based on their strength of correlation with the dependent variable. This means that if an independent variable does not have a significant contribution to the dependent variable, it is excluded from the regression equation. Due to questions regarding whether or not sex and resource availability had any impact on PSSA-M performance, it was decided a stepwise regression analysis would be the most appropriate analysis. This approach was preferred over other regression analysis methods which forced the inclusion of all the independent variables. "A stepwise regression analysis enters predictor variables into the equation based on their ability to predict unique variance in the outcome variable – variance that is not already predicted by predictor variables that are already in the equation." (Leary, 2001, p. 166).

In this study, the independent variables are sex, resource availability (measured by free and reduced lunch status), and MBSP-C probes in the fall, winter, and spring of first, second, and third grades. Dependent variables include PSSA-M Composite scores and five PSSA-M subtests (Numbers and Operations, Measurement, Geometry, Algebraic Concepts, and Data Analysis and Probability). Stepwise MLR will generate multiple regression models, up to five if all of the

independent variables have a significant contribution. Each model generated from stepwise MLR is compared. A determination is made whether or not the full regression model is a good fit for the data and regression coefficients are reported. The full or final regression model is used to determine if the model is a good fit for the data; in other words, whether or not the independent variables are able to predict the criterion measures in a statistically significant way. Regression coefficients are also reported for the final regression model. The unstandardized regression coefficient, *B*, represents the strength of the relationship between the predictor and the outcome of the in same unit of measurement as the predictor. "It is the change in the outcome associated with a unit change in the predictor" (Fields, 2013, p. 870). The standardized regression coefficient, β, indicates the strength of the relationship between a predictor and the outcome in a standardized manner. This standardization allows for comparisons between β values (Fields, 2013; Laerd Statistics, 2015).

In addition to *B* and β values, MLR generates the multiple correlation coefficient (*R),* the coefficient of determination (R^2) and the adjusted coefficient of determination (adjusted R^2). *R* depicts the correlation between the dependent and all independent variables. R^2 represents the how much of the variance in the dependent variable can be explained by all of the independent variables (Fields, 2013; Laerd Statistics, 2015). The adjusted R^2 is defined as,

> A measure of the loss of predictive power or shrinkage in regression. The adjusted R^2 tells us how much variance in the outcome would be accounted for if the model had been derived from the population from which the sample was taken. (Fields, 2013, pp. 870)

The *R*, R^2, and adjusted R^2 provide additional information and are included in the interpretation of the MLR results below. *R* values and adjusted R^2 values are highlighted more frequently than the R^2. This is because the adjusted R^2 is thought to represent a more accurate value of variance

since it corrects positive bias that may occur from the sample population. The adjusted R^2 accounts for any sampling bias that may have inflated R^2 (Laerd Statistics, 2015). The R^2 and adjusted R^2 are also referred to as the effect size. R^2 and adjusted R^2 values equal to or greater than 0.26 are considered substantial, less than 0.26 to 0.13 are considered moderate, and less than 0.13 to 0.02 are categorized as weak effect sizes (Cohen, 1988).

First Grade Multiple Linear Regression

First grade data from all the independent variables were entered into a stepwise MLR for each of the PSSA-M dependent variables. Table 14 provides a summary of the regression models with corresponding, R, R^2, and adjusted R^2 values, which are reviewed in subsequent sections.

Table 14

Summary of First Grade Stepwise Regression Models

Dependent Variable	Model and Predictors Variables	*R*	R^2	Adjusted R^2
PSSA-M Composite	Model 1: MBSP-C spring	.513	.264	.261
	Model 2: MBSP-C spring, MBSP-C fall	.559	.312	.307
	Model 3: MBSP-C spring, MBSP-C fall, MBSP-C winter	.575	.331	.323
	Model 4: MBSP-C spring, MBSP-C fall, MBSP-C winter, sex	.586	.343	.333
Numbers and Operations	Model 1: MBSP-C spring	.490	.240	.239
	Model 2: MBSP-C spring, MBSP-C Winter	.501	.251	.248
Measurement	Model 1: MBSP-C spring	.397	.158	.156
	Model 2: MBSP-C spring, MBSP-C fall	.409	.167	.164
Geometry	Model 1: MBSP-C winter	.409	.167	.164
	Model 2: MBSP-C winter, MBSP-C Spring	.430	.185	.178
	Model 3: MBSP-C winter, MBSP-C spring, MBSP-C fall	.447	.200	.191
Algebraic Concepts	Model 1: MBSP-C spring	.400	.160	.158
	Model 2: MBSP-C spring, MBSP-C Winter	.427	.182	.179
	Model 3: MBSP-C spring, MBSP-C winter, MBSP-C fall	.436	.190	.185
Data Analysis and Probability	Model 1: MBSP-C spring	.320	.103	.101
	Model 2: MBSP-C spring, MBSP-C fall	.336	.113	.109
	Model 3: MBSP-C spring, MBSP-C fall, MBSP-C winter	.350	.122	.117

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics. PSSA Composite $N = 274$, Numbers and Operations $N = 499$, Measurement $N = 496$, Geometry $N = 274$, Algebraic Concepts $N = 493$, Data Analysis and Probability $N = 493$.

Multiple linear regression of PSSA-M Composite with first grade independent variables. A stepwise multiple linear regression was performed to predict performance on the third grade PSSA-M Composite from MBSP-C probes administered in the fall, winter, and spring of first grade, sex, and resource availability. The analysis generated four regression

models. First grade MBSP-C data in the fall, winter, and spring and sex statistically significantly predicted performance on the third grade PSSA-M Composite, F (4, 256) = 33.424, $p < .0005$. The model is a good fit for the data. The full regression model explained 33% of the variance on PSSA-M Composite scores.

The R value for the first regression model which included MBSP-C in the spring of first grade was .513. The second regression model for PSSA-M Composite included MBSP-C data for the spring and fall of first grade generated an R value of .559. The third regression model added MBSP-C in the winter of first grade and increased R to .575. The fourth regression model included sex with $R = .586$. These R values indicate a strong positive predictive relationship between the first independent variables included in the regression model and PSSA-M composite scores in third grade (Cohen, 1977).

Based on the first model, first grade MBSP-C spring data accounts for the largest variance, 26% of overall PSSA-M performance which was administered in the spring of third grade. The second regression model included first grade MBSP-C fall which increased the variance accounted for by the model to 31%. The adjusted R^2 increased to 32% with the addition of MBSP-C in the winter of first grade. The full regression model which included MBSP-C in the fall, winter, and spring and sex generated a R^2 value of .343 and adjusted R^2 value of .333 or 33%. An effect size of 33% is considered substantial (Cohen, 1988).

The increase in variance with the addition of MBSP-C in the winter and sex of student is statistically significant ($p = .007$; $p = .032$). MBSP-C winter data are easily available in systems where universal screenings are administered three times a year. Therefore, while the percentage of additional variance explained appears small, it is worth including in the regression model. However, educational systems may choose to exclude student sex data without significantly

impacting the predictive power of the regression model. Resource availability in first grade did not significantly contribute to the overall variance of PSSA-M Composite in the spring of third grade.

Regression coefficients and standard errors for the full model, Model 4, are summarized in Table 15. The full model is reported because it accounts for the most variance and includes the independent variables that have a statistically significant impact on the dependent variable.

Table 15

Stepwise Multiple Regression Predicting Third Grade PSSA-M Composite From First Grade MBSP-C Fall, Winter, and Spring Data and Sex of Student.

Variable	*B*	*SE B*	β
Constant	1200.723**	32.088	
MBSP-C spring	7.453	1.659	.316**
MBSP-C fall	8.296	2.514	.188**
MBSP-C winter	5.439	1.987	.202*
Sex	-33.610	15.594	-.110*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics*; SE B* = Standard error of unstandardized regression coefficient. *N* = 274. * $p < .05$, ** $p < .001$

The standardized regression coefficient (β) represents the impact an independent variable has on the dependent variable. Since the β is standardized, comparisons can be made between β values. This comparison is not possible with the unstandardized regression coefficient, *B*. All of the β values from the independent variables included in the full regression model are statistically significant. However, a comparison of the β values indicate first grade MBSP-C spring data has the most significant impact on third grade PSSA-M Composite scores. First grade MBSP-C fall and winter scores have a similar impact on third grade PSSA-M Composite scores. The relationship between the independent variables and dependent variable is positive. This means that in a prediction model, an increase in MBSP-C scores results in an increase in PSSA-M Composite scores. The β value of sex, which is a dichotomous variable, is interpreted to mean females in the first grade cohort demonstrated higher scores on the third grade PSSA-M

composite. Based on first grade data, PSSA-M Composite scores increased 7.453 points for each MBSP-C spring digit correct, 8.296 points for each MBSP-C fall digit correct, and 5.439 points for each MBSP-C spring digit correct. Females performed 33.610 points higher than males.

Multiple linear regression of PSSA-M Numbers and Operations subtest with first grade independent variables. A stepwise multiple linear regression was performed to predict performance on the third grade PSSA-M Numbers and Operations subtest from MBSP-C probes administered in the fall, winter, and spring of first grade, sex, and resource availability. The stepwise regression generated two regression models, both good fits to the data. MBSP-C in the spring and winter of first grade were found to statistically significantly predict performance on the third grade PSSA-M Numbers and Operations subtest, F (2, 490) = 82.184, $p < .0005$. The R^2 for the full regression model was 25% with an adjusted R^2 of 25%, which indicates a moderate effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C in the spring of first grade accounted for the most amount of variance on third grade PSSA-M Numbers and Operations performance (R = .490; adjusted R^2 = .239). An R value of .490 indicated a moderately strong predictive relationship (Cohen, 1977). An effect size of 24% is considered moderate (Cohen, 1988). The second regression model included the MBSP-C in the winter of first grade which resulted in a statistically significant (p = .008) increase in total variance to 25% (R = .501; adjusted R^2 = .251). This is interpreted to mean 25% of variance on third grade PSSA-M Numbers and Operations subtest can be explained by MBSP-C scores administered in the winter and spring of first grade. The amount of variance explained by the regression model is considered moderate. The inclusion of MBSP-C winter in the regression model increased the strength of the relationship to over .50 which indicates a strong relationship (Cohen, 1977).

Resource availability, sex, and MBSP-C in the fall of first grade did not significantly predict performance on third grade PSSA-M Numbers and Operations subtest or significantly contribute to the variance accounted for by the model.

Regression coefficients and standard errors for the full model, Model 2, are summarized in Table 16. The full model is reported because it accounts for the most variance and includes the independent variables which have a statistically significant impact on the dependent variable.

Table 16

Stepwise Multiple Regression Predicting Third Grade PSSA-M Numbers and Operations subtest from First Grade MBSP-C Winter and Spring Data.

Variable	*B*	*SE B*	β
Constant	21.714**	.691	
MBSP-C spring	.354	.049	.391**
MBSP-C winter	.142	.054	.143*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics*; SE B* = Standard error of unstandardized regression coefficient. $N = 510$. * $p < .05$, ** $p < .001$

All of the β values from the independent variables included in the full regression model are statistically significant. A comparison of β values indicate first grade MBSP-C spring data has the most significant impact on third grade PSSA-M Numbers and Operations scores (β = .391). First grade MBSP-C winter scores have a lower, but still significant impact on third grade PSSA-M Numbers and Operations scores. In the prediction model, the impact of MBSP-C in the spring is 2.73 times stronger than MBSP-C in the winter. The relationship between the independent variables and dependent variable is positive. This means an increase in MBSP-C scores represents a score increase on the PSSA-M Numbers and Operations subtest. Based on first grade data, PSSA-M Numbers and Operations subtest scores increased .354 points for each MBSP-C spring digit correct and .142 points for each MBSP-C winter digit correct.

Multiple linear regression of PSSA-M Measurement subtest with first grade independent variables. A stepwise multiple linear regression was performed to predict

performance on the third grade PSSA-M Measurement subtest from MBSP-C probes administered in the fall, winter, and spring of first grade, sex, and resource availability. The stepwise regression created two models. MBSP-C in the fall and spring of first grade were found to statistically significantly predict performance on the third grade PSSA-M Measurement subtest, F (2, 490) = 49.124, p < .0005. The R^2 for the full regression model was 17% with an adjusted R^2 of 16%. This indicates a moderate effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C in the spring of first grade accounted for the most amount of variance on third grade PSSA-M Measurement performance with a relationship that is moderate in magnitude (R = .397; adjusted R^2 = .156). The second regression model included the MBSP-C in the fall of first grade which increased total variance only slightly, remaining at 16% (R = .409; adjusted R^2 = .164). While the increase in variance was statistically significant (p = .021), it is not so significant that systems should delay analysis until spring data are available. Resource availability, sex, and MBSP-C in the winter of first grade did not significantly predict performance on third grade PSSA-M Measurement subtest.

Regression coefficients and standard errors for the full regression model are summarized in Table 17. The full model is reported because it accounts for the most variance and includes the independent variables which have a statistically significant impact on the dependent variable.

Table 17

Stepwise Multiple Regression Predicting Third Grade PSSA-M Measurement Subtest From First Grade MBSP-C Fall and Spring Data.

Variable	*B*	*SE B*	β
Constant	5.975**	.242	
MBSP-C spring	.111	.014	.358**
MBSP-C fall	.063	.027	.104*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics*; SE B* = Standard error of unstandardized regression coefficient. N = 510. * p < .05, ** p <.001

All of the β values from the independent variables included in the full regression model are statistically significant. A comparison of β values indicate first grade MBSP-C spring scores have the most significant impact on third grade PSSA-M Measurement scores. First grade MBSP-C fall scores have a lower, but still statistically significant impact on third grade PSSA-M Measurement scores. The relationship between the independent variables and dependent variables are positive. This positive relationship suggests when looking at the prediction model an increase in MBSP-C scores would result in increased PSSA-M Measurement scores. Based on first grade data, PSSA-M Measurement subtest scores increased .111 points for each MBSP-C spring digit correct and .063 points for each MBSP-C fall digit correct.

Multiple linear regression of PSSA-M Geometry subtest with first grade independent variables. A stepwise multiple linear regression was performed to predict performance on the third grade PSSA-M Geometry subtest from MBSP-C probes administered in the fall, winter, and spring of first grade, sex, and resource availability. The stepwise regression created three models, all with a good fit to the data. In the full regression model, MBSP-C in the fall, winter, and spring of first grade were found to statistically significantly predict performance on the third grade PSSA-M Geometry subtest, F (3, 257) = 21.431, p < .0005. The R^2 for the full regression model was 20% with an adjusted R^2 of 19%. This indicates a moderate effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C in the winter of first grade accounted for the most amount of variance on third grade PSSA-M Geometry performance with a moderately strong relationship (R = .409; adjusted R^2 = .164). The second regression model included the MBSP-C in the spring of first grade which resulted in a statistically significant (p = .019) increase in variance of the PSSA-M Geometry subtest, accounting for 18% of the variance

of MBSP-C (R = .430; adjusted R^2 = .185). The third regression model included fall of first grade MBSP-C data (R = .447; adjusted R^2 = .191). The increase in variance for this model was statistically significant (p = .027). The magnitude of the relationship remained moderate for the full regression model. Resource availability and sex in first grade did not significantly contribute to the prediction of third grade PSSA-M Geometry. Regression coefficients and standard errors for the full regression model, Model 3, are summarized in Table 18.

Table 18

Stepwise Multiple Regression Predicting Third Grade PSSA-M Geometry Subtest From First Grade MBSP-C Fall, Winter, and Spring Data.

Variable	*B*	*SE B*	β
Constant	7.086**	.233	
MBSP-C winter	.054	.020	.225*
MBSP-C spring	.037	.016	.173*
MBSP-C fall	.055	.025	.140*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics*; SE B* = Standard error of unstandardized regression coefficient. N = 274. * p < .05, ** p <.001

All of the β values from the independent variables included in the full regression model are statistically significant. A comparison of the β values indicate first grade MBSP-C winter scores have the most significant impact on third grade PSSA-M Geometry scores. First grade MBSP-C spring and fall scores have a lower, but still statistically significant impact on third grade PSSA-M Geometry scores. The relationship between the independent variables and dependent variable is positive. This means that in a prediction model, an increase in MBSP-C scores results in score increases on the PSSA-M Geometry subtest. Based on first grade data, PSSA-M Geometry subtest scores increased .054 points for each MBSP-C winter digit correct, .037 points for each MBSP-C spring digit correct, and .055 points for each MBSP-C fall digit correct.

Multiple linear regression of PSSA-M Algebraic Concepts subtest with first grade independent variables. A stepwise multiple linear regression was performed to predict performance on the third grade PSSA-M Algebraic Concepts subtest from MBSP-C probes administered in the fall, winter, and spring of first grade, sex, and resource availability. The stepwise regression created three models. MBSP-C in the fall, winter, and spring of first grade were found to statistically significantly predict performance on the third grade PSSA-M Algebraic Concepts subtest, F (3, 489) = 38.297, p < .0005. The R^2 for the full regression model was 19% with an adjusted R^2 of 19%. This indicates a moderate effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C in the spring of first grade accounted for the most amount of variance on third grade PSSA-M Algebraic Concepts performance (R = .400; adjusted R^2 = .158). The second regression model included the MBSP-C in the winter of first grade which increased total variance of PSSA-M Algebraic Concepts that can be accounted for by MBSP-C in the winter and spring of first grade to 18% (R = .427; adjusted R^2 = .179). The increase in variance accounted for with the addition of MBSP-C winter data into the regression model is statistically significant (p < .0005). The third regression model added first grade MBSP-C fall data (R = .436; adjusted R^2 = .190). The increase in total variance accounted for with the addition of MBSP-C fall data remained statistically significant, (p = .027). The R values indicate a relationship that is significant but moderate in magnitude. Resource availability and sex in first grade did not significantly contribute to the prediction of third grade PSSA-M Algebraic Concepts performance. Regression coefficients and standard errors for the full regression model are summarized in Table 19.

Table 19

Stepwise Multiple Regression Predicting Third Grade PSSA-M Algebraic Concepts Subtest From First Grade MBSP-C Fall, Winter, and Spring Data.

Variable	*B*	*SE B*	β
Constant	6.668**	.199	
MBSP-C spring	.060	.014	.242**
MBSP-C winter	.047	.016	.173*
MBSP-C fall	.050	.022	.101*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematic*s; SE B* = Standard error of unstandardized regression coefficient. $N = 510$. * $p < .05$, ** $p < .001$

All of the β values from the independent variables included in the full regression model are statistically significant. A comparison of β values indicate first grade MBSP-C spring data have the most significant impact on third grade PSSA-M Algebraic Concepts scores. First grade MBSP-C winter scores and MBSP-C fall scores have a lower, but still significant impact on third grade PSSA-M Algebraic Concepts outcomes. The relationship between the independent variables and dependent variables is positive, which means an increase in MBSP-C results in a score increase on the PSSA-M Algebraic Concepts subtest. Based on first grade data, PSSA-M Algebraic Concepts subtest scores increased .060 points for each MBSP-C spring digit correct, .047 points for each MBSP-C winter digit correct, and .050 points for each MBSP-C fall digit correct.

Multiple linear regression of PSSA-M Data Analysis and Probability subtest with first grade independent variables. A stepwise multiple linear regression was performed to predict performance on the third grade PSSA-M Data Analysis and Probability subtest from MBSP-C probes administered in the fall, winter, and spring of first grade, sex, and resource availability. The stepwise regression created three models. MBSP-C in the fall, winter, and spring of first grade were found to statistically significantly predict performance on the third grade PSSA-M Data Analysis and Probability subtest with a good model fit, $F(3, 489) = 22.708$,

$p < .0005$. The R^2 for the full regression model was 12% with an adjusted R^2 of 12%. This indicates a weak effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C in the spring of first grade accounted for the most amount of variance on third grade PSSA-M Data Analysis and Probability performance ($R = .320$; adjusted $R^2 = .101$). The second regression model included the MBSP-C in the fall of first grade which increased total variance of PSSA-M Data Analysis and Probability explained by MBSP-C in the fall and spring of first grade to 11% ($R = .336$; adjusted $R^2 = .109$). This increase in total variance is statistically significant ($p = .016$). The third regression model included winter of first grade MBSP-C data ($R = .350$; adjusted $R^2 = .117$). The increase in total variance explained by the addition of MBSP-C winter data is statistically significant ($p = .024$). However, the addition of MBSP-C winter had minimal impact on the variance for practical purposes.

The strength of the relationship between the full model and performance on the Data Analysis and Probability subtest is moderate in magnitude. Resource availability and sex in first grade did not significantly contribute to the prediction of third grade PSSA-M Data Analysis and Probability performance. Regression coefficients and standard errors for the full regression model are summarized in Table 20.

Table 20

Stepwise Multiple Regression Predicting Third Grade PSSA-M Data Analysis and Probability Subtest From First Grade MBSP-C Fall, Winter, and Spring Data.

Variable	*B*	*SE B*	β
Constant	6.892**	.234	
MBSP-C spring	.079	.017	.279**
MBSP-C winter	-.077	.026	-.139*
MBSP-C fall	.043	.019	.137*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics*; SE B* = Standard error of unstandardized regression coefficient. $N = 510$. * $p < .05$, ** $p < .001$

All of the β values from the independent variables included in the full regression model are statistically significant. When β values are compared, first grade MBSP-C spring scores have the largest impact on third grade PSSA-M Data Analysis and Probability scores. First grade MBSP-C winter scores have a lower, but still significant impact on the prediction of third grade PSSA-M Data Analysis and Probability scores. The relationship between MBSP-C in the spring and fall with PSSA-M Data Analysis and Probability performance is positive. This means an increase in MBSP-C spring or fall scores represents a score increase on the PSSA-M Data Analysis and Probability subtest. MBSP-C winter data have a negative β values. This is interpreted to mean that in a prediction model, an increase in MBSP-C winter performance would predict a decrease in performance on the PSSA-M Data Analysis and Probability subtest. Based on first grade data, PSSA-M Data Analysis and Probability subtest scores increased .079 points for each MBSP-C spring digit correct, .043 points for each MBSP-C fall digit correct, and decreased .08 points for each MBSP-C winter digit correct.

Second Grade Multiple Linear Regression

Second grade data from all the independent variables were entered into a stepwise linear regression for each of the PSSA-M dependent variables. The results for each dependent variable are reported in the following sections. Table 21 provides a summary of all second grade regression models with corresponding R, R^2, and adjusted R^2 values.

Table 21

Summary of Second Grade Stepwise Multiple Linear Regression Models

Dependent Variable	Model and Predictor Variables	*R*	R^2	Adjusted R^2
PSSA-M Composite	Model 1: MBSP-C winter	.509	.259	.257
	Model 2: MBSP-C winter, MBSP-C Spring	.534	.285	.283
	Model 3: MBSP-C winter, MBSP-C spring, MBSP-C fall	.548	.301	.297
	Model 4: MBSP-C winter, MBSP-C spring, MBSP-C fall, resource availability	.560	.313	.308
Numbers and Operations	Model 1: MBSP-C winter	.440	.194	.193
	Model 2: MBSP-C winter, MBSP-C Spring	.473	.223	.221
	Model 3: MBSP-C winter, MBSP-C spring, resource availability	.481	.231	.228
Measurement	Model 1: MBSP-C winter	.355	.126	.125
	Model 2: MBSP-C winter, MBSP-C Fall	.387	.150	.148
	Model 3: MBSP-C winter, MBSP-C fall, MBSP-C spring	.402	.161	.158
	Model 4: MBSP-C winter, MBSP-C fall, MBSP-C spring, sex	.408	.166	.162
	Model 5: MBSP-C winter, MBSP-C fall, MBSP-C spring, sex, resource availability	.413	.171	.165
Geometry	Model 1: MBSP-C winter	.294	.086	.085
	Model 2: MBSP-C winter, resource availability	.313	.098	.095
Algebraic Concepts	Model 1: MBSP-C winter	.388	.151	.150
	Model 2: MBSP-C winter, MBSP-C Spring	.415	.173	.170
	Model 3: MBSP-C winter, MBSP-C spring, resource availability	.427	.182	.179
Data Analysis and Probability	Model 1: MBSP-C spring	.297	.088	.087
	Model 2: MBSP-C spring, MBSP-C Fall	.311	.096	.094

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics. PSSA Composite $N = 569$, Numbers and Operations $N = 804$, Measurement $N = 787$, Geometry $N = 569$, Algebraic Concepts $N = 804$, Data Analysis and Probability $N = 789$.

Multiple linear regression of PSSA-M Composite with second grade independent variables. A stepwise multiple linear regression was completed to predict performance on the third grade PSSA-M Composite from MBSP-C probes administered in the fall, winter, and spring of second grade, sex, and resource availability. The analysis generated four regression models. The sex of second grade students did not significantly contribute to the overall variance of PSSA-M Composite in the spring of third grade. Second grade MBSP-C data in the fall, winter, and spring and resource availability statistically significantly predicted performance on the third grade PSSA-M Composite. The model is a good fit for the data, F (4, 539) = 61.468, p < .0005. The R^2 for the full regression model was 31% with an adjusted R^2 of 31%. This indicates a substantial effect size (Cohen, 1988).

The R value for the first regression model identified MBSP-C in the winter of second grade as having the strongest relationship with third grade PSSA-M Composite (R = .509). The second regression model for PSSA-M Composite included winter and spring MBSP-C data for second grade and generated an R value of .534. The third regression model added MBSP-C in the fall of second grade and increased R to .548. The fourth regression model included resource availability with R = .560. These R values indicate a strong predictive relationship between third grade PSSA-M composite scores and MBSP-C data in the fall, winter, spring and resource availability in second grade.

Based on the first regression model, second grade MBSP-C winter data accounts for the largest variance, 26% of overall PSSA-M performance administered in the spring of third grade. The second regression model included second grade MBSP-C spring data which increased the variance to 28%. The increase in the total variance accounted for is statistically significant (p <

.0005). The adjusted R^2 increased to 30% with the addition of second grade MBSP-C fall data (p = .001) and 31% (p = .002) when resource availability was added to the regression model.

Regression coefficients and standard errors for the full regression model are summarized in Table 22. The full model is reported because it accounts for the most variance and includes the independent variables which have a statistically significant impact on the dependent variable.

Table 22

Stepwise Multiple Regression Predicting Third Grade PSSA-M Composite From Second Grade Resource Availability and MBSP-C Fall, Winter, and Spring Data.

Variable	*B*	*SE B*	β
Constant	1105.438**	25.570	
MBSP-C winter	5.885	1.232	.265**
MBSP-C spring	3.645	1.023	.190**
MBSP-C fall	4.821	1.523	.145*
Resource availability	42.387	13.496	.114*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics; *SE B* = Standard error of unstandardized regression coefficient. N = 569. * $p < .05$, ** $p < .001$

All of the β values from the independent variables included in the full regression model are statistically significant. When β values were compared, second grade MBSP-C winter scores have the most significant impact on third grade PSSA-M Composite scores. Second grade MBSP-C fall and winter scores have a similar impact on third grade PSSA-M Composite scores, followed by resource availability, which had the least amount of impact on third grade PSSA-M Composite scores. The relationship between the independent variables and dependent variable is positive. This means an increase in MBSP-C scores represents a score increase on the PSSA-M Composite. The β value of resource availability, which is a dichotomous variable, is interpreted to mean student who did not receive free and reduced lunch in the second grade cohort demonstrated higher scores on the third grade PSSA-M composite. Based on second grade data, PSSA-M Composite scores increased 5.885 points for each MBSP-C winter digit correct, 3.645 points for each MBSP-C spring digit correct, and 4.821 points for each MBSP-C fall digit

correct. Students who were not receiving free or reduced lunch performed 42.387 points higher than those who were receiving free or reduced lunch.

Multiple linear regression of PSSA-M Numbers and Operations subtest with second grade independent variables. A stepwise multiple linear regression was conducted to predict performance on the third grade PSSA-M Numbers and Operations subtest from MBSP-C probes administered in the fall, winter, and spring of second grade, sex, and resource availability. The stepwise regression generated three regression models. MBSP-C in the winter, spring and resource availability of second grade were found to statistically significantly predict performance on the third grade PSSA-M Numbers and Operations subtest, F (3, 783) = 78.352, p < .0005. The R^2 for the full regression model was 23% with an adjusted R^2 of 23%. This indicates a moderate effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C in the winter of second grade accounted for the most amount of variance on third grade PSSA-M Numbers and Operations performance (R = .440; adjusted R^2 = .193). The second regression model included the MBSP-C spring data which increased total variance to 22% (R = .473; adjusted R^2 = .221). The increase in total variance accounted for with the addition of second grade MBSP-C spring data is statistically significant (p <.0005). The third regression model included resource availability in second grade as having a statistically significant relationship to third grade PSSA-M Numbers and Operations performance. Resource availability accounted for a significant increase in the total variance explained (R = .481; adjusted R^2 = .228; p = .002). Educational systems may opt to exclude resource availability from the regression model. The addition of resource availability did result in a statistically significant increase to the adjusted R^2. However, the increase in

variance may not be worth the time and effort for educational systems to obtain resource availability data.

An *R* value of .481 indicates a moderately strong relationship between the full regression model and performance on the Numbers and Operations subtest on the third grade PSSA-M. The sex of second grade students and MBSP-C in the fall of second grade did not significantly predict performance on third grade PSSA-M Numbers and Operations subtest or significantly contribute to the variance. Regression coefficients and standard errors for the full regression model are summarized in Table 23.

Table 23

Stepwise Multiple Regression Predicting Third Grade PSSA-M Numbers and Operations Subtest From Second Grade Resource Availability and MBSP-C Winter and Spring Data.

Variable	*B*	*SE B*	β
Constant	21.015**	.747	
MBSP-C winter	.185	.034	.251**
MBSP-C spring	.158	.030	.241**
Resource availability	1.088	.391	.088*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics *SE B* = Standard error of unstandardized regression coefficient. $N = 821$. * $p < .05$, ** $p < .001$

All of the β values from the independent variables included in the full regression model are statistically significant. A comparison of β values indicate second grade MBSP-C winter scores have the most significant impact on third grade PSSA-M Numbers and Operations subtest scores, with a β value of .251. Second grade MBSP-C spring scores have a similar, but slightly less, impact on third grade PSSA-M Numbers and Operations subtest scores (β = .241). When compared to other variables included in the prediction model, resource availability had the least amount of impact on the third grade Numbers and Operations subtest. The relationship between the independent variables and dependent variable is positive. This means an increase in MBSP-C scores represents a score increase on the PSSA-M Numbers and Operations subtest. The β

value of resource availability, which is a dichotomous variable, is interpreted to mean students who did not receive free and reduced lunch in the second grade cohort demonstrated higher scores on the third grade PSSA-M Numbers and Operations subtest. Based on second grade data, PSSA-M Number and Operations subtest scores increased .185 points for each MBSP-C winter digit correct and .158 points for each MBSP-C spring digit correct. Students who were not receiving free or reduced lunch performed 1.088 points higher than those who were receiving free or reduced lunch.

Multiple linear regression of PSSA-M Measurement subtest with second grade independent variables. A stepwise multiple linear regression was performed to predict performance on the third grade PSSA-M Measurement subtest from MBSP-C probes administered in the fall, winter, and spring of second grade, sex, and resource availability. The stepwise regression created five models. This suggests that all five second grade variables are statistically significantly related to performance on the third grade PSSA-M Measurement subtest, F (5, 781) = 32.162, p < .0005. The R^2 for the full regression model was 17% with an adjusted R^2 of 17%. This indicates a moderate effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C in the winter of second grade accounted for the most amount of variance on third grade PSSA-M Measurement performance (R = .355; adjusted R^2 = .125). The second regression model included MBSP-C in the fall of second grade which increased total variance explained to 15% (R = .387; adjusted R^2 = .148). The increase in total variance explained with the addition of MBSP-C fall data to the regression model was statistically significant (p <.0005). The third independent variable included in the regression model was spring MBSP-C data (R = .402; adjusted R^2 = .158) followed by sex (R = .408; adjusted R^2 = .162) and resource availability (R = .413; adjusted R^2 = .165). All increases

in variance are statistically significant ($p = .001$; $p = .032$; $p = .039$). However, educational systems may choose to exclude sex and resource availability data without significantly impacting the adjusted R^2. Regression coefficients and standard errors for the full regression model are summarized in Table 24.

Table 24

Stepwise Multiple Regression Predicting Third Grade PSSA-M Measurement Subtest From Second Grade Resource Availability, Sex, and MBSP-C Fall, Winter, and Spring Data.

Variable	*B*	*SE B*	β
Constant	5.398**	.325	
MBSP-C winter	.034	.013	.135**
MBSP-C fall	.056	.016	.149**
MBSP-C spring	.035	.011	.157**
Sex	.285	.128	.073*
Resource availability	.285	.138	.069*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics; *SE B* = Standard error of unstandardized regression coefficient. $N = 821$. * $p < .05$, ** $p < .001$

All of the β values from the independent variables included in the full regression model are statistically significant. A comparison of β values indicate second grade MBSP-C spring scores have the most significant impact on third grade PSSA-M Measurement subtest scores. Second grade MBSP-C fall and winter scores have a similar impact on third grade PSSA-M Measurement subtest scores, followed by sex and resource availability. The relationship between the independent variables and dependent variable is positive. This means an increase in MBSP-C scores represents a score increase on the PSSA-M Measurement subtest. The β value of resource availability, which is a dichotomous variable, is interpreted to mean students who did not receive free and reduced lunch in the second grade cohort demonstrated higher scores on the third grade PSSA-M Measurement subtest. The β value of sex, also a dichotomous variable, indicates males out performed females on the PSSA-M Measurement subtest. Based on second grade data, PSSA-M Measurement subtest scores increased .034 points for each MBSP-C winter

digit correct, .056 points for each MBSP-C fall digit correct, and .035 points for each MBSP-C spring digit correct. Students who were not receiving free or reduced lunch performed .285 points higher than those who were receiving free or reduced lunch. Males performed .285 higher than females on the PSSA-M Measurement subtest.

Multiple linear regression of PSSA-M Geometry subtest with second grade independent variables. A stepwise multiple linear regression was conducted to predict performance on the third grade PSSA-M Geometry subtest from MBSP-C probes administered in the fall, winter, and spring of second grade, sex, and resource availability. The stepwise regression generated two regression models, both having a good fit to the data. MBSP-C in the winter of second grade and resource availability were found to statistically significantly predict performance on the third grade PSSA-M Geometry subtest, F (2, 541) = 29.439, $p < .0005$. The R^2 for the full regression model was 10% with an adjusted R^2 of 10%. This indicates a weak effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C in the winter of second grade accounted for the most amount of variance on third grade PSSA-M Geometry performance (R = .294; adjusted R^2 = .085). The second regression model included resource availability of second grade students. The addition of resource availability to the regression model increased total variance explained to 10% (R = .313; adjusted R^2 = .095). The increase in total variance explained when resource availability was added to the regression model was statistically significant (p = .008). The full regression model has a moderate relationship to the PSSA-M Geometry subtest. MBSP-C in the fall and spring and sex of students in second grade did not significantly predict performance on third grade PSSA-M Geometry subtest. Regression coefficients and standard errors for the full regression model are summarized in Table 25.

Table 25

Stepwise Multiple Regression Predicting Third Grade PSSA-M Geometry Subtest From Second Grade MBSP-C Winter Data and Resource Availability.

Variable	*B*	*SE B*	β
Constant	7.470**	.207	
MBSP-C winter	.044	.007	.277**
Resource availability	.295	.110	.110*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics; *SE B* = Standard error of unstandardized regression coefficient. $N = 569$. * $p < .05$, ** $p < .001$

All of the β values from the independent variables included in the full regression model are statistically significant. The β values indicate second grade MBSP-C winter scores have the most significant impact on third grade PSSA-M Geometry subtest scores. The independent variable and dependent variable have a positive relationship. This means that in a prediction model, an increase in MBSP-C winter scores results in an increase on the PSSA-M Geometry subtest. The β value of resource availability, which is a dichotomous variable, is interpreted to mean students who did not receive free and reduced lunch in the second grade cohort demonstrated higher scores on the third grade PSSA-M Geometry subtest. Based on second grade data, PSSA-M Geometry subtest scores increased .044 points for each MBSP-C winter digit correct. Students who were not receiving free or reduced lunch performed .295 points higher than those who were receiving free or reduced lunch.

Multiple linear regression of PSSA-M Algebraic Concepts subtest with second grade independent variables. A stepwise multiple linear regression was performed to predict performance on the third grade PSSA-M Algebraic Concepts subtest from MBSP-C probes administered in the fall, winter, and spring of second grade, sex, and resource availability. The stepwise regression generated three regression models. MBSP-C in the winter and spring and resource availability of second grade were found to statistically significantly predict performance on the third grade PSSA-M Algebraic Concepts subtest, $F(3, 783) = 58.040$, $p < .0005$. The R^2

for the full regression model was 18% with an adjusted R^2 of 18%. This indicates a moderate effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C in the winter of second grade accounted for the most amount of variance on third grade PSSA-M Algebraic Concepts performance (R = .388; adjusted R^2 = .150). The second regression model entered MBSP-C in the spring of second grade, this increased total variance of PSSA-M Algebraic Concepts explained by MBSP-C in the winter and spring of second grade to 17% (R = .415; adjusted R^2 = .170). This increase is statistically significant ($p < .0005$.) The third regression model, also statistically significant (p = .003), included resource availability (R = .427; adjusted R^2 = .179). The R values indicate a moderately strong predictive relationship between the independent variables in the full regression model and performance on the Algebraic Concepts subtest of the third grade PSSA-M. MBSP-C in the fall of second grade and sex of student did not significantly contribute to the variance or predict performance on third grade PSSA-M Algebraic Concepts. Regression coefficients and standard errors for the full regression model are summarized in Table 26.

Table 26

Stepwise Multiple Regression Predicting Third Grade PSSA-M Algebraic Concepts Subtest From Second Grade Resource Availability and MBSP-C Winter and Spring Data.

Variable	*B*	*SE B*	β
Constant	6.445**	.218	
MBSP-C winter	.047	.010	.224**
MBSP-C spring	.038	.009	.205**
Resource availability	.342	.114	.098*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics; *SE B* = Standard error of unstandardized regression coefficient. N = 821. * $p < .05$, ** $p < .001$

All of the β values from the independent variables included in the full regression model are statistically significant. The β values indicate second grade MBSP-C winter scores have the

most significant impact on third grade PSSA-M Algebraic Concepts subtest scores. MBSP-C spring scores also have a significant, but smaller impact on PSSA-M Algebraic Concepts scores. The β value of resource availability, which is a dichotomous variable, is interpreted to mean students who did not receive free and reduced lunch in the second grade cohort demonstrated higher scores on the third grade PSSA-M Algebraic Concepts subtest. Resource availability had the least amount of impact on third grade PSSA-M Algebraic outcomes (β = .098). The independent variables and dependent variable have a positive relationship. This means that in a prediction model, an increase in MBSP-C winter or spring scores results in an increase on the PSSA-M Algebraic Concepts subtest. Based on second grade data, PSSA-M Algebraic Concepts subtest scores increased .047 points for each MBSP-C winter digit correct and .038 points for each MBSP-C spring digit correct. Students who were not receiving free or reduced lunch performed .342 points higher on the PSSA-M Algebraic Concepts subtest than those who were receiving free or reduced lunch.

Multiple linear regression of PSSA-M Data Analysis and Probability subtest with second grade independent variables. A stepwise multiple linear regression was conducted to predict performance on the third grade PSSA-M Data Analysis and Probability subtest from MBSP-C probes administered in the fall, winter, and spring of second grade, sex, and resource availability. The stepwise regression generated two models. MBSP-C in the fall and spring of second grade were found to statistically significantly predict performance on the third grade PSSA-M Data Analysis and Probability subtest, F (2, 784) = 41.836, p < .0005. The R^2 for the full regression model was 10% with an adjusted R^2 of 9%. This indicates a weak effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C in the spring of second grade accounted for the most amount of variance on third grade PSSA-M Data Analysis and Probability performance (R = .297; adjusted R^2 = .087). The second regression model included the MBSP-C in the fall of second grade which increased total variance of PSSA-M Data Analysis and Probability that can be accounted for by MBSP-C in the fall and spring of second grade to 9% (R = .311; adjusted R^2 = .094). The increase in adjusted R^2 is statistically significant (p = .009). MBSP-C in the fall and spring were moderately predictive of student performance the Data Analysis and Probability subtest of third grade PSSA-M. MBSP-C winter of second grade, resource availability and student sex did not significantly contribute to the variance or predict performance on third grade PSSA-M Data Analysis and Probability. Regression coefficients and standard errors for the full regression model are summarized in Table 27.

Table 27

Stepwise Multiple Regression Predicting Third Grade PSSA-M Data Analysis and Probability Subtest From Second Grade MBSP-C Fall and Spring Data.

Variable	*B*	*SE B*	β
Constant	7.499**	.147	
MBSP-C spring	.045	.008	.235**
MBSP-C fall	.035	.013	.109*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics; *SE B* = Standard error of unstandardized regression coefficient. N = 821. * $p < .05$, ** $p < .001$

The β values from the independent variables included in the full regression model are statistically significant. A comparison of β values indicate second grade MBSP-C spring scores have the most significant impact on third grade PSSA-M Data Analysis and Probability subtest scores. MBSP-C fall scores also have a significant impact on PSSA-M Data Analysis and Probability scores, but less than half the impact of MBSP-C spring scores. The independent variables and dependent variable have a positive relationship. This means that in a prediction model, an increase in MBSP-C spring or fall scores results in an increase on the PSSA-M Data

Analysis and Probability subtest. Based on second grade data, PSSA-M Data Analysis and Probability subtest scores increased .045 points for each MBSP-C spring digit correct and .035 points for each MBSP-C fall digit correct.

Third Grade Multiple Linear Regression

Third grade data from all the independent variables were entered into a stepwise linear regression for each of the PSSA-M dependent variables. The results for each dependent variable are reported in the following sections. Please refer to Table 28 for a summary of each regression model with corresponding R, R^2, and adjusted R^2 values.

Table 28

Summary of Third Grade Stepwise Regression Models

Dependent Variable	Model and Predictor Variables	*R*	R^2	Adjusted R^2
PSSA-M Composite	Model 1: MBSP-C winter	.569	.324	.323
	Model 2: MBSP-C winter, MBSP-C spring	.613	.375	.374
	Model 3: MBSP-C winter, MBSP-C spring, MBSP-C fall	.623	.388	.386
	Model 4: MBSP-C winter, MBSP-C spring, MBSP-C fall, resource availability	.630	.397	.394
Numbers and Operations	Model 1: MBSP-C spring	.535	.286	.286
	Model 2: MBSP-C spring, MBSP-C winter	.580	.336	.335
	Model 3: MBSP-C spring, MBSP-C winter, sex	.585	.342	.340
	Model 4: MBSP-C spring, MBSP-C winter, sex, MBSP-C fall	.589	.347	.345
	Model 5: MBSP-C spring, MBSP-C winter, sex, MBSP-C, resource availability	.592	.351	.348
Measurement	Model 1: MBSP-C spring	.387	.150	.149
	Model 2: MBSP-C spring, MBSP-C winter	.418	.175	.173
	Model 3: MBSP-C spring, MBSP-C winter, sex	.427	.182	.180
	Model 4: MBSP-C spring, MBSP-C winter, sex, resource availability	.432	.187	.184
	Model 5: MBSP-C spring, MBSP-C winter, sex, resource availability, MBSP-C fall	.436	.190	.186
Geometry	Model 1: MBSP-C spring	.349	.122	.121
	Model 2: MBSP-C spring, resource availability	.361	.130	.128
	Model 3: MBSP-C spring, resource availability, MBSP-C winter	.369	.136	.133
Algebraic Concepts	Model 1: MBSP-C winter	.413	.171	.170
	Model 2: MBSP-C winter, MBSP-C Spring	.430	.185	.184

Table 28 *Summary of Third Grade Stepwise Regression Models*

Dependent Variable	Model and Predictor Variables	*R*	R^2	Adjusted R^2
	Model 3: MBSP-C winter, MBSP-C spring, resource availability	.436	.190	.188
	Model 4: MBSP-C winter, MBSP-C spring, resource availability, sex	.439	.193	.190
Data Analysis and Probability	Model 1: MBSP-C spring	.369	.136	.136
	Model 2: MBSP-C spring, MBSP-C Winter	.396	.157	.155
	Model 3: MBSP-C spring, MBSP-C winter, MBSP-C fall	.399	.159	.157

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics. PSSA Composite $N = 886$, Numbers and Operations $N = 1144$, Measurement $N = 1144$, Geometry $N = 886$, Algebraic Concepts $N = 1177$, Data Analysis and Probability $N = 1144$.

Multiple linear regression of PSSA-M Composite with third grade independent variables. A stepwise multiple linear regression was conducted to predict performance on the third grade PSSA-M Composite from MBSP-C probes administered in the fall, winter, and spring of third grade, sex, and resource availability. The analysis generated four regression models. Third grade MBSP-C data in the fall, winter, and spring and resource availability statistically significantly predicted performance on the third grade PSSA-M Composite, F (4, 881) = 145.083, $p < .0005$ for the full regression model. The R^2 for the full regression model was 40% with an adjusted R^2 of 39%. This indicates a substantial effect size (Cohen, 1988).

The R value for the first regression model identified MBSP-C in the winter of third grade as having the strongest relationship with third grade PSSA-M Composite ($R = .569$). The second regression model for PSSA-M Composite included winter and spring MBSP-C data for third grade and generated an R value of .613. The third regression model added MBSP-C in the fall of third grade and increased R to .623. The fourth regression model included resource availability with $R = .630$. These R values indicate a strong predictive relationship.

Based on the first regression model, third grade MBSP-C winter data accounts for the largest amount of variance, 32% on overall PSSA-M performance, administered in the winter of third grade. The second regression model included third grade MBSP-C spring data which increased the amount of explained variance to 37%. The increase in variance from the first to second regression model is statistically significant ($p < .0005$). The adjusted R^2 increased to 39% with the addition of third grade MBSP-C fall data and remained at 39% when resource availability was added to the regression model. Given the relatively small increase in variance explained when resource availability was added to the regression model, it would be acceptable to exclude it from the prediction model. The sex of third grade students did not significantly contribute to the overall variance of PSSA-M Composite in the spring of third grade.

Regression coefficients and standard errors for the full regression model are summarized in Table 29. The full model is reported because it accounts for the most variance and includes the independent variables which have a statistically significant impact on the dependent variable.

Table 29

Stepwise Multiple Regression Predicting PSSA-M Composite From Third Grade Resource Availability and MBSP-C Fall, Winter, and Spring Data.

Variable	*B*	*SE B*	β
Constant	1022.820**	20.478	
MBSP-C winter	5.904	.853	.275**
MBSP-C spring	5.151	.690	.274**
MBSP-C fall	3.786	.873	.145**
Resource availability	34.185	9.58	.095**

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics; *SE B* = Standard error of unstandardized regression coefficient. $N = 886$. * $p < .05$, ** $p < .001$

The β value for each independent variables included in the full regression model were statistically significant. A comparison of β values indicate third grade MBSP-C winter scores have the most significant impact on third grade PSSA-M Composite scores. It should be noted, third grade MBSP-C spring scores have a nearly identical impact on third grade PSSA-M

Composite scores, with only a .001 difference in β values. MBSP-C in the fall has a statistically significant, but smaller, impact on PSSA-M Composite scores. The impact of resource availability was also statistically significant, but had the least impact on third grade PSSA-M outcomes. The relationship between the independent variables and dependent variable is positive. This means that in the prediction model, an increase in MBSP-C scores results in an increase on PSSA-M Composite scores. The β value of resource availability, which is a dichotomous variable, is interpreted to mean students who received free or reduced lunch performed lower on the third grade PSSA-M composite than those who did not. Based on third grade data, PSSA-M Composite scores increased 5.904 points for each MBSP-C winter digit correct, 5.151 points for each MBSP-C spring digit correct, and 3.786 points for each MBSP-C fall digit correct. Students who were not receiving free or reduced lunch performed 34.185 points higher than those who were receiving free or reduced lunch.

Multiple linear regression of PSSA-M Numbers and Operations subtest with third grade independent variables. A stepwise multiple linear regression was performed to predict performance on the third grade PSSA-M Numbers and Operations subtest from MBSP-C probes administered in the fall, winter, and spring of third grade, sex, and resource availability. The stepwise regression generated five regression models. MBSP-C in the winter, spring and fall in addition to, sex and resource availability of third grade were found to statistically significantly predict performance on the third grade PSSA-M Numbers and Operations subtest, F (5, 1138) = 122.924, $p < .0005$. The R^2 for the full regression model was 35% with an adjusted R^2 of 35%. This indicates a substantial effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C in the spring of third grade accounted for the most amount of variance on third grade PSSA-M Numbers and Operations

performance (R = .535; adjusted R^2 = .286). The second regression model entered MBSP-C winter data which increased total variance to 34% (R = .580; adjusted R^2 = .335). This increase is statistically significant ($p < .0005$). The third regression model indicated sex had a statistically significant relationship to third grade PSSA-M Numbers and Operations (R = .585; adjusted R^2 = .340). The increase in variance when sex was added to the regression model was statistically significant (p = .002). The fourth regression model entered MBSP-C fall data (R = .589; adjusted R^2 = .345). The fifth regression model identified resource availability as having statistically significant contribution to the overall variance (R = .592; adjusted R^2 = .348). The increase in variance explained from the fourth to fifth regression model is statistically significant (p = .011). However, the increase is variance (.003) may not be relevant from an application standpoint. An R value of .589 signifies a strong relationship between the full regression model and student performance on the Numbers and Operations subtest.

Regression coefficients and standard errors for the full regression model are summarized in Table 30. The full model is reported because it accounts for the most variance and includes the independent variables which have a statistically significant impact on the dependent variable.

Table 30

Stepwise Multiple Regression Predicting PSSA-M Numbers and Operations Subtest From Third Grade Resource Availability, Sex, and MBSP-C Fall, Winter, and Spring Data.

Variable	*B*	*SE B*	β
Constant	16.735**	.716	
MBSP-C spring	.178	.020	.302**
MBSP-C winter	.177	.023	.267**
Sex	.790	.254	.074*
MBSP-C fall	.069	.024	.084*
Resource availability	.694	.274	.062*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics; *SE B* = Standard error of unstandardized regression coefficient. N = 1143. * $p < .05$, ** $p < .001$

The β values for each independent variables included in the full regression model are statistically significant. A comparison of β values indicate third grade MBSP-C spring scores have the most significant impact on third grade PSSA-M Numbers and Operations subtest scores. Third grade MBSP-C winter scores have the second most significant impact on third grade PSSA-M Numbers and Operations subtest scores. MBSP-C fall scores have less of an impact on PSSA-M Numbers and Operations scores when compared to MBSP-C in the spring and winter. Sex and resource availability have the least amount of influence on third grade Numbers and Operations outcomes. The relationship between the independent variables and dependent variable is positive. This means that in the prediction model, an increase in MBSP-C scores represents a score increase on the PSSA-M Numbers and Operations subtest. The β value of sex, which is a dichotomous variable, is interpreted to mean males in the third grade cohort demonstrated higher scores on the third grade PSSA-M Numbers and Operations subtest. The β value of resource availability, also a dichotomous variable, is interpreted to mean students who received free or reduced lunch performed lower on the third grade PSSA-M Numbers and Operations subtest than those who did not. Based on third grade data, PSSA-M Numbers and Operations subtest scores increased .178 points for each MBSP-C spring digit correct, .177 points for each MBSP-C winter digit correct, and .069 points for each MBSP-C fall digit correct. Students who were not receiving free or reduced lunch performed .694 points higher on the PSSA-M Numbers and Operations subtest than those who were receiving free or reduced lunch. Males performed .790 points higher than females.

Multiple linear regression of PSSA-M Measurement subtest with third grade independent variables. A stepwise multiple linear regression was performed to predict performance on the third grade PSSA-M Measurement subtest from MBSP-C probes

administered in the fall, winter, and spring of third grade, sex, and resource availability. The stepwise regression created five models. This suggests that all five variables are statistically significantly related to performance on the third grade PSSA-M Measurement subtest with a moderately strong relationship, F (5, 1138) = 53.363, p < .0005. The R^2 for the full regression model was 19% with an adjusted R^2 of 19%. This indicates a moderate effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C in the spring of third grade accounted for the most amount of variance on PSSA-M Measurement performance (R = .387; adjusted R^2 = .149). The second regression model included MBSP-C in the winter of third grade which increased total variance to 17% (R = .418; adjusted R^2 = .173). The increase in adjusted R^2 from the first regression model is statistically significant (p < .0005). The third independent variable included in the regression model was sex of student (R = .427; adjusted R^2 = .180). The fourth regression model included MBSP-C fall data (R = .432; adjusted R^2 = .184). The fifth regression model identified resource availability as having a statistically significant contribution to the overall variance explained (R = .436; adjusted R^2 = .186). The increase in total variance explained from the third to the fifth regression model is 0.6%. For practical purposes the increase in variance, while statistically significant, may not be worth the additional time and effort required to collect additional data. It is likely MBSP-C fall data will be readily available from normal universal screening procedures. However, sex and resource availability can be excluded without significantly decreasing the amount of variance explained. Regression coefficients and standard errors for the full regression model are summarized in Table 31.

Table 31

Stepwise Multiple Regression Predicting PSSA-M Measurement Subtest From Third Grade Resource Availability, Sex, and MBSP-C Fall, Winter, and Spring Data.

Variable	*B*	*SE B*	β
Constant	4.658**	.295	
MBSP-C spring	.048	.008	.219**
MBSP-C winter	.044	.010	.182**
Sex	.329	.105	.084*
Resource availability	.285	.113	.069*
MBSP-C fall	.021	.010	.069*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics; *SE B* = Standard error of unstandardized regression coefficient. $N = 1143$. * $p < .05$, ** $p < .001$

The β values for the independent variables included in the full regression model are statistically significant. A comparison of β values indicate third grade MBSP-C spring scores have the most significant impact on third grade PSSA-M Measurement subtest scores. Third grade MBSP-C winter scores have the second most significant impact on third grade PSSA-M Measurement subtest scores. MBSP-C fall scores (β = .069) and resource availability (β = .069) have much less impact on PSSA-M Measurement scores than MBSP-C in the spring and winter. In fact, in the prediction model, sex has more influence on third grade PSSA-M Measurement outcomes (β = .084) than MBSP-C in the fall and resource availability. The relationship between the independent variables and dependent variable is positive. This means that in the prediction model, an increase in MBSP-C scores results in increased PSSA-M Measurement subtest scores. The β value of sex, which is a dichotomous variable, is interpreted to mean males in the third grade cohort demonstrated higher scores on the third grade PSSA-M Measurement subtest. The β value of resource availability, also a dichotomous variable, is interpreted to mean students who received free or reduced lunch performed lower on the third grade PSSA-M Measurement subtest than those who did not. Based on third grade data, PSSA-M Measurement subtest scores increased .048 points for each MBSP-C spring digit correct, .044 points for each MBSP-C winter

digit correct, and .021 points for each MBSP-C fall digit correct. Students who were not receiving free or reduced lunch performed .285 points higher on the PSSA-M Measurement subtest than those who were receiving free or reduced lunch. Males performed .329 points higher than females.

Multiple linear regression of PSSA-M Geometry subtest with third grade independent variables. A stepwise multiple linear regression was performed to predict performance on the third grade PSSA-M Geometry subtest from MBSP-C probes administered in the fall, winter, and spring, sex, and resource availability. The stepwise regression generated three regression models. MBSP-C in the spring and winter of third grade and resource availability were found to statistically significantly predict performance on the third grade PSSA-M Geometry subtest, F (3, 882) = 46.200, $p < .0005$. The R^2 for the full regression model was 14% with an adjusted R^2 of 13%. This indicates a moderate effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C in the spring of third grade accounted for the most amount of variance on third grade PSSA-M Geometry performance (R = .349; adjusted R^2 = .121). The second regression model added resource availability which increased total variance explained to 13% (R = .361; adjusted R^2 = .128). The third regression modeled identified MBSP-C winter as a significant contributor to the variance of PSSA-M Geometry performance, with a moderately strong predictive relationship (R = .369; adjusted R^2 = .133). While the increase in adjusted R^2 is statistically significant, it may not be necessary to include both MBSP-C winter and resource availability in a prediction model. From an application standpoint, MBSP-C winter data will likely be available. It may not be worth the additional time and effort to collect resource availability data given the overall small increase in variance when it was added in the second regression model.

MBSP-C in the fall and sex of students in third grade did not significantly predict performance on third grade PSSA-M Geometry subtest or significantly contribute to the variance. Regression coefficients and standard errors for the full regression model are summarized in Table 32.

Table 32

Stepwise Multiple Regression Predicting PSSA-M Geometry Subtest From Third Grade Resource Availability and MBSP-C Winter and Spring Data.

Variable	*B*	*SE B*	β
Constant	7.079**	.178	
MBSP-C spring	.036	.006	.260**
Resource availability	.240	.084	.092*
MBSP-C winter	.016	.007	.103*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics; *SE B* = Standard error of unstandardized regression coefficient. $N = 886$. * $p < .05$, ** $p < .001$

The β value of the independent variables included in the full regression model are statistically significant. A comparison of β values indicate third grade MBSP-C spring scores had the most significant impact on third grade PSSA-M Geometry subtest scores. Third grade MBSP-C winter scores had the second most significant impact on third grade PSSA-M Geometry subtest scores. MBSP-C fall scores had a statistically significant ($p = .018$) but much less impact on PSSA-M Geometry scores than MBSP-C in the spring. Resource availability has the least amount of influence on third grade PSSA-M geometry scores with a β value of .092. The relationship between the independent variables and dependent variable is positive. This means that in the prediction model, an increase in MBSP-C scores results in a score increase on the PSSA-M Geometry subtest. The β value of resource availability, a dichotomous variable, is interpreted to mean students who received free or reduced lunch performed lower on the third grade PSSA-M Geometry subtest than those who did not. Based on third grade data, PSSA-M Geometry subtest scores increased .036 points for each MBSP-C spring digit correct and .016

points for each MBSP-C winter digit correct. Students who were not receiving free or reduced lunch performed .240 points higher on the PSSA-M Geometry subtest than those who were receiving free or reduced lunch.

Multiple linear regression of PSSA-M Algebraic Concepts subtest with third grade independent variables. A stepwise multiple linear regression was conducted to predict performance on the third grade PSSA-M Algebraic Concepts subtest from MBSP-C probes administered in the fall, winter, and spring of third grade, sex, and resource availability. The stepwise regression generated four regression models. MBSP-C in the winter and spring, resource availability, and sex of student in third grade had a statistically significant predictive relationship with performance on the third grade PSSA-M Algebraic Concepts subtest, F (4, 1139) = 67.912, $p < .0005$. The R^2 for the full regression model was 19% with an adjusted R^2 of 19%. This indicates a moderate effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C in the winter data accounted for the most amount of variance on third grade PSSA-M Algebraic Concepts performance (R = .413; adjusted R^2 = .170). The second regression model entered MBSP-C spring data which increased total variance of PSSA-M Algebraic Concepts that can be accounted for by MBSP-C in the winter and spring of third grade to 18% (R = .430; adjusted R^2 = .184). The third regression model added resource availability (R = .436; adjusted R^2 = .188). Sex was identified as a significant additional contributor to variance in the fourth regression model (R = .439; adjusted R^2 = .190). While the increase in adjusted R^2 is statistically significant, it may not be necessary to include sex and resource availability in the prediction model. From an application standpoint, it is difficult to justify the additional data collection given the small contribution to overall variance.

The full regression model has a moderately strong predictive relationship to third grade performance on the Algebraic Concepts subtest. MBSP-C in the fall of third grade did not significantly contribute to the variance or predict performance on third grade PSSA-M Algebraic Concepts. Regression coefficients and standard errors for the full regression model are summarized in Table 33.

Table 33

Stepwise Multiple Regression Predicting PSSA-M Algebraic Concepts Subtest From Third Grade Resource Availability, Sex, and MBSP-C Winter and Spring Data.

Variable	*B*	*SE B*	β
Constant	5.129**	.265	
MBSP-C winter	.066	.008	.300**
MBSP-C spring	.030	.007	.153**
Resource availability	.260	.102	.070*
Sex	.190	.094	.054*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics; *SE B* = Standard error of unstandardized regression coefficient. $N = 1143$. * $p < .05$, ** $p < .001$

Each independent variable included in the full regression model generated statistically significant β values. A comparison of β values indicated third grade MBSP-C winter scores had the most significant impact on third grade PSSA-M Algebraic Concepts subtest scores. Third grade MBSP-C spring scores had almost half the impact of MBSP-C winter scores on third grade PSSA-M Algebraic Concepts scores. Resource availability had a small, but statistically significant impact on third grade PSSA-M Algebraic Concepts outcomes. In the prediction model, sex had the least amount of influence on third grade PSSA-M Algebraic Concepts outcomes. The relationship between the independent variables and dependent variable is positive. This means that in the prediction model, an increase in MBSP-C scores results in an increase on PSSA-M Algebraic Concepts subtests scores. The β value of resource availability, which is a dichotomous variable, is interpreted to mean students receiving free or reduced lunch performed lower on the third grade PSSA-M Algebraic Concepts than those who did not. The β

value of sex, also a dichotomous variable, is interpreted to mean males in the third grade cohort demonstrated higher scores on the third grade PSSA-M Algebraic Concepts subtest than females. Based on third grade data, PSSA-M Algebraic Concepts subtest scores increased .066 points for each MBSP-C winter digit correct and .030 points for each MBSP-C spring digit correct. Students who were not receiving free or reduced lunch performed .260 points higher on the PSSA-M Algebraic Concepts subtest than those who were receiving free or reduced lunch. Males performed .190 points higher than females.

Multiple linear regression of PSSA-M Data Analysis and Probability subtest with third grade independent variables. A stepwise multiple linear regression was conducted to predict performance on the third grade PSSA-M Data Analysis and Probability subtest from MBSP-C probes administered in the fall, winter, and spring of third grade, sex, and resource availability. The stepwise regression generated three regression models. MBSP-C in the fall, winter, and spring were found to statistically significantly predict performance on the PSSA-M Data Analysis and Probability subtest with a moderately strong relationship, F (3, 1140) = 72.106, $p < .0005$. The R^2 for the full regression model was 16% with an adjusted R^2 of 16%. This indicates a moderate effect size (Cohen, 1988).

The outcome of the stepwise regression indicated MBSP-C spring data accounted for the most amount of variance on third grade PSSA-M Data Analysis and Probability performance (R = .369; adjusted R^2 = .136). The second regression model included MBSP-C winter data. This increased total variance on PSSA-M Data Analysis and Probability performance explained by MBSP-C in the winter and spring to 16% (R = .396; adjusted R^2 = .155). MBSP-C fall data were added into the third regression model and resulted in a statistically significant contribution to the prediction model (R = .399; adjusted R^2 = .157). The increase in variance from the second to

third regression model with the addition of MBSP-C fall is statistically significant but results in only a small increase to the total variance explained by the model.

Resource availability and student sex did not significantly contribute to the variance or predict performance on third grade PSSA-M Data Analysis and Probability. Regression coefficients and standard errors for the full regression model are summarized in Table 34.

Table 34

Stepwise Multiple Regression Predicting PSSA-M Data Analysis and Probability Subtest From Third Grade MBSP-C Fall, Winter, and Spring Data.

Variable	*B*	*SE B*	β
Constant	6.611**	.154	
MBSP-C spring	.039	.007	.224**
MBSP-C winter	.032	.008	.165**
MBSP-C fall	.016	.008	.066*

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics; *SE B* = Standard error of unstandardized regression coefficient. $N = 1143$. * $p < .05$, ** $p < .001$

Each independent variable included in the full regression model generated statistically significant β values. A comparison of the β values indicate third grade MBSP-C spring scores have the most significant impact on third grade PSSA-M Data Analysis and Probability subtest scores. Third grade MBSP-C winter scores have the second most significant impact on third grade PSSA-M Data Analysis and Probability subtest scores. MBSP-C fall scores had the least amount of impact on PSSA-M Data Analysis and Probability scores. The relationship between the independent variables and dependent variable is positive. This means that in the prediction model, an increase in MBSP-C scores results in a score increase on the PSSA-M Data Analysis and Probability subtest. Based on third grade data, PSSA-M Data Analysis and Probability subtest scores increased .039 points for each MBSP-C spring digit correct, .032 points for each MBSP-C winter digit correct, and .016 points for each MBSP-C fall digit correct.

Pearson Correlations for Independent and Dependent Variables

MLR analysis was used to explore the broad research question of the present study: To what extent does a universal mathematics screening, MBSP-C in first, second, and third grade, sex, and SES predict math achievement as reported on the five subtests of the PSSA-M in third grade? However, of the several hypotheses generated by this broad research question required correlations between the dependent and independent variables at specific points in times. These hypotheses are as follows, (a) it is hypothesized that student performance in the fall of first grade will have the weakest correlation with PSSA performance and student performance in the spring of third grade will have the strongest correlation with third grade PSSA-M achievement due to time proximity between MBSP-C and PSSA-M administration, (b) it is hypothesized that MBSP-C will have the strongest correlation with the Numbers and Operations subtest of the PSSA-M, (c) it is hypothesized that resource availability will account for a significant amount of variance on PSSA-M achievement, with the potential to decrease the longer students are in a high quality educational setting, and (d) it is hypothesized that sex and resource availability will have a moderate association with math achievement, based on highlights from the 2007 TIMSS (Gonzales et al., 2009).

The relationship between the independent and dependent variables were analyzed with Pearson correlations. Pearson correlations were generated to determine the strength of the relationship between each independent variable with PSSA-M scores. The strength of the relationship between sex, resource availability and MBSP-C in the fall, winter, and spring of first, second, and third grade were also examined. Please refer to Tables 35 and 36 for a summary of the Pearson correlation coefficients (*r)*. An *r* value of greater than 0.1 to less than 0.3 indicates a weak correlation. Correlation coefficients ranging from greater than 0.3 to less

than 0.5 indicate a moderate correlation. Correlation coefficients greater than 0.5 indicate a strong correlation. The relationship between variables can be positive and negative in nature (Fields, 2013; Laerd Statistics, 2015).

Table 35

Pearson Correlations for MBSP-C Fall, Winter, and Spring of First, Second, and Third Grade with PSSA-M Scores

	First Grade MBSP-C			Second Grade MBSP-C			Third Grade MBSP-C		
	Fall	Winter	Spring	Fall	Winter	Spring	Fall	Winter	Spring
PSSA-M Composite	.388**	.501**	.522**	.431**	.514**	.489**	.469**	.572**	.561**
Numbers and Operations	.204**	.420**	.501**	.352**	.450**	.451**	.399**	.536**	.546**
Measurement	.240**	.339**	.406**	.343**	.374**	.380**	.295**	.392**	.406**
Geometry	.309**	.418**	.396**	.190**	.287**	.259**	.190**	.297**	.349**
Algebraic Concepts	.272**	.387**	.416**	.314**	.402**	.403**	.301**	.418**	.379**
Data Analysis and Probability	.023	.277**	.324**	.245**	.281**	.306**	.273**	.355**	.379**

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics. First Grade Cohort PSSA-M Composite and Geometry subtest MBSP-C Fall $N = 266$, MBSP-C Winter $N = 269$, MBSP-C Spring $N = 272$; Numbers and Operations, Measurement, Algebraic Concepts, and Data Analysis and Probability MBSP-C Fall $N = 498$, MBSP-C Winter $N = 501$, MBSP-C Spring $N = 508$. Second Grade Cohort PSSA-M Composite and Geometry subtest MBSP-C Fall $N = 549$, MBSP-C Winter $N = 556$, MBSP-C Spring $N = 564$; Numbers and Operations, Measurement, Algebraic Concepts, and Data Analysis and Probability MBSP-C Fall $N = 793$, MBSP-C Winter $N = 807$, MBSP-C Spring $N = 817$. Third Grade Cohort PSSA-M Composite and Geometry subtest MBSP-C Fall $N = 901$, MBSP-C Winter $N = 917$, MBSP-C Spring $N = 927$; Numbers and Operations, Measurement, Algebraic Concepts, and Data Analysis and Probability MBSP-C Fall $N = 1160$, MBSP-C Winter $N = 1187$, MBSP-C Spring $N = 1201$. ** $p < .01$

The correlation coefficients summarized in Table 35 indicate a moderate to strong positive correlation between MBSP-C in the fall, winter, and spring of first, second, and third grade with PSSA-M composite scores in third grade. The strength of correlation between MBSP-C and PSSA-M subtests varied ($r = .023$ to .561) based on the time of administration and grade level.

Correlation between MBSP-C in first grade and PSSA-M subtests in third grade. A weak, positive association was observed between first grade MBSP-C fall data and third grade PSSA-M Numbers and Operations subtest scores, $r = .204$, $p < .0005$. The correlation between MBSP-C and PSSA-M Numbers and Operations was moderate based on winter data ($r = .420$, $p < .0005$) and strong based on spring data ($r = .501$, $p < .0005$).

There was a weak, positive correlation between first grade MBSP-C fall data and scores on the third grade PSSA-M Measurement subtest ($r = .240$, $p < .0005$). A moderate correlation was observed between MBSP-C in the winter of first grade and performance on the third grade PSSA-M Measurement subtest ($r = .339$, $p < .0005$). The relationship between MBSP-C and PSSA-M Measurement remained moderate in the spring ($r = .406$, $p < .0005$). The correlation between first grade MBSP-C in the fall ($r = .309$, $p < .0005$), winter ($r = .418$, $p < .0005$), and spring ($r = .396$, $p < .0005$) with performance on the PSSA-M Geometry subtest in third grade was moderate.

A weak, positive correlation was observed between first grade MBSP-C scores in the fall and third grade PSSA-M Algebraic Concepts scores ($r = .272$, $p < .0005$). There was a positive, moderate correlation between MBSP-C in the winter ($r = .387$, $p < .0005$) and spring ($r = .416$, $p < .0005$) of first grade with performance on the PSSA-M Algebraic Concepts subtest in third grade.

A significant correlation was not observed between first grade MBSP-C fall data and performance on the third grade PSSA-M Data Analysis and Probability subtest. However, a weak, positive associate was observed in the winter of first grade ($r = .277$, $p < .0005$). There was a positive, moderate correlation between MBSP-C in the spring of first grade and

performance on the PSSA-M Data and Analysis subtest administered in the spring of third grade ($r = .324$, $p < .0005$).

Correlation between MBSP-C in second grade and PSSA-M subtests in third grade. A moderate, positive association was observed between second grade MBSP-C in the fall ($r = .352$, $p < .0005$), winter ($r = .450$, $p < .0005$), and spring ($r = .451$, $p < .0005$) with third grade PSSA-M Numbers and Operations subtest scores. There was a moderate, positive correlation between second grade MBSP-C fall data and scores on the third grade PSSA-M Measurement subtest ($r = .343$, $p < .0005$). A moderate, positive correlation was observed between MBSP-C in the winter of second grade and performance on the third grade PSSA-M Measurement subtest ($r = .374$, $p < .0005$). The relationship between MBSP-C and PSSA-M Measurement remained moderate in the spring ($r = .380$, $p < .0005$).

Weak, positive correlations were observed between second grade MBSP-C in the fall ($r = .190$, $p < .0005$), winter ($r = .287$, $p < .0005$), and spring ($r = .259$, $p < .0005$) with performance on the PSSA-M Geometry subtest in third grade. There was a moderate, positive correlation between second grade MBSP-C scores in the fall and third grade PSSA-M Algebraic Concepts scores ($r = .314$, $p < .0005$). There was a positive, moderate correlation between MBSP-C in the winter ($r = .402$, $p < .0005$) and spring ($r = .403$, $p < .0005$) of second grade with performance on the PSSA-M Algebraic Concepts subtest in third grade.

A weak, positive correlation was observed between second grade MBSP-C fall data and performance on the third grade PSSA-M Data Analysis and Probability subtest ($r = .245$, $p < .0005$). The correlation between MBSP-C in the winter of second grade and performance on the PSSA-M Data Analysis subtest in third grade remained weak and positive ($r = .281$, $p < .0005$). There was a positive, moderate correlation between MBSP-C in the spring of second grade and

performance on the PSSA-M Data and Analysis subtest administered in the spring of third grade ($r = .306$, $p < .0005$).

Correlation between MBSP-C in third grade and PSSA-M subtests in third grade. A moderate, positive association was observed between third grade MBSP-C fall data and third grade PSSA-M Numbers and Operations subtest scores, $r = .399$, $p < .0005$. The correlation between MBSP-C and PSSA-M Numbers and Operations was strong based on winter data ($r = .536$, $p < .0005$) and spring data ($r = .546$, $p < .0005$).

There was a weak, positive correlation between third grade MBSP-C fall data and scores on the third grade PSSA-M Measurement subtest ($r = .295$, $p < .0005$). A moderate correlation was observed between MBSP-C in the winter of third grade and performance on the third grade PSSA-M Measurement subtest ($r = .392$, $p < .0005$). The relationship between MBSP-C and PSSA-M Measurement remained moderate in the spring ($r = .406$, $p < .0005$).

There was a weak, moderate correlation between MBSP-C in the fall of third grade and performance on the PSSA-M Geometry subtest ($r = .190$, $p < .0005$). The relationship between MBSP-C in the winter and PSSA-M Geometry scores remained weak ($r = .297$, $p < .0005$). There was a moderate, positive correlation between MBSP-C in the spring with performance on the Geometry subtest ($r = .349$, $p < .0005$).

A moderate, positive correlation was observed between third grade MBSP-C scores in the fall and third grade PSSA-M Algebraic Concepts scores ($r = .301$, $p < .0005$). There was a positive, moderate correlation between MBSP-C in the winter ($r = .418$, $p < .0005$) and spring ($r = .379$, $p < .0005$) of third grade with performance on the PSSA-M Algebraic Concepts subtest in third grade.

A weak, positive correlation was observed between third grade MBSP-C fall data and performance on the third grade PSSA-M Data Analysis and Probability subtest ($r = .273$, $p < .0005$). The correlation between MBSP-C in the winter of third grade and performance on the PSSA-M Data Analysis subtest in third grade was moderate and positive ($r = .355$, $p < .0005$). There was a positive, moderate correlation between MBSP-C in the spring of third grade and performance on the PSSA-M Data and Analysis subtest administered in the spring of third grade ($r = .379$, $p < .0005$).

Correlation between sex, resource availability and mathematical achievement. Pearson correlations were generated to determine what, if any, correlation exists between sex, resource availability and mathematical achievement. Please refer to the correlation coefficients (r) summarized in Table 36.

Table 36

Pearson Correlations Between Sex, Resource Availability, PSSA-M Scores and MBSP-C

	First Grade		Second Grade		Third Grade	
	Sex	Resource Availability	Sex	Resource Availability	Sex	Resource Availability
PSSA-M Composite	-.028	.179**	.031	.201*	.015	.228**
Numbers and Operations	.074	.110*	.063	.156**	.037	.201**
Measurement	.124**	.099*	.098**	.135**	.058*	.176**
Geometry	-.039	.144*	-.016	.142**	-.020	.168**
Algebraic Concepts	.000	.125**	.026	.170**	.027	.165**
Data Analysis and Probability	.024	.064	.034	.104**	.016	.136**
MBSP-C Fall	.023	.131**	.074*	.174**	-.007	.137**
MBSP-C Winter	.084	.168**	.071*	.153**	-.032	.178**
MBSP-C Spring	.096*	.124**	.049	.162**	-.042	.215**

Note. MBSP-C = Monitoring Basic Skills Progress- Computation Probe; PSSA-M = Pennsylvania System of School Assessment, Mathematics. First Grade Cohort PSSA-M Composite and Geometry subtest MBSP-C Fall $N = 266$, MBSP-C Winter $N = 269$, MBSP-C Spring $N = 272$; Numbers and Operations, Measurement, Algebraic Concepts, and Data Analysis and Probability MBSP-C Fall $N = 498$, MBSP-C Winter $N = 501$, MBSP-C Spring $N = 508$. Second Grade Cohort PSSA-M Composite and Geometry subtest MBSP-C Fall $N = 549$, MBSP-C Winter $N = 556$, MBSP-C Spring $N = 564$; Numbers and Operations, Measurement, Algebraic Concepts, and Data Analysis and Probability MBSP-C Fall $N = 793$, MBSP-C Winter $N = 807$, MBSP-C Spring $N = 817$. Third Grade Cohort PSSA-M Composite and Geometry subtest MBSP-C Fall $N = 901$, MBSP-C Winter $N = 917$, MBSP-C Spring $N = 927$; Numbers and Operations, Measurement, Algebraic Concepts, and Data Analysis and Probability MBSP-C Fall $N = 1160$, MBSP-C Winter $N = 1187$, MBSP-C Spring $N = 1201$. * $p < .05$, ** $p < .01$

Student sex did not demonstrate a significant correlation with mathematical outcomes in first, second, or third grade, with the exception of performance on the PSSA-M Measurement subtest. Sex demonstrated a weak, positive correlation with Measurement outcomes in third grade for first, second, and third grade cohorts. There was also a weak, positive correlation between sex and MBSP-C in the spring of first grade and fall and winter of second grade.

Resource availability in first grade demonstrated a weak, positive correlation with performance on MBSP-C in the fall ($r = .131$, $p = .003$), winter ($r = .168$, $p < .0005$), and spring ($r = .124$, $p = .005$) of first grade. Resource availability in first grade demonstrated a weak, positive correlation with the PSSA-M Composite ($r = .179$, $p = .003$), Numbers and Operations subtest ($r = .110$, $p = .013$), Geometry subtest ($r = .144$, $p = .017$), and Algebraic Concepts subtest ($r = .125$, $p = .005$) in third grade.

Resource availability in second grade demonstrated a weak, positive correlation with performance on MBSP-C in the fall ($r = .174$, $p < .0005$), winter ($r = .153$, $p < .0005$), and spring ($r = .162$, $p < .0005$) of second grade. Resource availability in second grade demonstrated a weak, positive correlation with the PSSA-M Composite ($r = .201$, $p < .0005$), Numbers and Operations subtest ($r = .156$, $p < .0005$), Measurement subtest ($r = .135$, $p < .0005$), Geometry subtest ($r = .142$, $p = .001$), Algebraic Concepts subtest ($r = .170$, $p <.0005$), and Data Analysis and Probability subtest ($r = .104$, $p = .003$) in third grade.

Resource availability in third grade demonstrated a weak, positive correlation with performance on MBSP-C in the fall ($r = .137$, $p < .0005$), winter ($r = .178$, $p < .0005$), and spring ($r = .215$, $p < .0005$) of third grade. Resource availability in third grade demonstrated a weak, positive correlation with the PSSA-M Composite ($r = .228$, $p < .0005$), Numbers and Operations subtest ($r = .201$, $p < .0005$), Measurement subtest ($r = .176$, $p < .0005$), Geometry subtest ($r = .168$, $p <.0005$), Algebraic Concepts subtest ($r = .165$, $p < .0005$), and Data Analysis and Probability subtest ($r = .136$, $p < .0005$) in third grade.

Summary

This chapter reviews the statistical analysis used to answer the research question. Complications to the study were acknowledged. Descriptive statistics, tests of assumption for MLR, and finally the results of the stepwise MLR models with first, second, and third grade data were reported and discussed. MBSP-C in the fall, winter, and spring of first, second, and third grade were found to have a strong predictive relationship with overall performance on the PSSA-M. The predictive relationship of the independent variables with the PSSA-M subtests varied based on grade level. However, the increases in the total variance explained on PSSA-M subtests consistently increased from first to third and from second to third grade. First grade predictor variables frequently accounted for more variance on PSSA-M subtests when compared to second grade variables. MBSP-C probes in the winter and spring consistently demonstrated moderate to strong predictive relationships with PSSA-M subtests. The relationship between the PSSA-M subtests MBSP-C in the fall, sex, and resource availability were less consistently predictive and demonstrated a weaker predictive relationship than MBSP-C winter and spring data.

These findings support the hypothesis that MBSP-C in the fall of first grade will have the weakest correlation with PSSA-M. The hypothesis that the strength of the predictive relationship will increase with proximity to the third grade PSSA-M is supported. However, R and adjusted R^2 values when predicting PSSA-M Composite in first and second grade are very similar and both indicate a strong predictive relationship and effect size. It was further hypothesized that MBSP-C would have the strongest predictive relationship with the Numbers and Operations subtest. This hypothesis is rejected, as MBSP-C demonstrated a stronger predictive relationship with PSSA-M Composite scores than the Numbers and Operations subtest. The predictive

relationship of sex and resource availability with PSSA-M scores are inconsistent. In the instances where sex was predictive of PSSA-M, the increase in variance was statistically significant but small in relevance for purposes of practical application. With the exception of first grade, males outperformed females when sex was identified as a significant predictor in the regression model. In the regression models with resource availability identified as predictive of PSSA-M, the increase in variance accounted for by the model was statistically significant, but small in relevance for purposes of practical application. It should be noted that resource availability was more frequently identified as a significant predictor the longer students were in school.

CHAPTER V

DISCUSSION

Introduction

The current research study investigated the predictive validity of a brief math computation probe, Monitoring Basic Skills Progress, Computation probe (MBSP-C) in the fall, winter, and spring of first, second, and third grades with the Pennsylvania System of School Assessment Mathematics exam (PSSA-M) administered in the spring of third grade. The relationship between mathematics achievement and resource availability and sex was also explored. This chapter offers a review of the research, findings and discussion of data analysis, study limitations, suggestions for future research, and implications for the field of school psychology.

Overview

Educational reform supported by Every Student Succeeds Act (ESSA) and initiated by its predecessor, No Child Left Behind (NCLB), pushed education systems to provide high quality, effective instruction to all students. Schools responded to this by developing systems to provide differentiated instruction and early intervention to students who are at-risk for academic, social, and emotional deficits. Multi-tiered Systems of Support (MTSS) is the term given to this service delivery model.

MTSS models focus on the improvement of student outcomes in academics and social/emotional development by providing high quality instruction with differentiation based on student need. There are six key features of successful MTSS models, outlined by the National Association of School Psychologists (NASP): (a) differentiated instruction within a high quality core curriculum, (b) universal screening, assessment, and monitoring progress; (c) focus on

prevention and intervention, (d) fidelity of interventions, (e) evidence-based practices, and (f) professional development (Cowan et al., 2013; Stoiber, 2014). Universal screenings are one of the key components of MTSS systems. Universal screenings are defined as "the systematic assessment of all children within a given class, grade, school building, or school district, on academic and/or social emotional indicators that the school personnel and community have agreed are important" (Ikeda, Neessen, & Witt, 2008, p. 103).

In comparison to reading, there is a relative dearth of research in the area of mathematics to provide educators appropriate direction (Clarke, Haymond, & Gersten, 2014; Methe, 2009). As multi-tiered models of service delivery continue to grow in implementation, and science, technology, engineering, and mathematics (STEM) skills become more of an educational focus, it is vital that research-based practices are also applied to the area of mathematics (Gersten et al., 2012; VanDerHeyden, 2010).

A comprehensive review of research indicates mathematical deficits remain persistent in students who are low-achieving. If mathematical deficits are not addressed with rigorous instruction and intervention the performance gap widens as students matriculate through their school career. Early identification and intervention to address mathematical difficulties should be a principal focus for educational systems. Research indicates students who initially place in the bottom 10^{th} percentile when entering kindergarten but were performing above the 10^{th} percentile upon exiting only had a 30% chance of performing below the 10^{th} percentile five years later while in fifth grade (Morgan et al., 2009, 2011). Without intervention in kindergarten, however, students who demonstrate math skills within the bottom 10^{th} percentile in kindergarten have a 70% likelihood of remaining below the 10^{th} percentile five years later (Martin et al., 2012; Morgan et al., 2009; 2011). This highlights the need for and efficacy of early identification and

intervention. MTSS models employ universal screening measures to identify students who may be at-risk for developing deficits for the purpose of early intervention.

The present study focuses on the predictive strength of a computation measure, MBSP-C with a criterion measure, the PSSA-M. This study builds upon previous research utilizing MBSP-C probes (Fuchs, Hamlett, & Fuchs, 1999) as a universal screening measure. Shapiro, Keller, Lutz, Santoro, and Hintze (2006) found MBSP-C, given the same school year as the PSSA-M, to have a moderate to strong predictive relationship with third grade PSSA-M Composite scores. Keller-Margulis, Shapiro, and Hintze (2008), later explored the predictive strength of MBSP-C in first and second grade, using the same data set previously collected for the Shapiro et al. (2006) study. Results of that analysis indicated a strong relationship between MBSP-C in the fall (r = .52), winter (r = .54), and spring (r = .60) of second grade and PSSA scores. First grade MBSP-C fall data demonstrated a weak (r = .27), but statistically significant relationship with PSSA scores in third grade. MBSP-C in the winter (r = .59) and spring (r = .50) demonstrated a strong relationship with third grade PSSA scores. Several concerns were noted with this study which warranted replication. Concerns include the exclusion of students identified as being in need of special education services and a relatively small sample size.

A review of research regarding the impact of resource availability on mathematical learning outcomes is inconsistent. However, there is a growing body of research that suggests students living in low socio-economic status (SES) homes are likely to demonstrate difficulty with mathematical learning. Mathematical deficits may also be more persistent for students living in poverty when compared to students who are not living in poverty. These findings suggest students living in low SES homes many benefit significantly from early, intensive math intervention (Reardon, 2013).

There is a mixed body of research on whether or not sex has an impact on mathematical learning outcomes. Some research indicates a significant difference in mathematical learning between males and females, but other research disputes an achievement gap between males and females (McGraw et al., 2006; Stoet & Geary, 2013). Given inconsistent findings regarding the role of sex and SES on students' mathematical proficiency, these factors are further investigated to determine what, if any, impact these have on mathematical learning outcomes.

Research Question and Hypotheses

The broad research question under investigation in the current study is: To what extent does a universal mathematics screening, MBSP-C, in first, second, and third grade, sex, and free and reduced meal status predict math achievement as reported on PSSA-M Composite scores and five subtests of the PSSA-M in third grade? The dependent variables considered within this research question included MBSP-C – fall, winter, and spring completed in first, second, and third grade; PSSA-M Numbers and Operations scores; PSSA-M Measurement scores; PSSA-M Geometry scores; PSSA-M Algebraic Concepts scores; PSSA-M Data Analysis and Probability scores; and PSSA-M Composite scores.

It was hypothesized that MBSP-C scores in first, second, and third grade would predict math achievement as measured by the composite score and five subtests of the PSSA-M in third grade. Based on previous research, it was hypothesized the correlation between MBSP-C and PSSA-M scores would be moderate to strong. It was hypothesized that student performance in the fall of first grade would have the weakest correlation with PSSA-M performance and student performance in the spring of third grade would have the strongest correlation with third grade PSSA-M achievement due to time proximity between MBSP-C and PSSA-M administration.

It was also hypothesized that MBSP-C would have the strongest correlation with the Numbers and Operations subtest of the PSSA-M. The Numbers and Operations subtest of the PSSA-M asks students to demonstrate an understanding of numbers, ways of representing numbers, relationships among numbers and number systems, an understanding of the meanings of operations, use of operations and understanding how they relate to each other, the ability to compute accurately and fluently, and the capacity to make reasonable estimates. These skills closely resemble those assessed on the MBSP-C probes. Therefore, it was predicted the strongest predictive relationship would exist between MBSP-C and the Numbers and Operations subtest of the PSSA-M.

It was further hypothesized that resource availability would account for a significant amount of variance on PSSA-M achievement, with the potential to decrease the longer students are in a high quality educational setting. Conversely, previous research indicated students living in poverty are more resistant to improvement in mathematics instruction, so there is potential for the amount of variance explained by resource availability to remain the same or increase the longer a student is in an educational setting. It was hypothesized that sex and resource availability would have a moderate association with math achievement, based on highlights from the 2007 Trends in International Mathematics and Science Study (TIMSS; Gonzales et al., 2009).

Hypotheses with MBSP-C as a Predictor Variable

To what extent does a universal mathematics screening, MBSP-C, in first, second, and third grade, sex, and free and reduced meal status predict math achievement as reported on PSSA-M Composite scores and five subtests of the PSSA-M in third grade? The results of the multiple linear regression (MLR) analysis indicate a strong predictive relationship between MBSP-C in the fall, winter, and spring of first, second, and third grade with PSSA-M

performance in the spring of third grade. MBSP-C in third grade demonstrated the strongest predictive relationship with PSSA-M composite scores with an *R* value of .630 and explained 39% of total variance. First grade demonstrates the second strongest predictive relationship with an *R* value of .586 for the full regression model. First grade data explained 33% of total variance of third grade PSSA-M Composite outcomes. Second grade variables demonstrated the weakest predictive relationship (*R* = .560) and accounted for 31% of total variance.

Pearson correlation were analyzed to determine the predictive relationship of each MBSP-C administration with third grade PSSA-M performance. Results of Pearson correlations indicate after the fall of first grade, the strength of the relationship of MBSP-C with third grade PSSA-M Composite outcomes remain relatively consistent with moderate to strong predictive relationships from the winter of first grade through the spring of third grade. It further was hypothesized that student performance in the fall of first grade would have the weakest correlation with PSSA-M performance and student performance in the spring of third grade would have the strongest correlation with third grade PSSA-M achievement. This was hypothesized given the proximity between MBSP-C and PSSA-M administration. Previous research indicated the strength of prediction increased the shorter the duration between administration of the screening tool and criterion measure. When validating a number sense screening tool for use in kindergarten and first grade, Jordan et al. (2010) found a significant increase in the main effect over the course of six administrations as students demonstrated age-appropriate changes in achievement.

MBSP-C in the fall of first grade (*r* = .388) demonstrates a moderate relationship with third grade PSSA-M composite scores. However, by the winter of first grade, MBSP-C demonstrated a strong predictive relationship with third grade PSSA-M composite scores (*r* =

.501). The relationship between MBSP-C in first, second, and third grade with PSSA-M composite scores increased from fall to winter, but then decreased slightly from winter to spring. MBSP-C winter scores consistently demonstrated the strongest predictive relationship with third grade PSSA-M composite scores. Therefore, this hypothesis is rejected when using MBSP-C to predict overall performance on the PSSA-M. Based on the results of the MLR, MBSP-C spring scores do not have the strongest predictive relationship with PSSA-M.

Pearson correlations indicated the strength of the relationship increased with proximity between MBSP-C and PSSA-M administration. Therefore, the hypothesis is accepted when looking at the predictive relationship between MBSP-C in the fall, winter, and spring of first, second, and third grades with third grade PSSA-M subtest scores. Correlation coefficients increased from fall to winter and winter to spring for each PSSA-M subtest at each grade level, with the exception of the relationship between first grade MBSP-C and Geometry. MBSP-C in the winter (r = .418) of first grade demonstrates a stronger strength of relationship than MBSP-C spring scores (r = .396).

It was hypothesized that MBSP-C would have the strongest correlation with the numbers and operations portion of the PSSA-M. The Numbers and Operations subtest of the PSSA-M requires students demonstrate an understanding of numbers, ways of representing numbers, relationships among numbers and number systems, meanings of operations, use of operations and understanding how they relate to each other, the ability to compute accurately and fluently, and the capacity to make reasonable estimates. These skills closely resemble those assessed on the MBSP-C probes. Therefore, it was predicted the relationship between MBSP-C and the Numbers and Operations subtest of the PSSA-M would be the strongest.

This hypothesis is rejected. MBSP-C demonstrates the strongest relationship with PSSA-M Composite scores. However, by the spring of first, second, and third grade the relationship between MBSP-C and Numbers and Operations is second only to PSSA-M composite scores at each grade level. The strength of the relationship between MBSP-C in the fall, winter, and spring at each grade level is consistently weakest with third PSSA-M Data Analysis and Probability subtest scores. As anticipated, these results suggest computation skills have the weakest relationship with skills represented on the third grade Data Analysis and Probability subtest. The Data Analysis and Probability subtest of the third grade PSSA-M requires students to formulate or answer questions that can be addressed with data and/or organize, display, interpret or analyze data (Data Recognition Corporation, 2014). An understanding of numbers and computation is required to successfully complete data analysis and probability problems. However, data analysis and probability require a level of conceptual understanding that is not captured with a simple computation probe (Locuniak & Jordan, 2008).

Hypotheses with Sex as a Predictor Variable

It was hypothesized that sex would have a moderate association with math achievement, based on highlights from the 2007 Trends in International Mathematics and Science Study (Gonzales et al., 2009). An analysis of the U.S National Assessment of Educational Progress (NEAP) from 1990 to 2003, found that sex gaps within math achievement continue to exist, with males performing slightly better than females, especially in the upper end of score distributions. Achievement gaps between the sexes were largest in the areas of measurement, number and operations, and geometry (McGraw et al., 2006).

Pearson correlations between student sex and math achievement as measured by performance on the third grade PSSA-M Composite, five PSSA-M subtests, and MBSP-C probes

in the fall, winter, and spring of first, second, and third grade, were reviewed. Correlation coefficients do not indicate a significant association between student sex and mathematical outcomes, with the exception of the Measurement subtest at each grade level and MBSP-C in the spring of first grade, fall of second grade, and winter of second grade. Although these correlations are considered statistically significant, the strength of the correlations were weak. This means it can be said with some certainty that the correlation is not zero. However, the majority of *r* values fall below 0.1. The exceptions were student sex in first grade and PSSA-M Measurement scores in third grade, which demonstrated a weak, positive association ($r = .124$). Based on these findings, sex has little to no relationship with math performance.

The inclusion of sex in full MLR models was inconsistent. Sex was found to be a statistically significant contributor to the overall variance explained in only five of the 18 MLR models generated in the present study. It is important to differentiate between statistical significance and practical significance. The instances in which sex was determined to be a statistically significant contribution to the regression model only increased the overall variance explained by very small amounts, 1% or less. Therefore, it is important to consider whether or not it is relevant to include sex as a variable for application purposes. The additional time required to collect these data combined with a potential to inadvertently potentiate the stigma around a mathematical gender gap by including sex in the regression model negate the small statistical contribution the inclusion of sex offers to the regression model. This hypothesis is rejected. Sex does not demonstrate a meaningful association with mathematical achievement.

Hypotheses with Resource Availability as a Predictor Variable

It was hypothesized resource availability will account for a significant amount of variance on PSSA-M achievement, with the potential to decrease the longer students are in a

high quality educational setting. However, previous research has indicated students living in poverty are more resistant to improvement in mathematics instruction, so there is potential for the amount of variance accounted for by resource availability to remain the same or increase the longer a student is in an educational setting (Aud et al., 2010). Resource availability was determined by whether or not a student received free and reduced lunch.

This hypothesis is accepted. Resource availability demonstrates a statistically significant, but weak, positive association with mathematical outcomes measured by MBSP-C while in first, second, and third grade and PSSA-M administered in the spring of third grade. The strength of the relationship increased slightly from first to second grade and from second to third grade. The number of full MLR models which included resource availability as a significant contributor to the overall variance on the PSSA-M went from zero with first grade data to five with second and third grade data. This suggests resource availability becomes a more relevant predictor the longer students are in school. However, more research is needed to determine whether the impact resource availability continues to increase, plateaus, or begins to decrease. More research is also warranted to determine whether free and reduced lunch status is an appropriate way to determine resource availability or socio-economic status.

Second and third grade data analysis included resource availability in the full MLR regression models for PSSA-M Composite and all PSSA-M subtests, with the exception of Data Analysis and Probability. While this contribution was considered statistically significant, the actual increase in variance explained by resource availability was relatively small. School systems could choose to exclude these data from prediction models without impacting the integrity of MBSP-C as a universal screening tool. It is likely the additional time and effort to include these data would not significantly improve student learning outcomes. Additionally,

resource availability is based on participation in the free or reduced lunch program which requires parents to complete an application and disclose their household income. It is possible that some students who would qualify for free or reduced lunch have not applied for participation in the program (National Forum on Educational Statistics, 2015).

Discussion

The purpose of this study was to examine the predictive validity of a computation probe, MBSP-C, in first, second, and third grade with PSSA-M performance in third grade to determine its utility as a universal screening measure. The results of this study have several findings relevant to the use of computation measures as universal screening tools for the identification of students who could benefit from additional math intervention and/or instruction.

First, MBSP-C demonstrates a strong predictive relationship with overall performance on the third grade PSSA-M as early as first grade. The full regression model for PSSA-M with first grade variables included MBSP-C in the fall, winter, and spring in addition to student sex. This MLR regression model accounted for 33% of variance explained on PSSA-M composite scores administered in third grade. The regression model without the inclusion of sex, accounted for 32% of the total variance explained, which is still substantial. Pearson correlation coefficients indicate a moderate correlation between MBSP-C in the fall of first grade with the PSSA-M in third grade and a strong correlation by the winter of first grade. These results are interpreted to mean that as early as winter of first grade, school systems can make a strong prediction regarding student learning outcomes for math in third grade. These findings support the use of a computation probes as a universal screening tool in the area of mathematics. Correlation coefficients of .40 or higher between the universal screening tool and a well-established criterion measure indicate adequate levels of predictive validity for the purpose of universally screening

students (Burns, Haegele, & Petersen-Brown, 2014). MBSP-C probes from the winter of first grade through the spring of third grade demonstrate a moderate to strong correlation with overall performance on the PSSA-M administered in the spring of third grade.

The Predictive Relationship between MBSP-C and PSSA-M Subtests

It was hypothesized that MBSP-C would have the strongest predictive relationship with the PSSA-M Numbers and Operations subtest. This hypothesis was rejected, as the strongest predictive relationship was observed between MBSP-C and PSSA-M Composite scores. However, the relationship between MBSP-C and the Numbers and Operations subtest was strong, and only slightly less than the predictive relationship with PSSA-M Composite scores. It is important to note, 40% to 50% of the items on the third grade PSSA-M fall under the Numbers and Operations subtest. Each remaining PSSA-M subtest accounts for approximately 12% to 15% of the items on the PSSA-M (Data Recognition Corporation, 2013). Therefore, it can be concluded the PSSA-M Composite score is heavily influenced by performance on the PSSA-M Numbers and Operations subtest.

There is an essential but not definite relationship between fluency of computation skills and successful application and understanding of math concepts that has been likened to the relationship between oral reading fluency and reading comprehension (Locuniak & Jordan, 2008). Given this relationship, it was rationalized that computation measures would have a moderate predictive relationship with the remaining four subtests PSSA-M subtests despite a low face validity. Measurement, Geometry, Algebraic Concepts, and Data Analysis and Probability subtests, demonstrate a low face validity with the MBSP-C probes. Previous research suggests there is still a moderate correlation between computation skills and higher level mathematical thinking (Codding et al., 2015). The present study supports these findings. Results of the MLR

analysis indicate a moderate predictive relationship between MBSP-C in the fall, winter, and spring of first, second, and third grade and PSSA-M subtests with low face validity, Measurement, Geometry, Algebraic Concepts, and Data Analysis and Probability. The moderate predictive relationship observed between MBSP-C and PSSA-M subtest with low face validity not only support the use of computation measures as a universal screener, but also has a general outcome measure in mathematics. There was one oddity in the prediction models produced by MLR analysis. Specifically, the MLR between first grade data and performance on the third grade PSSA-M Data Analysis and Probability subtest.

In the prediction model an increase in first grade MBSP-C winter scores resulted in lower scores on the PSSA-M Data Analysis and Probability subtest in third grade. It is unknown what factors contributed to this atypical outcome. It may be related to the instruction first grade students received from the fall to winter of first grade, which is focused primarily on computation skills. This could have resulted in score improvements on the computation measure and rates of growth that were not maintained past this time period. Mean scores from the first grade fall to first grade winter probe increased substantially, with a mean score of 3.67 in the fall and 14.55 in the winter. This study would have to be replicated with a different data set to determine if score improvements on winter administration of MBSP-C in first grade really predict a score decrease on third grade PSSA-M Data Analysis and Probability performance.

When the predictive relationships between MBSP-C and PSSA-M subtests were analyzed without the consideration of other predictor variables the findings were similar to MLR results, with a few exceptions. Based on the results of Pearson correlation analysis, there is a weak predictive relationship between MBSP-C in the fall of first grade with third grade performance on the Numbers and Operations subtest, Measurement subtest, and Algebraic Concepts subtest.

The relationship between fall MBSP-C in first grade and third grade Data Analysis and Probability is so small, it is negligible. MBSP-C in the winter of first grade demonstrated a weak predictive relationship with Data Analysis and Probability scores in third grade. However, the remaining predictive relationships between MBSP-C in first grade and PSSA-M subtests are moderate to strong, based on winter and spring data.

Based on the results of Pearson correlations, weak predictive relationships were also observed between MBSP-C in the fall, winter, and spring of second grade with third grade performance on the Geometry subtest. There was a weak predictive relationship between second grade MBSP-C fall and winter administration with third grade Data Analysis and Probability subtest.

The results of Pearson correlation analysis with third grade data indicated a weak predictive relationship between the fall and winter administration of MBSP-C and performance on the Geometry subtest, MBSP-C in the fall and performance on the Measurement subtest, and MBSP-C in the fall and performance on the Data Analysis and Probability subtest.

Pearson correlations illustrate the strength of the predictive relationship is not as strong between MBSP-C at specific points in time as when looked at collectively with MLR. However, a strong predictive relationship is observed between MBSP-C and PSSA-M composite scores and most subtests from the winter administration through the end of third grade. School systems should interpret fall data with caution and consider reviewing spring data from the previous year in conjunction with fall data. For example, in the fall data analysis teams may want to look at first grade spring data in addition to fall second grade data when making determinations about which student would benefit from additional math intervention. Consequently, MBSP-C in the

winter and spring demonstrated a moderate to strong predictive relationship with overall math achievement despite low face validity.

Sex as a Predictor Variable

Previous research regarding whether or not there is a mathematical achievement gap between males and females generated inconsistent findings. Research that supports an achievement gap, with males outperforming females, indicated small, but statistically significant differences in mathematical achievement (McGraw et al., 2006; Stoet & Geary, 2013). These findings are supported by the present study. There was a weak, but statistically significant, contribution to the overall variance explained when first grade data were used to predict third grade PSSA-M Composite scores, when second grade data were used to predict third grade PSSA-M Measurement scores, and when third grade data were used to predict performance on the PSSA-M Numbers and Operations subtest, Measurement subtest, and Algebraic Concepts subtest.

However, similar to the results reported by Else-Quest et al. (2010), differences between male and female achievement were negligible. The inclusion of sex as an independent variable indicates no to very minimal differences between male and female mathematical outcomes. When student sex was identified as a statistically significant contributor to variance in mathematical outcomes, it increased total variance explained by no more than 1%. These findings suggest sex is not a meaningful predictor of math skills. Therefore, the results of this study indicate student sex is not an adequate predictor variable.

Resource Availability as a Predictor Variable

Previous research which investigated the correlation between SES (resource availability) indicated students from low-income households do not reach mathematical proficiency (NEAP,

U.S. Department of Education, 2015). Disparities between the mathematical learning of students in low-income households and middle-class to high-income households has increased over recent years. Rates of growth indicate that the improvement in mathematical performance that has been observed nationally cannot be generalized to students living in low-income households. Students living in low-income households were also found to be more resistant to intervention (Reardon, 2013; Reardon & Bischoff, 2011).

The results of the present study indicate resource availability demonstrated a weak correlation with future mathematical outcomes. The variance accounted for by resource availability is minimal, but does increase slightly from first grade to second grade and from second grade to third grade. This cautiously supports the hypothesis that the impact of poverty increases the longer students are in school. However, more research is needed in this area. It is important to note, the impact of resource availability on total variance accounted for on PSSA-M scores was statistically significant, but minimal when considering practical implications for collecting additional data and targeting students living in poverty for intervention. Resource availability could be excluded from the prediction model without drastically altering its functionality.

Limitations of the Study

There are several limitations to this study which should be acknowledged and considered when generalizing the results. First, the archival data used in this study were from a convenience sample. Use of archival data and a convenience sample created the potential for threats to internal and external validity. Threats to internal validity included the variability in the quality and intensity of math instruction students received during the period of data collection. While curricular variability is limited somewhat by the adoption of the PA Common Core curriculum,

the use of archival data does not permit for quality control checks of instruction. Secondly, standardized administration of MBSP-C is assumed but cannot be guaranteed due to the use of archival data. The school district did not retain the completed MBSP-C probe sheets. As a result, scoring accuracy and correct data entry into the data warehouse system could not be independently verified.

The 2014 PSSA technical manual indicated Geometry as a reported domain, but no Geometry scores were reported for PSSA-M administered to third grade students in 2014. According to the PSSA technical manual this was due to the transition from the previously-used standards to the newly-adopted Pennsylvania Core Standards. Consequently, PSSA data from the 2013-2014 year were excluded from the data set for analysis of PSSA-M Composite score and Geometry data, which resulted in smaller sample sizes for these two dependent variables. All other scores on the 2014 PSSA-M were comparable to their counterpart scores in the previous PSSA-M administrations included in this study.

Threats to external validity stem primarily from the use of a convenience sample. All data were collected from the same rural school district in Pennsylvania. The sample may not be representative of the population as a whole, which limits generalizability to more diverse populations. A large majority of the sample population (91% to 93%) were composed of students who identify as Caucasian. It would be beneficial to replicate this study within a population that is more diverse and representative of the general population.

Students receiving special education services were included in the present study, with the exception of those who took the modified format of the PSSA. This exclusion was intentional and because PSSA Mathematics data were not available for students who took the modified version of the PSSA. This means that students with the most severe disabilities were not

included in the data set, which could have skewed the results. This has the potential to effect generalization of the results to other populations, therefore this exclusion should be noted. However, from an application standpoint, it is unlikely educational systems are interested in predicting how students with severe disabilities will perform on a test they are not expected to take.

Other potential limitations stem from curricular changes which have occurred since Monitoring Basic Skills Progress was originally published. The MBSP-C probes were based on state standards from two to three decades ago (Fuchs, Hamlett, & Fuchs, 1998, 1999). During that time, there has been a significant change in mathematics curriculums and an increase in excepted learning outcomes for students at each grade level. This suggests MBSP-C may not accurately reflect what students are expected to learn throughout the school year as well as more recently developed math computation probes.

While the idea of basing a universal screening and progress monitoring measure off of academic standards certainly has merit, it is likely this measure needs updating to reflect the Common Core State Standards (Clarke et al., 2014). The published normative scores for the MBSP-C are also quite dated, having been published in 1999. Therefore, any school system considering the use of MBSP-C as a universal screening tool should seriously consider generating local normative data with data from their own student population. Best practices suggest normative data be updated every five to seven years (Stewart & Silberglitt, 2008).

Resource availability was determined by whether or not a student received free or reduced lunch. It is important to acknowledge participation in the free or reduced lunch program is voluntary and requires parents to complete an application disclosing financial information. Therefore, it is likely there are students who would qualify for free or reduced lunch but did not

apply to participate in this program. As a result, it is assumed not all students living in low-income environments were correctly identified and included in the resource availability independent variable. There is a movement within the field of educational research to move away from the use of free or reduced lunch status as an indicator of SES. The National Forum on Educational Statistics (2015) identified three concerns when free or reduced lunch status is used as an indicator of SES. These concerns include inaccurate use and interpretation of free or reduced lunch data, limited access to free or reduced lunch status data, and the consideration of only household income when determining economic need.

Recommendations for Future Research

Future research is needed to expand upon the findings of this study and more globally, universal screening for mathematical deficits. Several areas of which require further research were highlighted by this study and comprehensive literature review. Topics that appear to be the most imperative relate to implementation of universal screening in math and include classification accuracy, identification of general outcome measures in mathematics, and gated evaluation systems.

Technical Adequacy and Classification Accuracy

Classification accuracy for universal screeners for mathematical deficits has been identified as an area which requires more research (Glover & Albers, 2007; VanDerHeyden, 2010; VanDerHeyden, 2011). The results of the present study indicate computation probes have a strong correlation with performance on the state administered achievement test administered two years later, while in third grade. It is recommended that studies such as this one further the usefulness of universal screening data by establishing appropriate cut-off scores for decision-making purposes. There is a significant amount of debate in the educational field regarding what

the cut-off is for a student being at-risk for academic failure (Christ & Nelson, 2014; VanDerHeyden, 2011). More research is needed to define this term and give educators some practical guidelines for determining at what point is a student really at risk.

In addition to appropriately identifying students who are at-risk for mathematical deficits, systems should also identifying students who are high achieving. The purpose of MTSS is to provide high quality instruction to all students. This primary function of MTSS combined with the country's lagging performance in the STEM fields make it important to consider how universal screening can improve the education for high achieving as well as low achieving students. More research is needed to determine if universal screening tools can be used to identify underachieving students for intervention but also students who would benefit from curriculum acceleration or curriculum compacting and enrichment in STEM.

General Outcome Measures in Mathematics

The results of this study indicate computation probes in first, second, and third grade are strongly predictive of third grade mathematical outcomes. It is recommended that similar studies be replicated with more diverse populations across a variety of locations to confirm the findings of this present study.

General outcome measures in math is an area that requires more research. The majority of research regarding universal screening is relatively recent, with most studies having been published after 2000 (Shin & Bryant, 2015). There is a fair amount of research attempting to identify the characteristics of students with mathematical deficits, both those identified as having a learning disability and those categorized as low achieving. Common deficits that have been observed include poor mathematical computation, weak retrieval of basic math facts, inefficient counting strategies, poor number sense, attention deficits, and weaknesses in working memory

(Geary, 2004; Geary, et al., 2012; Jordan & Hanich, 2003; Martin et al., 2012). While the correlation between these characteristics and mathematical deficits can vary based on the time of administration and student age, it would be a strong starting point to develop general outcome measures for the purpose of identifying students who are likely to need additional support in mathematics on top of the generic curriculum all students access.

Christ and Vining (2006) indicate that curriculum-based computation tools can be used as a general outcome measure. This is supported by the current study, but requires replication across a more diverse population and multiple locations. In order to support the use of computation skills as a general outcome measure in mathematics, studies similar to the present one should be replicated with computation measures other than MBSP-C and different well-established criterion measures.

Based on the results of this study, MBSP-C demonstrates a moderate predictive relationship with performance on PSSA-M subtests with low face validity with computation skills. There is some evidence that computation measures do not correlate highly with mathematical outcomes for students at the upper elementary and secondary levels. However, measures of concept and application do correlate with future mathematical outcomes for older students (Anselmo, 2014). Preliminary research using fractions to engage in problem solving also correlated highly with future mathematical outcomes (Hansen, Jordan & Rodrigues, 2015). More research is needed to confirm initial findings and to determine at what grade-level, computation ceases to demonstrate a strong correlation with future math outcomes.

Initial findings support the use of computation fluency skills as a general outcome measure for mathematical performance, especially in a gated evaluation system. However, given the relatively weak predictive relationship between MBSP-C in the fall of first grade with future

mathematical outcomes and questions regarding the appropriateness of computation measures to determine the need for additional intervention for older students, more research is recommended to determine how computation measures would fit into alterative universal screening models such as gated evaluation systems and threshold decision making models.

Gated Evaluation Systems

Gated evaluation is defined the process of "involving multiple assessments that cost efficiently identify a subset of individuals from a larger pool of target participants with a combination of methods and measures generally arranged in sequential order" (Walker, Small, Severson, Seeley, & Feil, 2014, p. 47). In gated evaluation systems, all students may complete a simple, broadband screener. A small proportion of those students, as identified on the initial broadband screener, are gated into the next stage of screening. In this next stage of the gated system, the small number of students complete a narrowband assessment. This is especially relevant when screening for mathematical deficits due to the complexity of mathematical skills.

For example, discrete skill measures are generally fluency based and designed to assess individual components and specific mathematical skills (Purpura, Reid, Eiland, & Baroody, 2015). These measures demonstrate good predictive validity and are sensitive to change over time (VanDerHeyden, Broussard, & Cooley, 2006). Given the moderate to strong predictive relationship with overall math achievement, measured by PSSA-M Composite scores, a computation measure would be an appropriate tool to use as the first step of a gated evaluation model.

However, two potential concerns or limitations when discrete mathematical measures are used to determine future academic risk were identified. Fluency-based mathematical measures correlate highly with non-math related measures such as reading fluency and school readiness

measures (Polognano & Hojnoski, 2012; VanDerHeyden, Broussard, Fabre, Stanley, Legendre & Creppell, 2004). This suggests fluency-based measures are assessing non-mathematical constructs in addition to math skills. These potential limitations were addressed with the use of a measure that sampled a broad range of mathematical skills and use of a gated evaluation system.

Automaticity of basic math facts and computation fluency are required for successfully engaging in higher level mathematical thinking and problem-solving. However, proficiency on computation measures does not always indicate proficiency with high-level mathematics (Locuniak & Jordan, 2008). Therefore, brief computation probes such as, MBSP-C, may be a good option for school systems to use as the first step in a gated evaluation system. While more research is needed, initial findings support gated evaluation procedures as an accurate method for identifying students at risk (Albers & Kettler, 2014; Fuchs, Compton, et al., 2011; VanDerHeyden, 2010).

Threshold Decision-Making Models

Another area which requires additional research is the use of threshold decision-making models. Threshold decision-making models may be considered in populations where a large percentage of students are not demonstrating academic proficiency. Threshold decision-making models are prevalent in the medical field but have not yet translated into educational practice. Medically, threshold decision-making is used to determine whether screening and/or intervention should be initiated based on the probability of being asymptomatic, probability of negative side effects for any given age for participation or lack of participation in screening, and probability of death (Hoffman et al., 2006).

VanDerHeyden (2013) cautioned against universally screening all students when other data sources indicate a majority of students are demonstrating academic, emotional, and/or behavioral difficulties. In educational systems that have a large percentage of students identified as at risk, universal screening systems are no longer effective and efficient. Typically, when over 20% of the student population is demonstrating a need, it is considered a systemic deficit and systemic interventions are recommended. The use of threshold decision making is promoted to take into account contextual factors that impact student outcomes. In an education setting, threshold decision-making would require educators to consider the probability of a false negative, probability of a false positive, probability of a false result for students who will not fail (specificity), benefit of intervention for students who will fail, risk of intervention for students who will not fail, and risk-of-test.

Potential benefits of this model include less strain on school resources and instructional time to conduct universal screenings with all students and decreased risk of flooding Tier 2 intervention with false positives. Potential downfalls include loss of universal screening data to evaluate the effectiveness of the educational system and potential for false negatives (VanDerHeyden, 2013). Empirical and longitudinal research is needed to determine whether or not threshold decision-making is applicable within an educational setting. A student's socio-economic status, or resource availability, falls under the purview of contextual factors considered when making decisions about universal screening systems.

Resource Availability

The results of the present study suggest resource availability, measured by participation in the free or reduced lunch program, has a weak but statistically significant correlation with the criterion measure, PSSA-M in third grade. This weak correlation increases slightly from first to

second and second to third grade. This suggests resource availability becomes a more relevant factor the longer students are in school. However, more research is needed to determine whether the impact resource availability continues to increase, plateaus, or begins to decrease.

The results of the present study suggest resource availability becomes more influential to math in the latter years. Previous research identified the achievement gap between students living in low-income households and those in high-income households decreased over the course of the school year. However, deficits were re-established over the summer months (Reardon, 2013). It would be beneficial to further explore the findings of Reardon (2013) by providing mathematical instruction to students living in low-income environments over the summer months to determine whether or not it mitigates regression of mathematical achievement observed in students from low-income households over the summer months.

More research is also needed to determine if the findings of the present study would differ if resource availability was determined in a more comprehension manner instead of free or reduced lunch status. The National Forum on Educational Statistics (2015) recommends eight alternative measures to be used to determine SES or resource availability. They are as follows, (a) eligibility for other means-tested programs, (b) information provided by the household, (c) student or family categorical status, (d) household income, (e) highest level of education completed by parent/guardian, (f) parent/guardian occupation, (g) SES of the neighborhood/community, and (h) school district poverty estimate. According to the National Forum on Educational Statistics, when these eight factors provide a more meaningful and accurate measure of SES than participation in the free or reduced lunch status alone. It is highly recommended future research use these eight factors or a combination of them to determine SES.

Implications for Practice

As MTSS service delivery models continue to become more prevalent, school psychologists play an important role to aid systems in developing efficient and effective methods for identifying students at-risk of academic, social, and emotional deficits. School psychologists and educational leadership should advocate for the development of data analysis teams and be fluent in analysis of data used by teams at a systemic, small group, and individual level. As members of data analysis teams, school psychologists are able to evaluate and improve programs on a systemic, small group, and individual student level with the use of universal screening data. In this way, school psychologists play a unique role when determining what tools should be used for universal screening and interpretation of universal screening data.

When choosing a mathematical universal screening tool, school psychologists should advocate for consideration of the following criteria, (a) is it suitable for the proposed use, (b) is a good fit for local needs, (c) is aligned with valued criterion, (d) is supported by research, (e) is a good fit for the intended population, (f) is able to demonstrate sufficient technical adequacy, (g) is user friendly, and (h) the data generated is easily understood and meaningful (Albers & Kettler, 2014).

If school psychologists are working within a system that chooses to administer computation probes as a universally screener, it is important they be well versed in the strengths and weakness of this tool. Advantages of MBSP-C include the low cost, ease of administration and scoring, number of alternative probes, technical adequacy, and moderate to high correlations with future mathematical outcomes from first through third grade.

Concerns with MBSP-C include the outdated published normative data. It is recommended local normative data be developed. School psychologists will need to be capable

of producing, analyzing, and distributing this data for use in a clear and understandable manner. When developing local normative data, it is recommended the sample consist of a minimum of 100 students per administration and should be updated every five to seven years (Stewart & Silberglitt, 2008). This may become problematic in small school districts.

Educational leadership should consider using computation probes as the first gate in a gated evaluation system. Students who demonstrate low performance on the computation probe would move to a second gate evaluation, such as a curriculum-based evaluation (CBE). CBE is a process used to identify specific skill deficits in order to drive instruction and intervention. CBE is defined as, "a problem-solving process used to determine what to teach and how to teach" (Kelley, 2008, p. 423). The first step of a CBE is problem identification. Followed by hypothesis generation, problem analysis, plan development, and implementation (Howell, Hosp, & Kurns, 2008; Kelley, 2008).

Sex and resource availability made negligible contributions to the total variance in PSSA-M Composite and subtest performance. It is recommended that these data are neither collected nor considered as part of universal screening procedures. However, school psychologists and educational leadership should remain current regarding future research related to the impact of resource availability on mathematical learning outcomes in order to revisit their universal practices as more information becomes available.

It is also recommended that school psychologists and educational leadership continue to review research regarding general outcome measures in mathematics as more becomes known about the relationship between early numeracy skills, computation skills, and concept and application skills as students advance through school. It is important to understand how these tools relate to future mathematical learning outcomes to inform decisions regarding universal

screening models and the data generated by these tools. When evaluating measures to utilize as universal screeners in mathematics, it is critical to remain focused on the purpose of these tools. It is meaningless to be able to very accurately and efficiently identify students in need of intervention if nothing is done with this information. Students need to have access to high quality math instruction within Tier 1, and supplemental intervention in Tiers 2 and 3. This continues to be an area of need in which school psychologists are able to provide guidance and evaluate program effectiveness on a systemic level.

Summary

The findings of this study support the use of computation measures as universal screening tool for students in first through third grade. MBSP-C probes in the fall, winter, and spring of first, second, and third grade demonstrate moderate to strong correlations with performance on the PSSA-M administered in the spring of third grade. Sex and resource availability had little impact on the overall variance accounted for in the prediction model as determined by multiple linear regression analyses and Pearson correlations.

This chapter highlighted the results of the statistical analysis and acceptance or rejection of the hypotheses. This discussion was followed by limitations to this study, which stem primarily from the use of a convenience sample and archival data. The use of a convenience sample impacts generalizability because the population lacked diversity. The use of archival data did not allow for control of administration procedures, scoring, data entry, and quality of math instruction. Study limitations were followed by suggestions for addition research and implications for school psychologists. Additional research and implications for school psychologists focus on the development and use of general outcome measures for math, technical adequacy of universal screening tools, and gated evaluation systems.

References

Albers, C. A., & Kettler, R. J. (2014). Best practices in universal screening. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 121-131). Bethesda, MD: National Association of School Psychologists.

Aldrich, J. O. & Cunningham, J. B. (2016). *Using IBM SPSS statistics* (2nd ed). Thousand Oaks, CA: SAGE Publications, Inc.

Allinder, R. M., Bolling, R. M., Oats, R. G., & Gagnon, W. A. (2000). Effects of teacher self-monitoring on implementation of curriculum-based measurement and mathematics computation achievement of students with disabilities. *Remedial and Special Education, 21,* 219-226. doi:10.1177/0741932510361265

Anderson, D., Lai, C., Alonzo, J., & Tindal, G. (2011). Examining a grade-level math CBM designed for persistently low-performing students. *Educational Assessment, 16,* 15-34. doi:10.1080/10627197.2011.551084

Anselmo, G.A. (2014). *Criterion validity of mathematics curriculum-based measurement* (Doctoral dissertation). Retrieved from http://hdl.handle.net.proxy-iup.klnpa.org/2069/2263

Aud, S., Fox, M., & KewalRamani, A. (2010). *Status and Trends in the Education of Racial and Ethnic Groups* (NCES 2010-015). National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Bachman, H. J., Votruba-Drzal, E., El Nokali, N. E., & Heatly, M. C. (2015). Opportunities for learning in elementary school: Implications for SES disparities in procedural and conceptual math skills. *American Educational Research Journal, 52*, 894-923. doi:10.3102/0002831215594877

Bailey, D. H., Watts, T. W., Littlefield, A. K., & Geary, D. C. (2014). State and trait effects on individual differences in children's mathematical development. *Psychological Science, 25,* 2017-2026. doi:10.1177/0956797614547539

Berninger, V. W. (2006). Research-supported ideas for implementing reauthorization IDEA with intelligent professional psychological services. *Psychology in the Schools, 43*, 782-796. doi: 10.1002/pits.20188

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology, 42,* 189-201. doi: 10.1037/00121649.41.6.189

Braden, J. P., & Schroeder, J. L. (2004). High-stakes testing and no child left behind: Information and strategies for educators. *Helping Children at Home and School II: Handouts for Families and Educators.* (pp. 73–77). Bethesda, MD: National Association of School Psychologists.

Braithwaite, D. W., Goldstone, R. L., van der Maas, H. L. J., & Landy, D. H. (2016). Non-formal mechanisms in mathematical cognitive development: The case of arithmetic. *Cognition, 149,* 40-55. doi: 10.1016/j.cognition.2016.01.004

Bricker, D., Yovanoff, P., Capt, B., & Allen, D. (2003). Use of a curriculum-based measurement to corroborate eligibility decisions. *Journal of Early Intervention, 26,* 20-30. doi: 10.1177/1053815108324422

Brown, J. D. (1997, April). Skewness and kurtosis. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *1*, 20-23.

Burns, M. K. (2010). Formative evaluation in school psychology: Fully informing the instructional practice. *School Psychology Forum: Research in Practice, 4,* 22-33.

Burns, M. K., Haegele, K., & Petersen-Brown, S. (2014). Screening for early reading skills: Using data to guide resources and instruction. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 171-198). Washington, DC: American Psychological Association. doi: 10.1037/14316-000

Byrne, B. M. (2010). *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming.* New York, NY: Routledge.

Christ, T. J., & Nelson, P. M. (2014). Developing and evaluating screening systems: Practical and psychometric considerations. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 79-110). Washington, DC: American Psychological Association. doi: 10.1037/14316-000

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

Christ, T. J., Johnson-Gros, K. N., & Hintze, J. M. (2005). An examination of alternate assessment durations when assessing multiple-skill computation fluency: The generalizability and dependability of curriculum-based outcomes within the context of educational decisions. *Psychology in the Schools, 42,* 615-622. doi: 10.1002/pits.20107

Christ, T. J., Scullin, A., Tolbize, A. & Jiban, C. L. (2008). Curriculum-based measurement of math computation. *Assessment for Effective Intervention, 33,* 198-205. doi: 10.1177/1534508407313480

Chu, F. W., vanMarle, K., & Geary, D. C. (2015). Early numerical foundations of young children's mathematical development. *Journal of Experimental Child Psychology, 132,* 205-212. doi: 10.1016/j.jecp.2015.01.006

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

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

Clarke, B., Doabler, C. T., & Nelson, N. J. (2014). Best practices in mathematics assessment and intervention with elementary students. In P. L. Harrison & A. Thomas (Eds.). *Best practices in school psychology: Data-based and collaborative decision making* (pp. 219-232). Bethesda, MD: National Association of School Psychologists.

Clarke, B., Haymond, K., & Gersten, R. (2014). Mathematics screening measures for the primary grades. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 199-222). Washington, DC: American Psychological Association. doi: 10.1037/14316-000

Clarke, B., Nese, J. F. T., Alonzo, J., Smith, J. L. M., Tindal. G., Kame'enui, E. J., & Baker, S. (2011). Classification accuracy of easyCBM first-grade mathematics measures: Findings and implications for the field. *Assessment for Effective Intervention, 36,* 243-255. doi: 10/1177/1534508411414153

Codding, R. S., Petscher. Y., & Truckenmiller, A. (2015). CBM reading, mathematics, and written expression at the secondary level: Examining latent composite relations among indices and unique predictions with state achievement. *Journal of Educational Psychology, 107,* 437-450. doi: 10.1037/a0037520

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* Routledge

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Cowan, K. C., Vaillancourt, K., Rossen, E., & Pollitt, K. (2013). *A framework for safe and successful schools* [Brief]. Bethesda, MD: National Association of School Psychologists.

Crawford, L. & Tindal, G. (2006). Policy and practice: Knowledge and beliefs of education professionals related to the inclusion of students with disabilities in a state assessment. *Remedial and Special Education, 27,* 208-217.

Current Population Survey. (2015). Annual Social and Economic Supplement. United States Census. Retrieved from https://www.census.gov/hhes/www/poverty/publications/pubscps.html

Daly, E. J., III, Martens, B. K., Barnett, D., Witt, J. C., & Olson, S. C. (2007). Varying intervention delivery in response to intervention: Confronting and resolving challenges with measurement, instruction, and intensity. *School Psychology Review, 36,* 562-581.

Data Recognition Corporation. (2014). *Technical report for the PSSA*. Maple Grove, MN: Author. Retrieved from http://www.education.pa.gov/Documents/K12/Assessment%20and%20Accountability/PSSA/Technical%20Reports/2014%20PSS%20Technical%20Report.pdf

Decker, S. L, & Roberts, A.M. (2015). Specific cognitive predictors of early math problem solving. *Psychology in the Schools, 52,* 477-488. doi: 10.1002/pits.21837

Dehaene, S., Molko, N., Cohen, L., & Wilson, A. J. (2004). Arithmetic and the brain. *Current Opinion in Neurobiology, 14,* 218-224. doi: 10.1016/j.conb.2004.03.008

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science, 284,* 970-974. doi: 10.1126/science.284.5416.970

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

De Santana, J. A., & Galera, C. (2014). Visual-spatial and verbal-spatial binding in working memory. *Psychology & Neuroscience, 7,* 399-406. doi: 10.3922/j.psns.2014.048

Desimone, L. M., & Long, D. (2010). Teacher effects and the achievement gap: Do teacher and teaching quality influence the achievement gap between Black and White and high- and low-SES students in the early grades? *Teachers College, 112,* 3024-3073.

DiPerna, J. C., Bailey, C. G., & Anthony, C. (2014). Broadband screening of academic and social behavior. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 223-248). Washington, DC: American Psychological Association. doi: 10.1037/14316-000

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43,* 1428-1446. doi: 10.1037/0012-1649.43.6.1428

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. doi: 10.1016/j.jsp.2006.11.002

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*, 103-127. doi: 10.1037/a0018053

Erturan, S., & Jansen, B. (2015). An investigation of boys' and girls' emotional experience of math, their math performance, and the relation between these variables. *European Journal of Psychology of Education, 30,* 421-435. doi: 10.1007/s10212-015-0248-7

Every Child a Chance Trust. (2009). *The long-term costs of numeracy difficulties*. Retrieved from http://www.shinetrust.org.uk/wp-content/uploads/ECC-Long-Term-Costs-Numeracy.pdf

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). London, England: SAGE Publications, Inc.

Fisher, P. H., Dobbs-Oates, J., Doctoroff, G. L., & Arnold, D. H. (2012). Early math interest and development of math skills. *Journal of Educational Psychology, 104,* 673-681. doi: 10.1037/a0027756

Foegen, A., & Deno, S. L. (2001). Identifying growth indicators for low-achieving students in middle school students. *The Journal of Special Education, 35,* 4-16. doi: 10.1177/002246690103500102

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

Frey, A. J., Lingo, A., & Nelson, C. M. (2010). Implementing positive behavior support in elementary schools. In M. R. Shinn & H. M. Walker (Eds.), *Interventions: For achievement and behavior problems in a three-tier model including RTI* (pp. 397-434). Bethesda, MD: National Association of School Psychologists.

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

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57,* 488-498. doi: 10.1177/001440299105700603

Fuchs, L. S., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *School Psychology Review, 28,* 659-671.

Fuchs, L. S., & Fuchs, D. (2005). Enhancing mathematical problem solving for students with learning disabilities. *The Journal of Special Education, 39,* 45-57. doi:10.1177.00224669050390010501

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2012). The early prevention of mathematics difficulty: Its power and limitation. *Journal of Learning Disabilities, 45,* 257-269. doi: 10.1177/0022219412442167

Fuchs, L. S., Fuchs, D., & Courey, S. J. (2005). Curriculum-based measurement of mathematics competence: From computation to concepts and applications to real-life problem solving. *Assessment for Effective Intervention, 30,* 33-46. doi: 10.1177/073724770503000204

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

Fuchs, L. S., Seethaler, P. M., Fuchs, D., & Hamlett, C. L. (2008). Using curriculum-based measurement to identify the 2% population. *Journal of Disability Policy Studies, 19,* 153-161. doi: 10.1177/1044207308327471

Fuchs, L. S., Compton, D. L., Fuchs, D. Paulsen, K., Bryant, J., & Hamlett, C. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology, 97,* 493-513. doi: 10.1037/0022-0663.97.3.493

Fuchs, L. S., Compton, D. L., Fuchs, D., Hollenbeck, K. N., Hamlett, C. L., & Seethaler, P. M. (2011). Two-stage screening for math-problem-solving difficulty using dynamic assessment of algebraic learning. *Journal of Learning Disabilities, 44,* 372-380. doi: 10.1177/0022219411407867

Fuchs, L. S., Fuchs, D., & Courey, S. J. (2005). Curriculum-based measurement of mathematics competence: From computation to concepts and applications to real-life problem solving. *Assessment for Effective Intervention, 30,* 33-46. doi: 10.1177/073724770503000204

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

Fuchs, L. S., Seethaler, P. M., Fuchs, D., & Hamlett, C. L. (2008). Using curriculum-based measurement to identify the 2% population. *Journal of Disability Policy Studies, 19,* 153-161. doi: 10.1177/1044207308327471

Geary, D. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities, 37*, 4-15. doi: 10.1177/00222194040370010201

Geary, D. C., Bailey, D. H., & Hoard, M. K. (2009). Predicting mathematical achievement and mathematical learning disability with a simple screening tool: The number sets test. *Journal of Psychoeducational Assessment, 27,* 265-279. doi: 10.1177/0734282908330592

Geary, D. C., Hoard, M. K., & Bailey, D. H. (2012). Fact retrieval deficits in low achieving children and children with mathematical learning disabilities. *Journal of Learning Disabilities,* 45, 291-307. doi: 10.1177/0022219410392046

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research, 79,* 1202-1242. doi:10.3102/0034654309334431

Gersten, R., Clarke, B., Jordan, N., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children, 78,* 423-445. doi: 10.1177/001440291207800403

Gersten, R., & Jordan, N. (2005). Early screening and intervention in mathematics difficulties: the need for action. Introduction to the special series. *Journal of Learning Disabilities, 38*, 291-292.

Gersten, R., Jordan, N., & Flojo, J. (2005). Early identification and interventions for students with math difficulties. *Journal of Learning Disabilities, 38,* 293-304.

Gersten, R., Rolfhus, E., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2015). Intervention for first graders with limited number knowledge: Large-scale replication of randomized controlled trial. *American Educational Research Journal, 52,* 516-546. doi: 10.3102/0002831214565787

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45,* 117-135. doi: 10.1016/j.jsp.2006.05.005

Glover, T. A., & DiPerna, J. C. (2007). Service delivery for response to intervention: Core components and directions for future research. *School Psychology Review, 36,* 526-540.

Gómez-Velázquez, F. R., Berumen, G., & González-Garrido, A. A. (2015). Comparisons of numerical magnitudes in children with different levels of mathematical achievement: An ERP study. *Brain Research, 1629,* 189-200. doi: 10.1016/j.brainres.2015.09.009

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., & Brenwald (2009). Highlights from TIMSS 2007: Mathematics and science achievement of U.S. fourth- and eighth grade students in an international context. Retrieved from http://nces.ed.gov/pubs2009/2009001.pdf

Gresham, F. M. (2006). Response to intervention. In G. G. Bear & K. M. Minke (Eds.), *Children's needs III: Development, prevention, and intervention* (pp. 525-540). Bethesda, MD: National Association of School Psychologists.

Gresham, F., Reschley, D., & Shinn, M. R. (2010). RTI as a driving force in educational improvement: Research, legal, and practice perspectives. In M. R. Shinn & H. M. Walker (Eds.), *Interventions: for achievement and behavior problems in a three-tier model including RTI* (pp. 47-77). Bethesda, MD: National Association of School Psychologists.

Hansen, N., Jordan, N. C., & Rodrigues, J. (2015). Identifying learning difficulties with fractions: A longitudinal study of student growth from third through sixth grade. *Contemporary Educational Psychology*. doi: 10.1016/j.cedpsych.2015.11.002

Hassinger-Das, B., Jordan, N. C., Glutting, J., Irwin, C., & Dyson, N. (2014). Domain-general mediators of the relation between kindergarten number sense and first-grade mathematics achievement. *Journal of Experimental Child Psychology, 118,* 78-92. doi: 10.1016/j.jecp.2013.09.008

Hintze, J. M., Christ, T. J., & Methe, S. A. (2006). Curriculum-based assessment. *Psychology in the Schools, 43,* 45-56. doi: 10.1002/pits.20128

Hoffman, J. R., Wilkes, M. S., Day, F. C., Bell, D. S., & Higa, J. K. (2006). The roulette wheel: An aid to informed decision making. *PLoS Med, 3,* 743-748. doi: 10.1371/journal.pmed.0030137

Howell, K. W., Hosp, J. L., & Kurns, S. (2008). Best practices in curriculum-based evaluation. In A. Thomas & J. Grimes (Eds). *Best practices in school psychology V* (pp. 349-362). Bethesda, MD: National Association of School Psychologists.

Huck, S. W. (2008). *Reading statistics and research* (5th ed). Boston, MA: Pearson Education.

Huck, S. W. (2014). *Reading statistics and research* (6th ed). London, England: Pearson Education Limited.

Huefner, D. S. (2006). *Getting comfortable with special education law: A framework for working with children with disabilities* (2nd ed.). Norwood, MA: Christopher-Gordon Publishers.

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., & Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect. *Psychology of Women Quarterly, 14*, 299-324. doi: 10.1111/j.1471-6402.1990.tb00022.x

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

Ikeda, M. J., Neesen, 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.

Ikeda, M. J., Paine, S. C., & Elliott, J. L. (2010). Supporting response to intervention (RTI) at school, district, and state levels. In M. R. Shinn & H. M. Walker (Eds.), *Interventions: for achievement and behavior problems in a three-tier model including RTI* (pp. 27-46). Bethesda, MD: National Association of School Psychologists.

Individuals with Disabilities Education Improvement Act. (2004). Federal Regulations Part 300. Retrieved from http://idea.ed.gov/download/statute.html

Jacobs, S. & Hartshorne, T.S. (2003). *Ethics and law for school psychologists* (4th ed.). Hoboken, NJ: John Wiley & Sons.

January, S. A., & Ardoin, S. P. (2015). Technical adequacy and acceptability of curriculum-based measurement and the measures of academic progress. *Assessment for Effective Intervention, 41,* 3-15. doi: 10.1177/1534508415579095

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36,* 582-600.

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

Johnson, K. R., & Layng, T. J. (1992). Breaking the structuralist barrier: Literacy and numeracy with fluency. *American Psychologist, 47,* 1475-1490. doi: 10.1037/0003- 066X.47.11.1475

Jordan, N. C., Glutting, J., & Ramineni, C. (2008). A number sense assessment tool for identifying children at risk for mathematical difficulties. In A. Dowker (Ed.). *Mathematical difficulties: Psychology and intervention* (pp. 45-58). San Diego, CA: Academic Press.

Jordan, N. C., Glutting, J., Ramineni, C., & Watkins, M. W. (2010). Validating a number sense screening tool for use in kindergarten and first grade: Prediction of mathematics proficiency in third grade. *School Psychology Review, 39,* 181 – 195.

Jordan, N. C., & Hanich, L. B. (2003). Characteristics of children with moderate mathematical deficiencies: A longitudinal perspective. *Learning Disabilities Research & Practice, 18,* 213-221. doi: 10.1111/1540-5826.00076

Jordan, N. C., Kaplan, D., & Hanich, L. B. (2002). Achievement growth in children with learning difficulties in mathematics: Findings of a two-year longitudinal study. *Journal of Educational Psychology, 94,* 586-597. doi: 10.1037//0022-0663.94.3.586

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). A longitudinal study of mathematical competencies in children with specific mathematical difficulties verse children with comorbid mathematical and reading difficulties. *Child Development, 74,* 834 – 850.

Jordan, N. C., Kaplan, D., & Hanich, L. B. (2002). Achievement growth in children with learning difficulties in mathematics: Findings of a two-year longitudinal study. *Journal of Educational Psychology, 94,* 586-597. doi: 10.1037//0022-0663.94.3.586

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematical outcomes. *Developmental Psychology, 45,* 850-867. doi: 10.1037/a0014939

Kamphaus, R. W., Reynolds, C. R., & Dever, B. V. (2014). Behavioral and mental health screening. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 249-274). Washington, DC: American Psychological Association. doi: 10.1037/14316000

Kelley, B. (2008). Best practices in curriculum-based evaluation and math. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 419-438). Bethesda, MD: National Association of School Psychologists.

Keller-Margulis, M. A., Mercer, S. H., & Shapiro, E. S. (2014). Differences in growth on math curriculum-based measures using triannual benchmarks. *Assessment for Effective Intervention, 39,* 146-155. doi: 10.1177/1534508412452750

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

Kettler, R. J, Glover, T. A., Albers, C. A., & Feeney-Kettler, K. A. (2014). An introduction to universal screening in educational settings. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 3-16). Washington, DC: American Psychological Association. doi: 10.1037/14316-000

Kincaid, D., Dunlap, G., Kern, L., Lane, K. L., Bambara, L. M., Brown, F., . . . Knoster, T. P. (2016). Positive behavior support: A proposal for updating and refining the definition. *Journal of Positive Behavior Interventions, 18,* 69-73. doi: 10.1177/1098300715604826

Kovaleski, J. F., & Pedersen, J. A. (2014). Best practices in data-analysis teaming. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 99-120). Bethesda, MD: National Association of School Psychologists.

Kovaleski, J. F., & Pedersen, J. A. (2008). Best practices in data-analysis teaming. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 115-129). Bethesda, MD: National Association of School Psychologists.

Kovaleski, J. F., VanDerHeyden, A. M., & Shapiro, E. S. (2013). *The RTI approach to evaluating learning disabilities*. New York, NY: The Guilford Press.

Laerd Statistics, Inc (2015). *Multiple regression using SPSS Statistics: Statistical tutorials and software guides.* Retrieved from https://statistics.laerd.com/premium/spss/mr/multiple-regression-in-spss.php

Landerl, K., & Kölle, C. (2009). Typical and atypical development of basic numerical skills in elementary school. *Journal of Experimental Child Psychology, 103,* 546-565. doi: 10.1016/j/jecp.2008.12.006

LeFevre, J.-A., Fast, L., Skwarcuk, S.-L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development, 81,* 1753-1767. doi: 10.1111/j.1467-8624.2010.01508.x

Lehrl, S., Kluczniok, K., & Rossbach, H.-G. (2016). Longer-term associations of preschool education: The predictive role of preschool quality for the development of mathematical skills through elementary school. *Early Childhood Research Quarterly, 36,* 475-488. doi: 10.1016/j.ecresq.2016.01.013

Lembke, E., & Foegen, A. (2009). Identifying early numeracy indicators for kindergarten and first-grade students. *Learning Disabilities Research and Practice, 24,* 12-20. doi: 10.1111/j.1540-5826.2008.012373.x

Lillenstein, J., Fritschmann, N., & Moran, L. (2012). Pennsylvania's secondary RTII initiative impact of a multi-tiered system of support in five middle schools. *Perspectives on Language and Literacy*, 38(2), 20

Locuniak, M. N., & Jordan, N. C. (2008). Using kindergarten number sense to predict calculation fluency in second grade. *Journal of Learning Disabilities, 41,* 451-459. doi: 10.1177/0022219408321126

Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L., & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1-6. *Developmental Sciences, 17,* 714-726. doi: 10.1111/desc.12152

Martin, R. B., Cirino, P. T., Barnes, M. A., Ewing-Cobbs, L., Fuchs, L. S., Stuebing, K. K., & Fletcher, J. M. (2012). Prediction and stability of mathematics skill and difficulty. *Journal of Learning Disabilities, 46,* 428-443. doi: 10.1177/0022219411436214

Mazzocco, M. M. M. (2003). Challenges in identifying target skills for math disabilities screening and intervention. *Journal of Learning Disabilities, 38,* 318 – 323.

Mazzocco, M. M. M. & Thompson, R .E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research and Practice, 20,* 142-155. doi: 10.1111/j.1540-5826.2005.00129.x

Mazzocco, M. M. M., Devlin, K. T., & McKenney, S. J. (2008). Is it a fact? Timed arithmetic performance of children with mathematical learning disability (MLD) varies as a function of how MLD is defined. *Developmental Neuropsychology, 33,* 318-344. doi: 10.1080/87565640801982403

MacPhee, D., Farro, S., & Canetto, S. S. (2013). Academic self-efficacy and performance of underrepresented STEM majors: Gender, ethnic and social class patterns. *Analyses of Social Issues and Public Policy, 13,* 347-369. doi: 10.1111/asap.12033

McGraw, R., Lubienski, S. T., & Strutchen, M. E. (2006). A closer look at gender in NAEP mathematics achievement and affect data: Intersections with achievement, race/ethnicity, and socioeconomic status. *Journal for Research in Mathematics Education, 37*, 129-150. doi: 10.2307/30034845

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749. doi: 10.1037/0003-066x.50.9.741

Methe, S. A., Begeny, J. C., & Leary, L. L. (2011). Development of conceptually focused early numeracy skill indicators. *Assessment for Effective Intervention, 36,* 230-242. doi: 10.1177/1534508411414150

Methe, S. A., Briesch, A. M., & Hulac, D. (2015). Evaluating procedures for reducing measurement error in math curriculum-based measurement probes. *Assessment for Effective Intervention, 40,* 99-113. doi: 10.1177/1534508414553295

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

Morgan, P., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning disabilities in mathematics. *Journal of Learning Disabilities, 42*, 306-321. doi: 10.1177/0022219408331037

Morgan, P. L., Farkas, G., & Wu, Q. (2011). Kindergarten children's growth trajectories in reading and mathematics: Who falls increasingly behind? *Journal of Learning Disabilities, 44,* 472-488. doi: 10.1177/0022219411414010

Murphy, M. M., Mazzocco, M. M. M., Hanich, L. B., & Early, M. C. (2007). Cognitive characteristics of children with mathematics learning disability (MLD) vary as a function of cutoff criterion used to define MLD. *Journal of Learning Disabilities, 40,* 458-478.

Namkung, J. M., & Fuchs, L. S. (2012). Early numerical competencies of students with different forms of mathematics disabilities. *Learning Disabilities Research and Practice, 27,* 2-11. doi: 10.1111/j.1540-5826.2011.00345.x

National Academy of Education. (2009). *Education policy white paper on standards, assessments, and accountability*. Washington, DC: Government Printing Office. Retrieved from http://www.naeducation.org/cs/groups/naedsite/documents/webpage/naed_080866.pdf

National Council of Teachers of Mathematics. (2013). *Supporting the common core state standards for mathematics*. Retrieved from http://www.nctm.org/uploadedFiles/Standards_and_Positions/Position_Statements/Common%20Core%20State%20Standards.pdf

National Education Association. (2015). *Less testing = More learning ESSA fact sheet*. Retrieved from National Education Association website: http://www.nea.org/assets/docs/ESSA%20Fact%20Sheet%20-%20Testing%20121415.pdf

National Forum on Education Statistics (2015). *Forum guide to alternative measures of socioeconomic status in education data systems.* (NFES 2015-158). U. S. Department of Education. Washington, D.C.: National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2015/2015158.pdf

National Research Council. (2002). J. Kilpatrick & J. Swafford (Eds.), *Helping children learn mathematics.* Washington, DC: National Academy Press.

National Science Foundation, National Center for Science and Engineering Statistics. (2015). *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2015*. Special Report NSF 15-311. Arlington, VA.

Nellis, L. M. (2012). Maximizing the effectiveness of building teams in response to intervention implementation. *Psychology in the Schools, 49,* 245-256. doi: 10.1002/pits.21594

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2001). Washington, DC: U.S. Department of Education.

Osborne, J. W. (2013). Normality of residuals is a continuous variable, and does seem to influence the trustworthiness of confidence intervals: A response to, and appreciation of, Williams, Grajales, and Kurkiewicz (2013). *Practical Assessment, Research & Evaluation, 18*(12). Retrieved from http://pareonline.net/getvn.asp?v=18&n=12

Parisi, D. M., Ihlo, T., & Glover, T. A. (2014). Screening within a multi-tiered prevention model: Using assessment to inform instruction and promote students' response to intervention. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 19 46). Washington, DC: American Psychological Association. doi: 10.1037/14316-000

Pennsylvania Department of Education, Bureau of Assessment and Accountability. (2011). *District report card 2010 – 2011.* Retrieved from http://www.sgasd.org/cms/lib2/PA01001732/Centricity/Domain/24/2010- 2011.pdf

Pennsylvania State Data Center. (2012). *Special education data report school year 2011-2012* Retrieved from http://penndata.hbg.psu.edu/BSEReports/Data%20Preview/2011_2012/PDF_Document Speced_Quick_Report_S 238_Final.pdf

Pennsylvania Training and Technical Assistance Network. (2015). *Local education agencies (LEAs) that have applied for approval to use response to instruction and intervention (RtII) to determine specific learning disabilities.* Retrieved from http://pattan.net website.s3.amazonaws.com/images/instructional/2015/08/27/RtII_Status_Chart0815.pdf

Petscher, Y., Kim, Y-S., & Foorman, B. R. (2011). The importance of predictive power in early screening assessments: Implications for placement in the response to intervention framework. *Assessment for Effective Intervention, 36,* 158-166. doi: 10.1177/1534508410396698

Pintea, S., & Moldovan, R. (2009). The receiver-operating characteristic (roc) analysis: Fundamentals and applications in clinical psychology. *Journal of Cognitive and Behavioral Psychotherapies, 9*, 49-66.

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

Price, G. R, Mazzocco, M. M. M., & Ansari, D. (2013). Why mental arithmetic counts: Brain activation during single digit arithmetic predicts high school math scores. *The Journal of Neuroscience, 33,* 156-163. doi: 10.1523/JNEUROSCI.2936-12.2013

Purpura, D. J., & Logan, J. A. R. (2015). The nonlinear relations of the approximate number system and mathematical language to early mathematics development. *Developmental Psychology, 51,* 1717-1724. doi: 10.1037/dev0000055

Purpura, D. J., & Reid, E. E. (2016). Mathematics and language: Individual and group differences in mathematical language skills. *Early Childhood Research Quarterly, 36,* 259-268. doi: 10.1016/j.ecres.2015.12.020

Purpura, D. J., Reid, E. E., Eiland, M. D., & Baroody, A. J. (2015). Using a brief preschool early numeracy skills screener to identify children with mathematical difficulties. *School Psychology Review, 44*, 41-59. doi: 10.17105/SPR44-1.41.59

Raju, P. K., & Clayson, A. (2010). The future of STEM education: An analysis of two national reports. *Journal of STEM Education*, *11*, 25-28.

Reardon, S. F., & Bischoff, K. (2011). Income inequality and income segregation. *The American Journal of Sociology, 116,* 1092-1153. doi: 10.1086/657114

Reardon, S. F., Shores, K. A., Kalogrides, D., & Weathers, E. S. (2014). Patterns of achievement gaps among school districts: New data, new measures, new insights. *Society for Research on Educational Effectiveness.* Retrieved from http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED562809

Roth, W-M. (2008). Mathematical cognition and the final report of the national mathematics advisory panel: A critical, cultural-historical activity theoretic analysis. *The Montana Mathematics Enthusiast, 5,* 371-386. Retrieved from http://scholarworks.umt.edu/cgi/viewcontent.cgi?article=1116&context=tme

Sansosti, F. J., & Noltemeyer, A. (2008). Viewing response-to-intervention through an educational change paradigm: What can we learn? *Contemporary School Psychologist*, *13*, 55-66. doi: 10.1007/BF03340942

Scheiber, C., Reynolds, M. R., Hajovsky, D. B., & Kaufman, A. S. (2015). Gender differences in achievement in a large, nationally representative sample of children and adolescents. *Psychology in the Schools, 52*, 335-348. doi: 10.1002/pits.21827

Seethaler, P. M., & Fuchs, L. S. (2010). The predictive utility of kindergarten screening for math difficulty. *Exceptional Children, 77,* 37-59.

Shapiro, E. S., Dennis, M. S., & Fu, Q. (2015). Comparing computer adaptive and curriculum based measures of math in progress monitoring. *School Psychology Quarterly, 30,* 470-487. doi: 10.1037/spq0000116

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

Shin, M., & Bryant, D. P. (2015). A synthesis of mathematical and cognitive performances of students with mathematics learning disabilities. *Journal of Learning Disabilities, 48*, 96-112. doi: 10.1177/0022219413508324

Sisco-Taylor, D., Fung, W., & Swanson, H. L. (2015). Do curriculum-based measures predict performance on word-problem solving measures? *Assessment for Effective Intervention, 40,* 131-142. doi: 10.1177/1534508414556504

Speece, D. L., Case, L. P., & Molloy, D. E. (2003). Responsiveness to general education instruction as the first gate to learning disability identification. *Learning Disabilities Research & Practice, 18,* 147-156.

Statistics Solutions (2013). *Statistical Analysis: A manual on dissertation statistics in SPSS*. Retrieved from https://www.statisticssolutions.com/wp-content/uploads/2013/10/SPSS-Manual.pdf

Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research and Practice, 15,* 128-134. doi: 10.1207/SLDRP1503_2

Stein, M., Kinder, D., Zapp, K.., & Feuerborn, L. (2010). Promoting positive math outcomes. In M. R. Shinn & H. M. Walker (Eds.), *Interventions: For achievement and behavior problems in a three-tier model including RTI* (pp. 527-551). Bethesda, MD: National Association of School Psychologists.

Stevens-Olinger, E. L. (2014). *Mathematics curriculum based measurement to predict state test performance: A comparison of measures and methods* (Doctoral dissertation). Retrieved from ProQuest. (3554952)

Stewart, L. H., & Silberglitt, B. (2008). Best practices in developing academic local norms. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 225-242). Bethesda, MD: National Association of School Psychologists.

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within- and across-nation assessment of 10 years of PISA data. *PLoS ONE,* 8: e57988. doi: 10.1371/journal.pone.0057988

Stoiber, K. C. (2014). A comprehensive framework for multitiered systems of support in school psychology. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 41-70). Bethesda, MD: National Association of School Psychologists.

Strait, G. G., Smith, B. H., Pender, C., Malone, P. S., Roberts, J., & Hall, J. D. (2015). The reliability of randomly generated math curriculum-based measurements. *Assessment for Effective Intervention, 40,* 247-253. doi: 10.1177/1534508415588075

Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristics curves. *Canadian Journal of Psychiatry, 52,* 121-128.

Tall, D. (2008). The transition to formal thinking in mathematics. *Mathematics Education Research Journal, 20,* 5-24. doi: 10.1007/BF03217474

The President's Council of Advisors on Science and Technology. (2011). *K-12 science, technology, engineering, and math (STEM) education for America's future*. Retrieved from Techdirections.com.

Toll, S. M. W., & Van Luit, J. E. H. (2014). Explaining numeracy development in weak performing kindergarteners. *Journal of Experimental Child Psychology, 124,* 97-111. doi: 10.1016/j.jecp.2014.02.001

U.S. Department of Education, National Center for Educational Statistics. (2012). *The condition of education 2012*. (NCES 2012-045 Indicator 47). Retrieved from http://nces.ed.gov/pubs2012/2012045.pdf

U.S. Department of Education. (2015). *Nation's report card: Mathematics 2015.* Washington, DC: National Center for Education Statistics.

VanDerHeyden, A. M. (2010). Determining early mathematical risk: Ideas for extending the research. *School Psychology Review, 39,* 196-202.

VanDerHeyden, A. M. (2013). Universal screening may not be for everyone: Using a threshold model as a smarter way to determine risk. *School Psychology Review, 42,* 402-414.

VanDerHeyden, A., Broussard, C., Fabre, M., Stanley, J., Legendre, J., & Creppell, R. (2004). Development and validation of curriculum-based measures of math performance for preschool children. *Journal of Early Intervention, 27,* 27–41. doi: 10.1177/105381510402700103

VanDerHeyden, A. M., Broussard, C., & Cooley, A. (2006). Further development of measures of early math performance for preschoolers. *Journal of School Psychology, 44,* 533-553. doi: 10.1016/j.jsp.2006.07.003

VanDerHeyden, A., Broussard, C., Snyder, P., George, J., LaFleur, S. M., & Williams, C. (2011). Measurement of kindergarteners' early mathematical concepts. *School Psychology Review, 40,* 296 – 306.

VanDerHeyden, A. M., Broussard, C., Fabre, M., Stanley, J., Legendre, J., & Creppell, R. (2004). Development and validation of curriculum-based measures of mat performance for preschool children. *Journal of Early Intervention, 27,* 27-41.

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

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

Vokovic, R. K., & Lesaux, N. K. (2013). The language of mathematics: Investigating the ways language counts for children's mathematical development. *Journal of Experimental Child Psychology, 115,* 227-244. doi: 10.1016/j.jecp.2013.02.002

Walker, M. H., & Shinn, M. R. (2010). Systemic, evidence-based approaches for promoting positive student outcomes within a multitier framework: Moving from efficacy to effectiveness. In M. R. Shinn & H. M. Walker (Eds.), *Interventions: for achievement and behavior problems in a three-tier model including RTI* (pp. 1-26). Bethesda, MD: National Association of School Psychologists.

Walker, H. M., Small, J. W., Severson, H. H., Seeley, J. R., & Feil, E. G. (2014). Multiple-gating approaches in universal screening within school and community settings. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 47-76). Washington, DC: American Psychological Association. doi: 10.1037/14316-000

Ward, A. R. (2015). Promoting strategic STEM education outreach programming using a systems-based STEM-EO model. *Research Management Review, 20*, 1-10.

White House Office of the Press Secretary. (2015). *Congress acts to fix no child left behind* [Fact sheet]. Retrieved from https://www.whitehouse.gov/the-press office/2015/12/03/press-briefing-press-secretary-josh-earnest-12215

Wynn, K. (1995). Infants possess a system of numerical knowledge. *Current Directions in Psychological Sciences, 4,* 172-177. doi: 10.1111/1467-8721

Appendix A

Institutional Review Board for the Protection of Human Subjects

Indiana University of Pennsylvania
www.iup.edu

Institutional Review Board for the
Protection of Human Subjects
School of Graduate Studies and Research
Stright Hall, Room 113
210 South Tenth Street
Indiana, Pennsylvania 15705-1048

P 724-357-7730
F 724-357-2715
irb-research@iup.edu
www.iup.edu/irb

November 20, 2015

Adelle Campbell
1651 Hampden Drive
York, PA 17408

Dear Ms. Campbell:

Your proposed research project, "Performance on Monitoring Basic Skills Progress-Computation Probes in First, Second, and Third Grade: Is it a Predictor of Pennsylvania System of School Assessment Mathematics Achievement in Third Grade?" (Log No. 15-282) has been reviewed by the IRB and is approved. In accordance with 45CFR46.101 and IUP Policy, your project is exempt from continuing review. This approval does not supersede or obviate compliance with any other University requirements, including, but not limited to, enrollment, degree completion deadlines, topic approval, and conduct of university-affiliated activities.

You should read all of this letter, as it contains important information about conducting your study.

Now that your project has been approved by the IRB, there are elements of the Federal Regulations to which you must attend. IUP adheres to these regulations strictly:

1. You must conduct your study exactly as it was approved by the IRB.
2. Any additions or changes in procedures must be approved by the IRB before they are implemented.
3. You must notify the IRB promptly of any events that affect the safety or well-being of subjects.
4. You must notify the IRB promptly of any modifications of your study or other responses that are necessitated by any events reported in items 2 or 3.

The IRB may review or audit your project at random *or* for cause. In accordance with IUP Policy and Federal Regulation (45CFR46.113), the Board may suspend or terminate your project if your project has not been conducted as approved or if other difficulties are detected

Although your human subjects review process is complete, the School of Graduate Studies and Research requires submission and approval of a Research Topic Approval Form (RTAF) before you can begin your research. If you have not yet submitted your RTAF, the form can be found at http://www.iup.edu/page.aspx?id=91683 .

While not under the purview of the IRB, researchers are responsible for adhering to US copyright law when using existing scales, survey items, or other works in the conduct of research. Information regarding copyright law and compliance at IUP, including links to sample permission request letters, can be found at http://www.iup.edu/page.aspx?id=165526.

I wish you success as you pursue this important endeavor.

Sincerely,

Jen Roberts

Jennifer Roberts, Ph.D.
Chairperson, Institutional Review Board for the Protection of Human Subjects
Professor of Criminology

JLR:jeb

Cc: Dr. Timothy Runge, Dissertation Advisor
Dr. Joseph Kovaleski, Graduate Coordinator
Ms. Brenda Boal, Secretary

Appendix B

Standardized Directions for MBSP-C Probe

It is time to take your math test. As soon as I give you your test, write your first name, your last name, and the date. After you have written your name and date on the test, turn your paper over and put your pencil down so I will know you are ready.

I want you to do as many problems as you can. Work carefully and do the best you can. Remember, start at top left. Work from left to right. Some problems will be easy for you; others will be harder. When you come to a problem you know you can do, do it right away. When you come to a problem that's hard, skip it and come back to it later.

Go through the entire test doing the easy problems. Then go back and try the harder ones. Remember, you might get points for getting part of a problem right. So, after you've done all the easy problems, try the harder problems. Try to do each problem even if you think you can't get the whole problem right.

When I say, "Begin", turn your test over and start to work. Work for the whole test time. You should have enough room to do your work in each block on the page. Write your answers so I can read them! If you finish early, check your answers. At the end of _____ minutes, I will say "Stop." Put your pencil down and turn your test face down.

Appendix C

Letter of Permission from Wiley and Sons and Copyright Clearance Center

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Oct 22, 2016

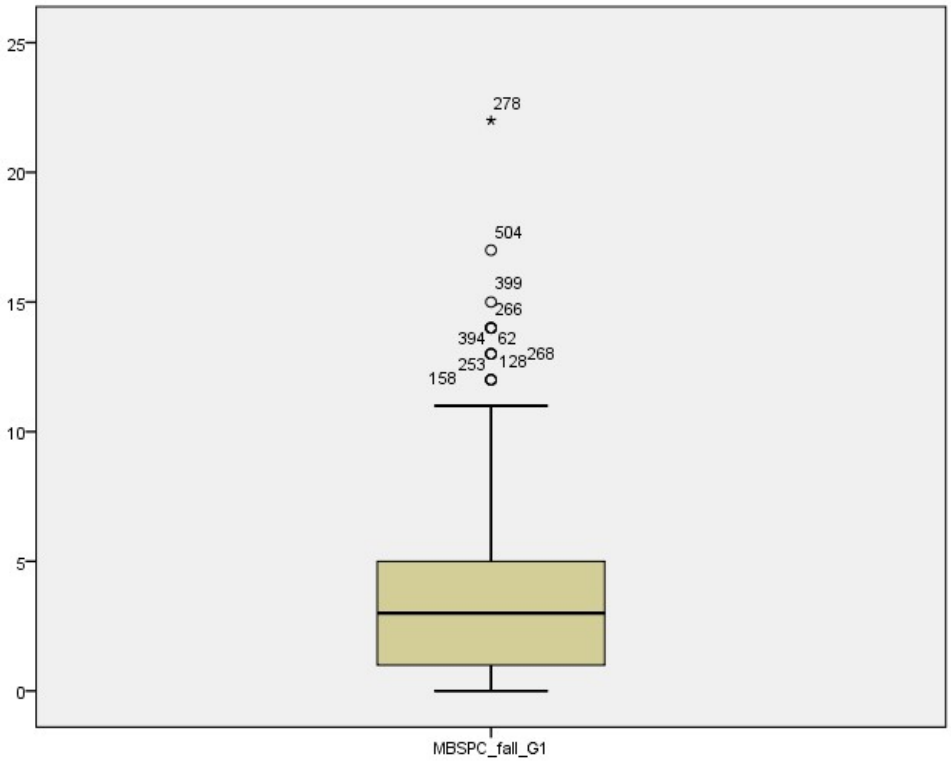
This Agreement between Adelle C Campbell ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

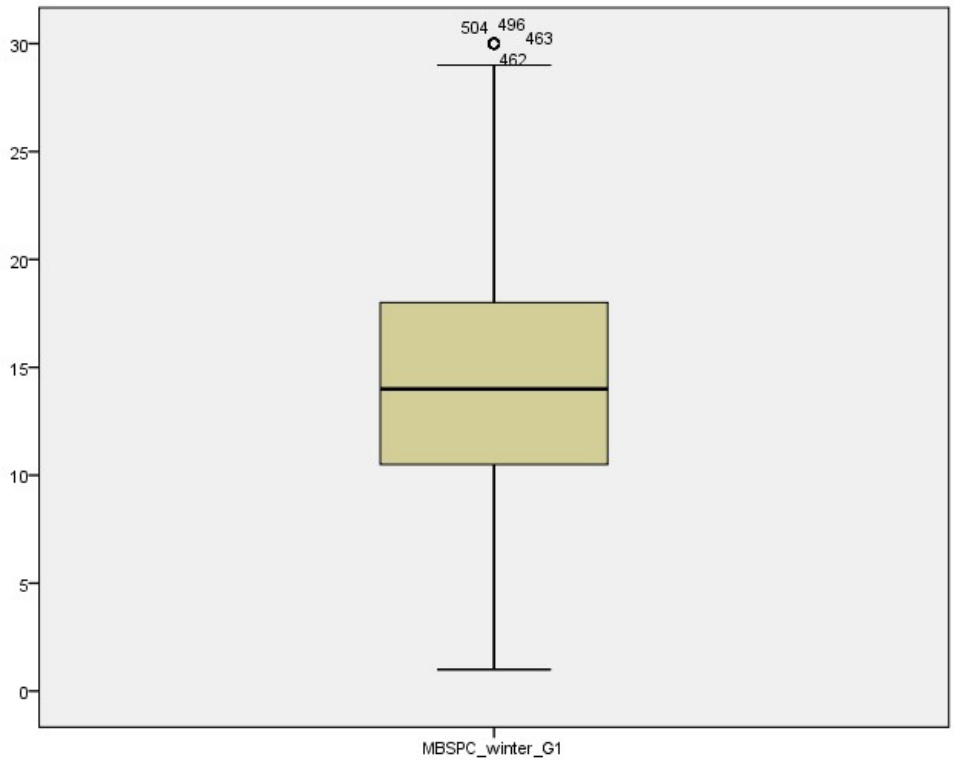
License Number	3974240335914
License date	Oct 22, 2016
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Child Development
Licensed Content Title	Pathways to Mathematics: Longitudinal Predictors of Performance
Licensed Content Author	Jo-Anne LeFevre,Lisa Fast,Sheri-Lynn Skwarchuk,Brenda L. Smith-Chant,Jeffrey Bisanz,Deepthi Kamawar,Marcie Penner-Wilger
Licensed Content Date	Nov 15, 2010
Licensed Content Pages	15
Type of use	Dissertation/Thesis
Requestor type	University/Academic
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	1
Original Wiley figure/table number(s)	Figure 1
Will you be translating?	No
Title of your thesis / dissertation	PERFORMANCE ON MONITORING BASIC SKILLS PROGRESS-COMPUTATION PROBES IN FIRST, SECOND, AND THIRD GRADE: IS IT A PREDICTOR OF PENNSYLVANIA SYSTEM OF SCHOOL ASSESSMENT MATHEMATICS ACHIEVEMENT IN THIRD GRADE?
Expected completion date	May 2017
Expected size (number of pages)	250
Requestor Location	Adelle C Campbell 1651 Hampden Drive YORK, PA 17408

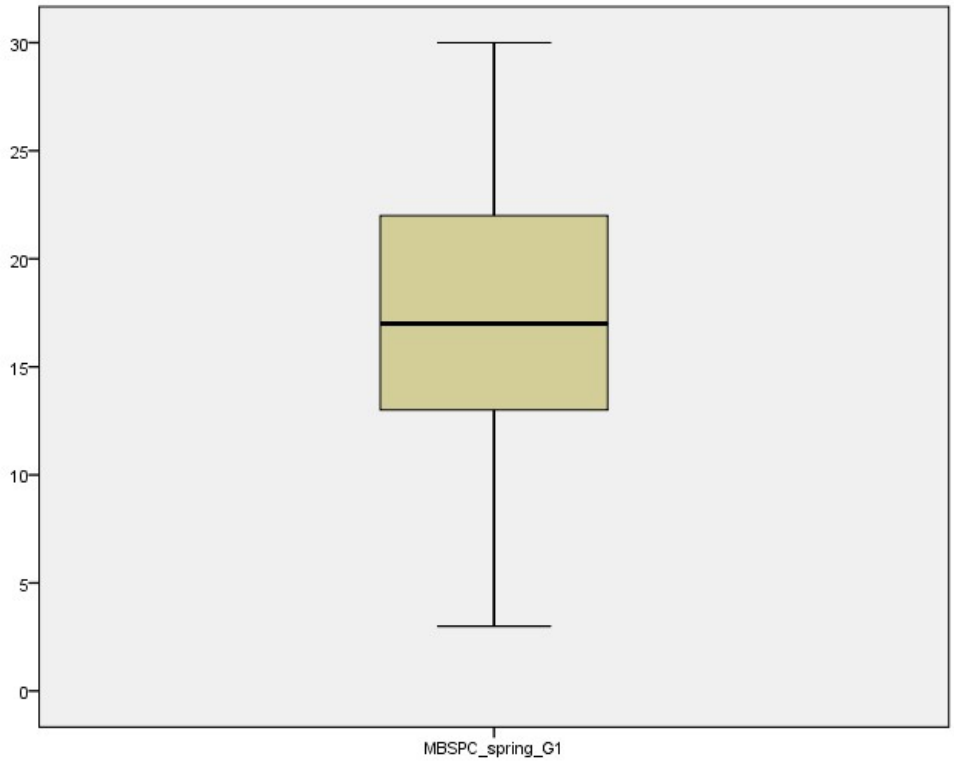
	United States Attn: Adelle C Campbell
Publisher Tax ID	EU826007151
Billing Type	Invoice
Billing Address	Adelle C Campbell 1651 Hampden Drive YORK, PA 17408 United States Attn: Adelle C Campbell
Total	0.00 USD

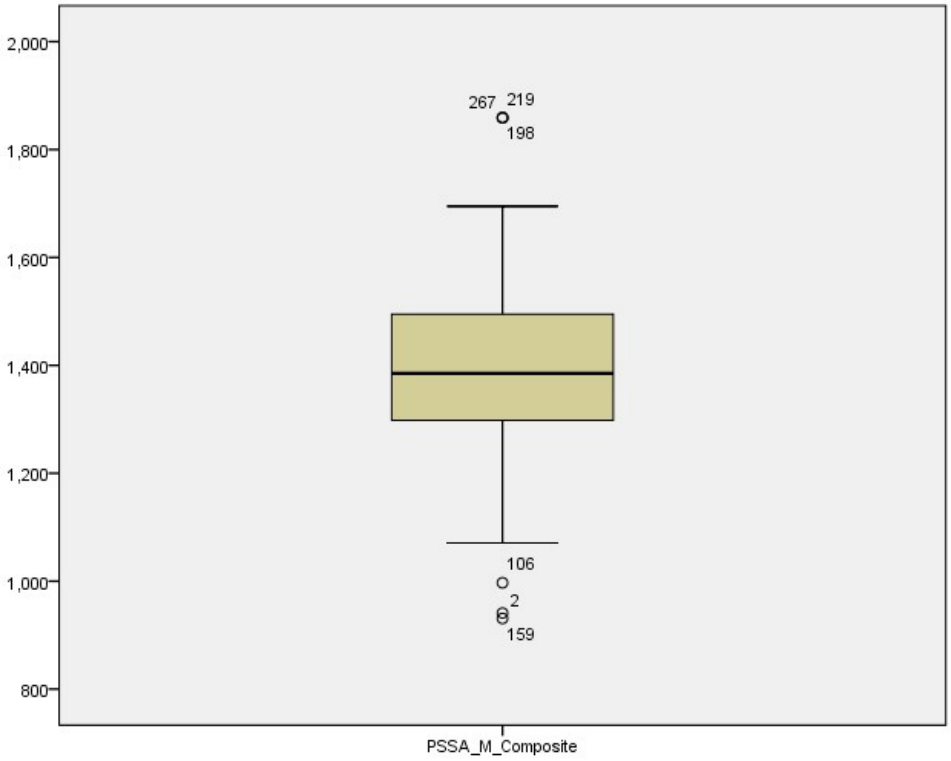
Appendix D

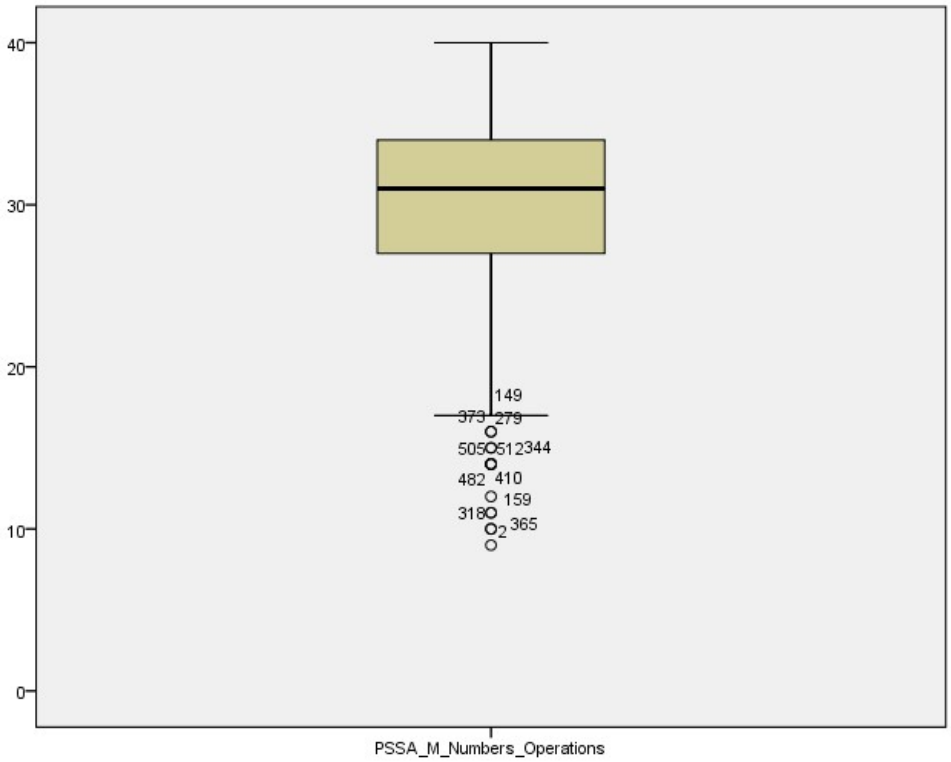
First Grade Boxplots for the Identification of Outliers

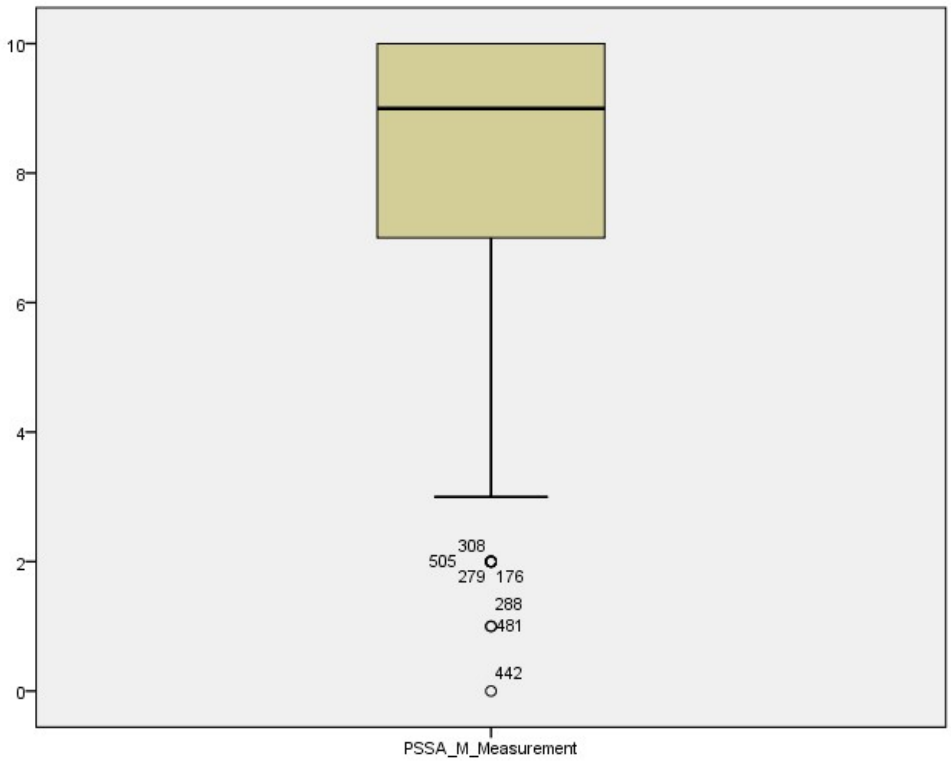


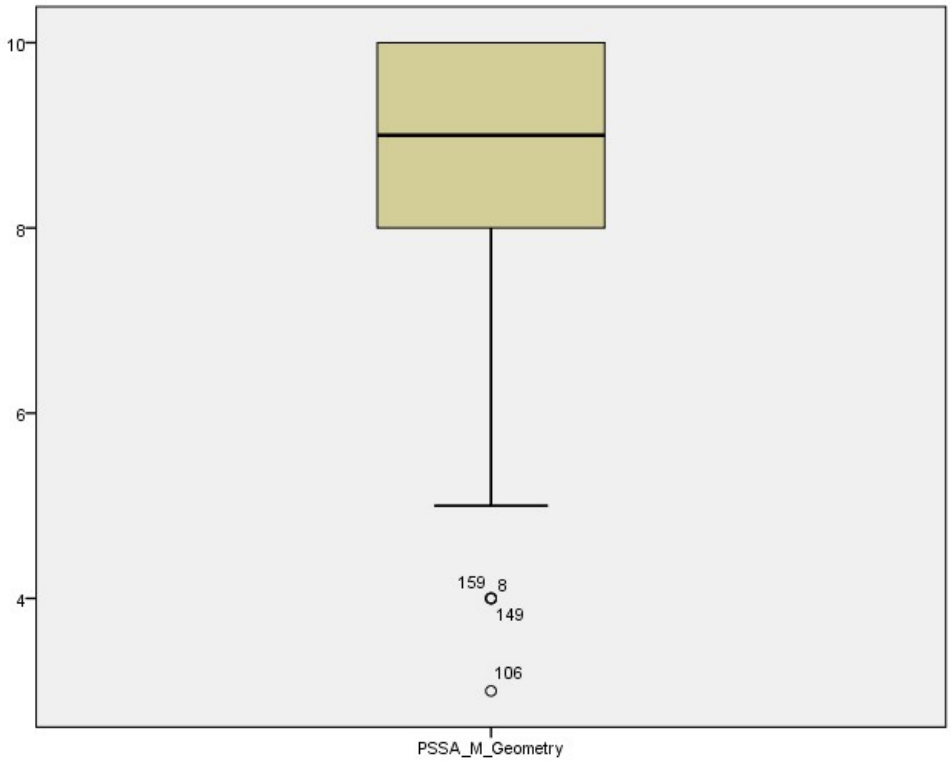

504
496
463
462
30
25
20
15
10
5
0
MBSPC_winter_G1

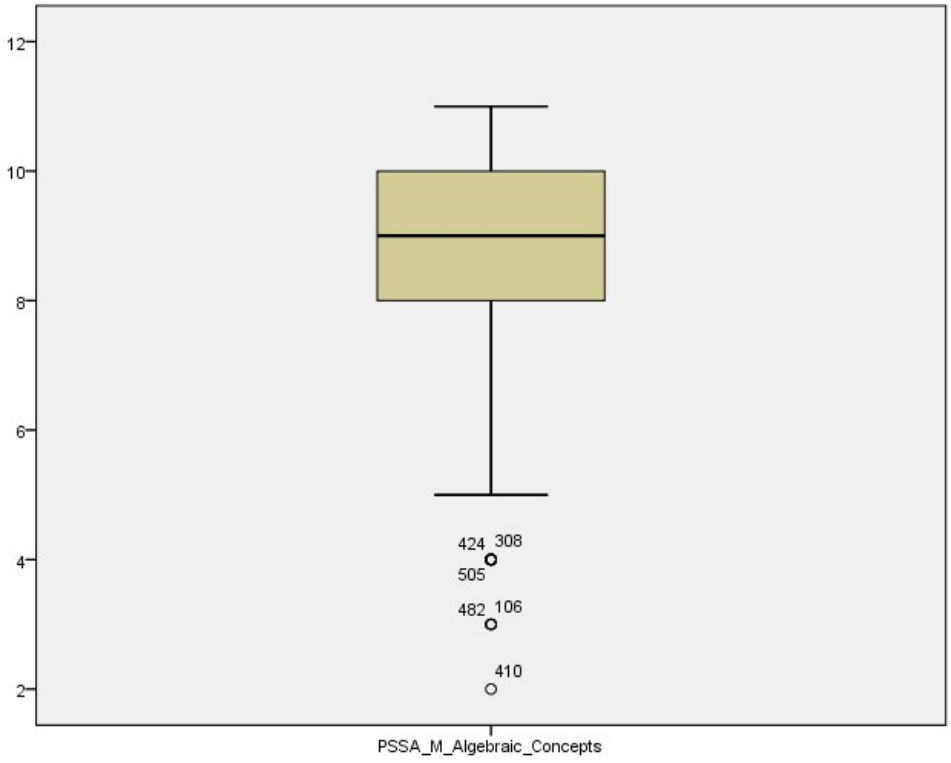

30
25
20
15
10
5
0
MBSPC_spring_G1

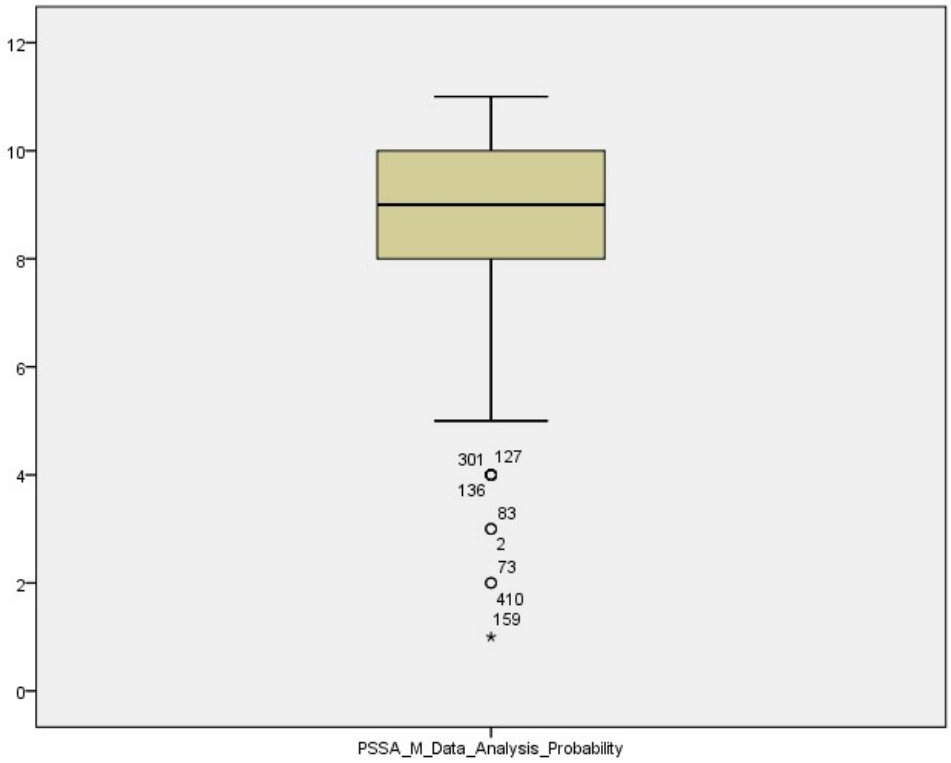

2,000
1,800
1,600
1,400
1,200
1,000
800
267
219
198
106
2
159
PSSA_M_Composite


40
30
20
10
0
149
373
279
505
512
344
482
410
159
318
365
2
PSSA_M_Numbers_Operations


10
8
6
4
2
0
308
505
279
176
288
481
442
PSSA_M_Measurement

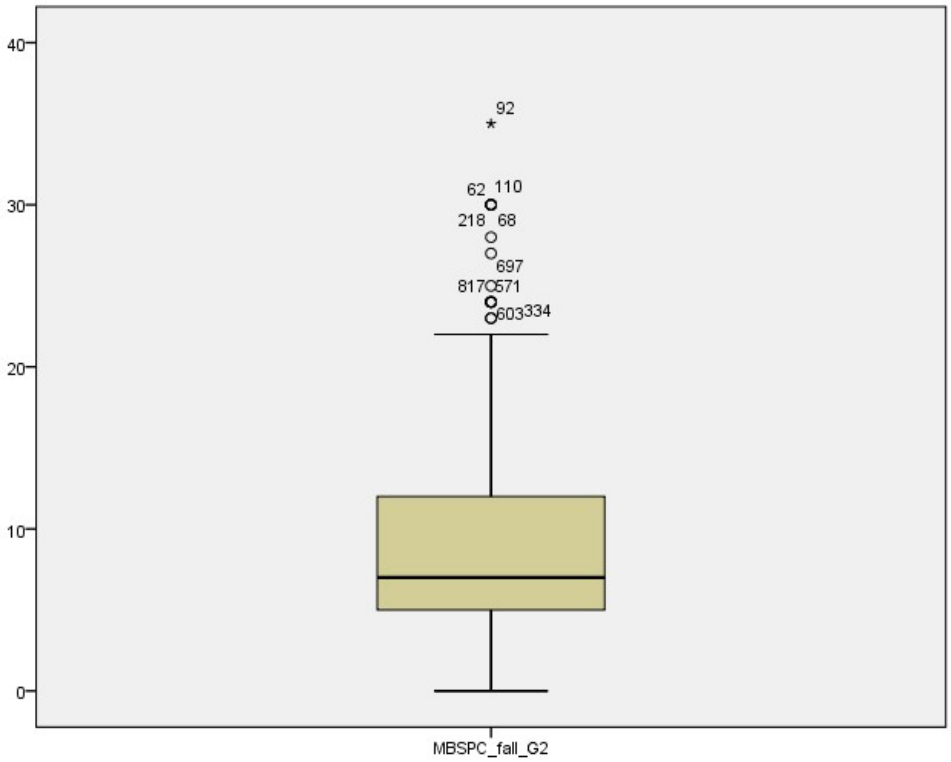

10
8
6
4
159
8
149
106
PSSA_M_Geometry

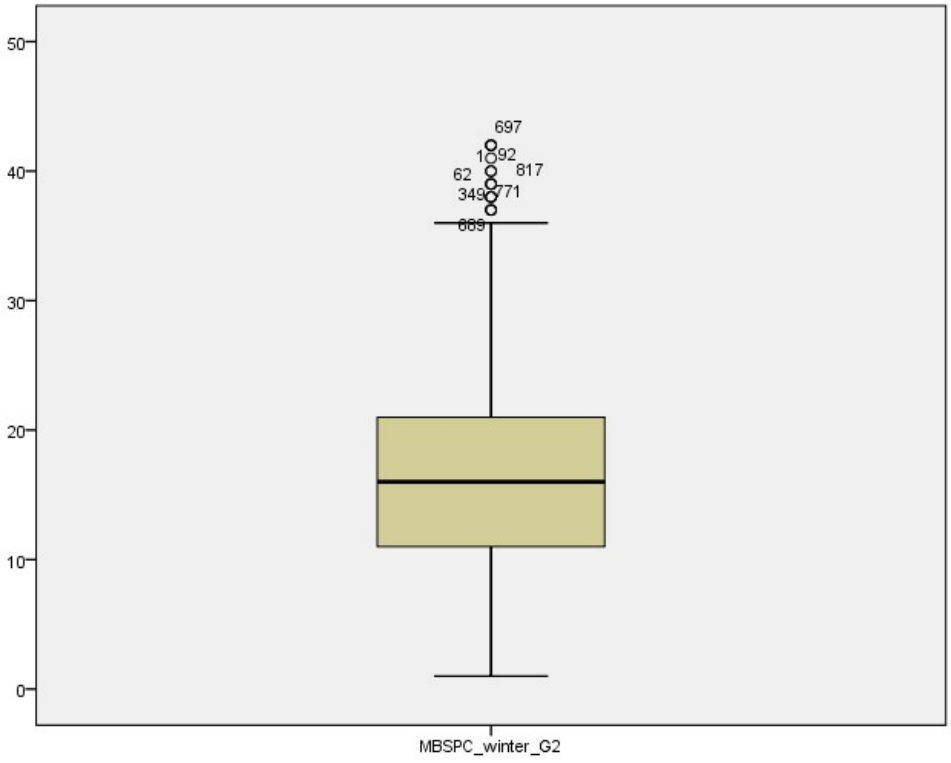

12
10
8
6
4
2
424
308
505
482
106
410
PSSA_M_Algebraic_Concepts

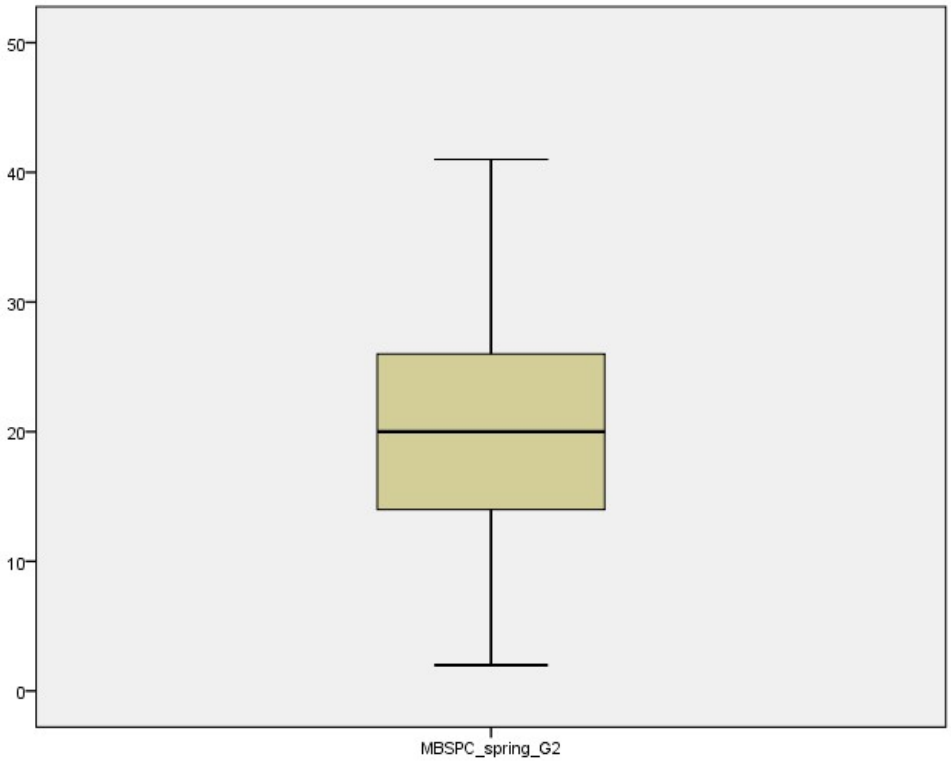

12
10
8
6
4
2
0
301
127
136
83
2
73
410
159
*
PSSA_M_Data_Analysis_Probability

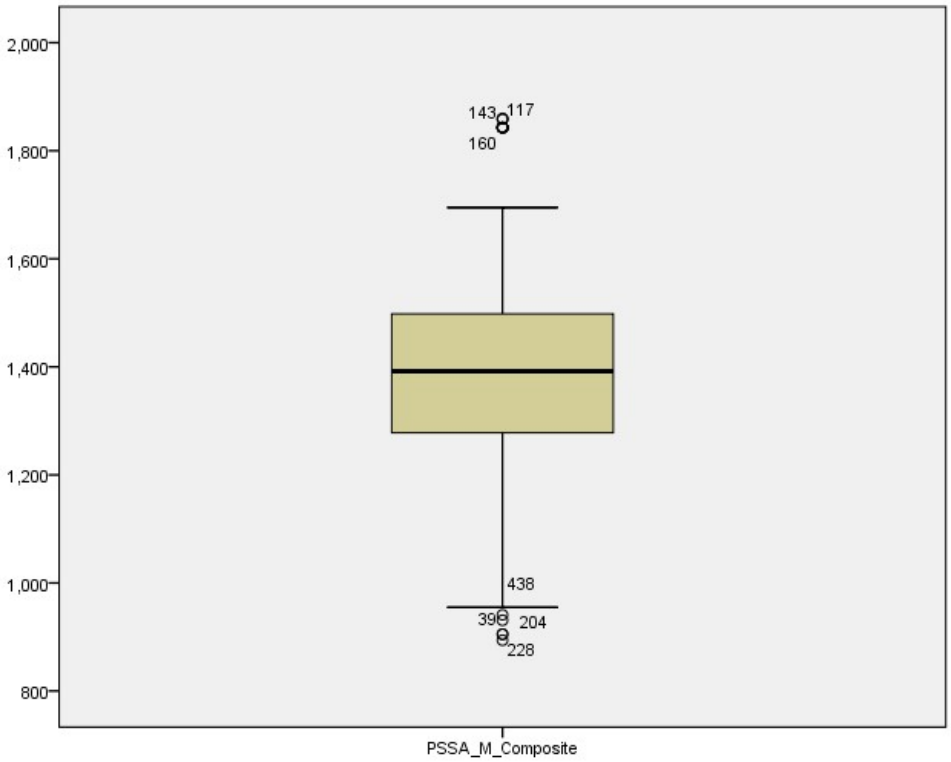
Appendix E

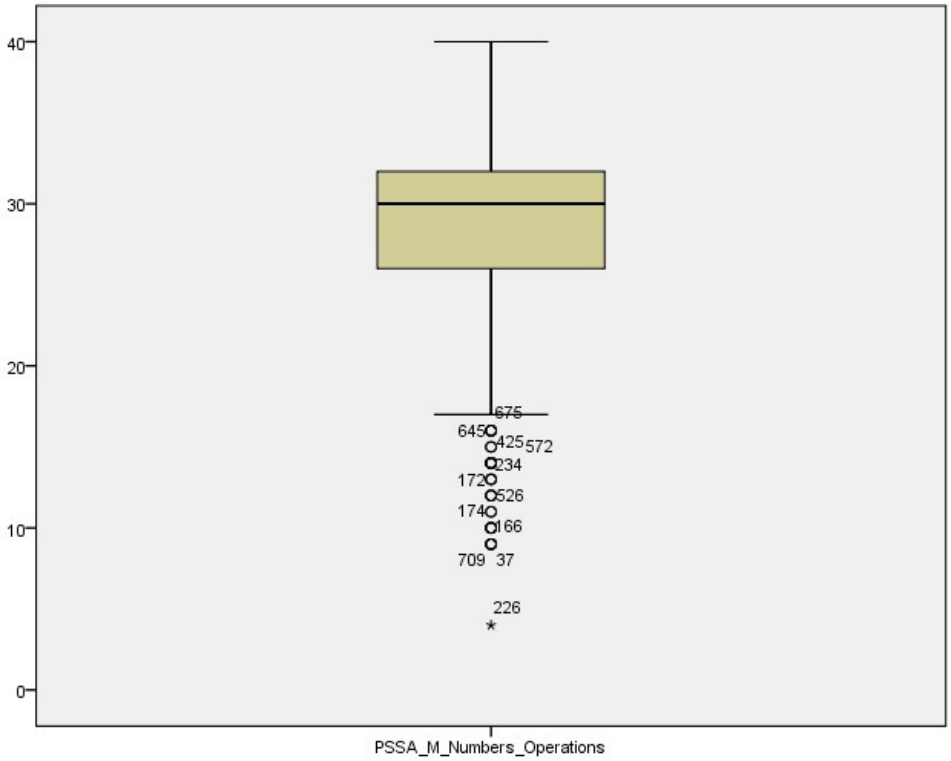
Second Grade Boxplots for the Identification of Outliers

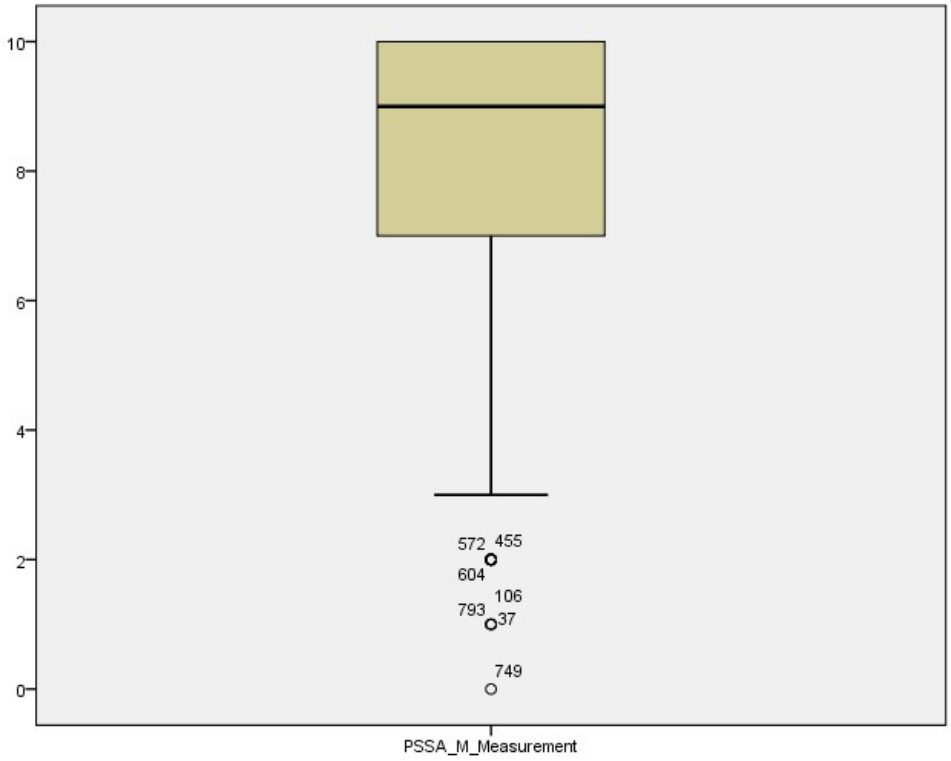


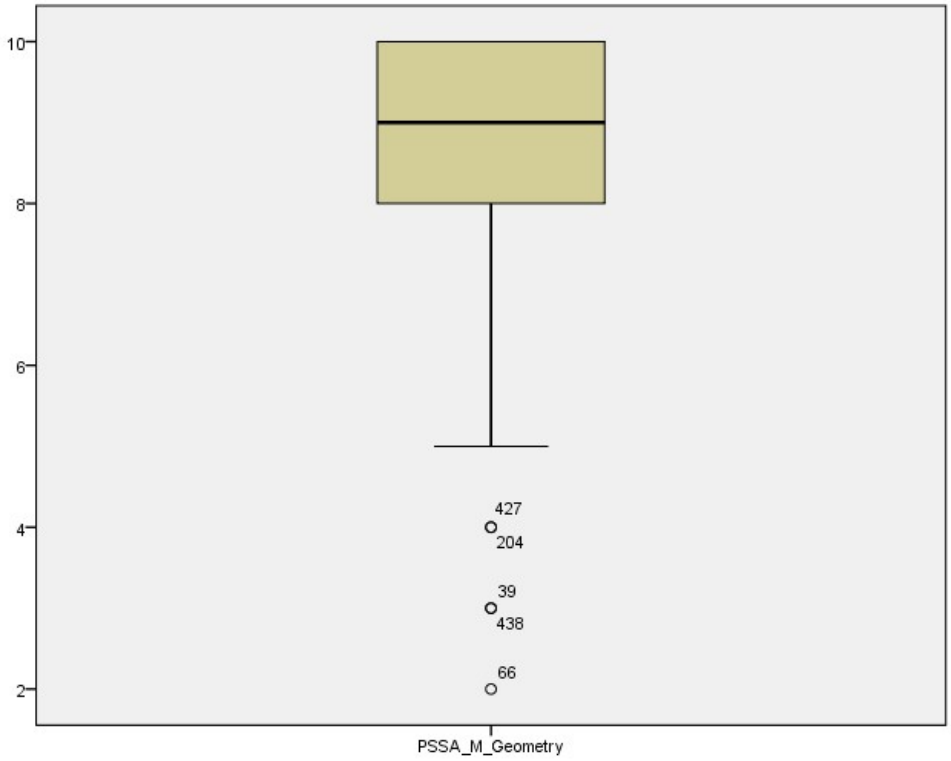

50
40
30
20
10
0
697
1092
62
817
349
771
669
MBSPC_winter_G2

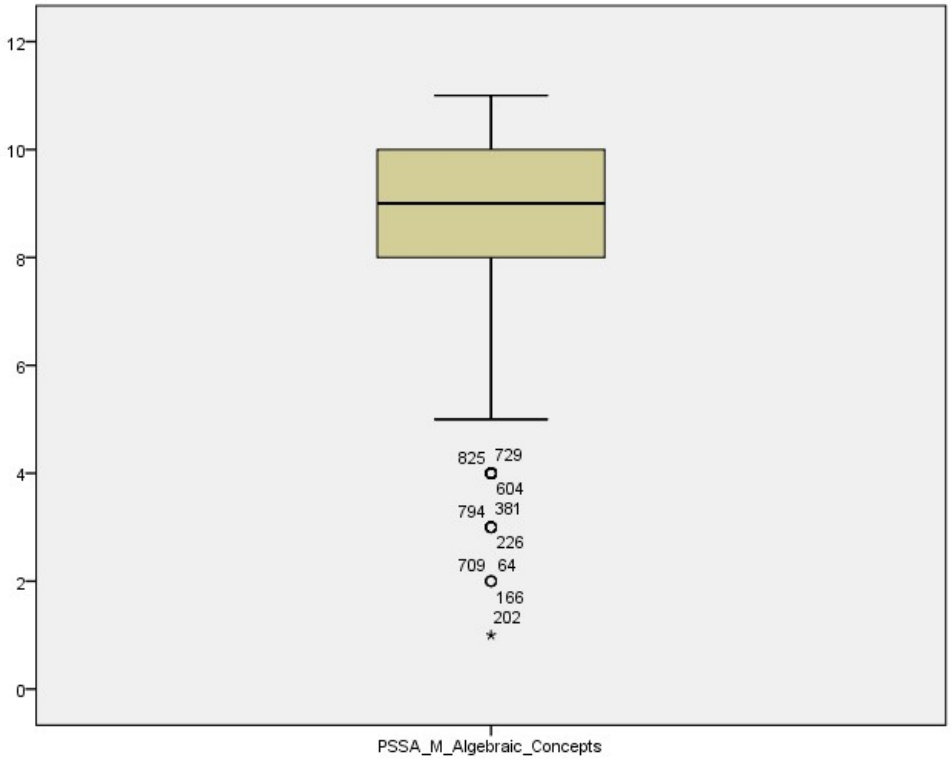

50
40
30
20
10
0
MBSPC_spring_G2

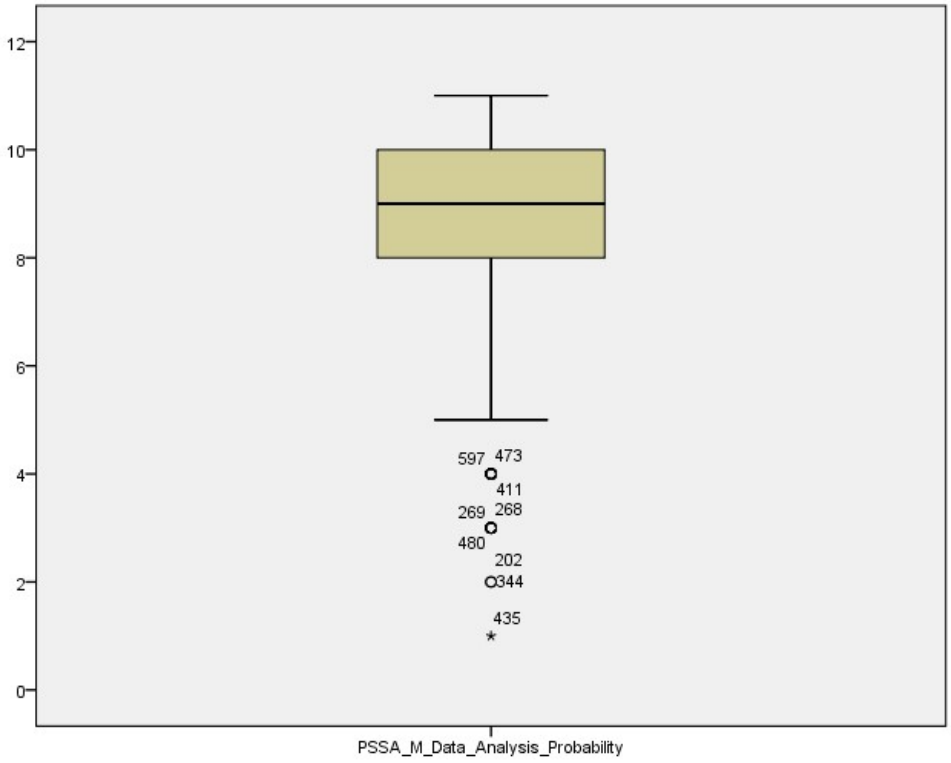

2,000
1,800
1,600
1,400
1,200
1,000
800
143
117
160
438
398
204
228
PSSA_M_Composite


40
30
20
10
0
675
645
425
572
234
172
526
174
166
709
37
226
PSSA_M_Numbers_Operations


10
8
6
4
2
0
572
455
604
106
793
37
749
PSSA_M_Measurement

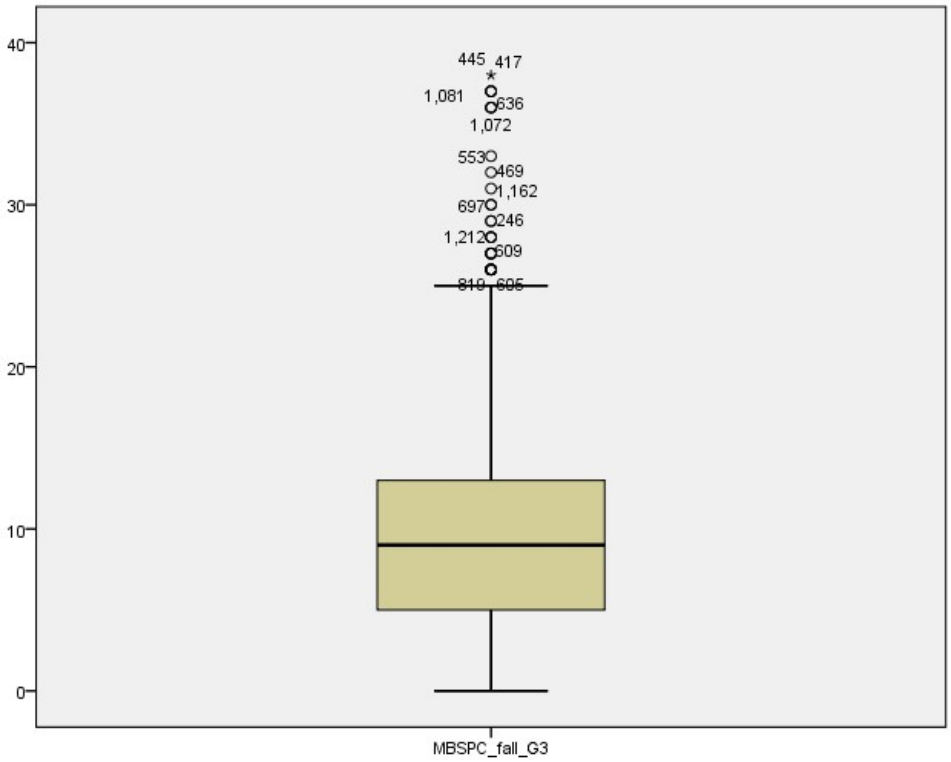

10
8
6
4
2
427
204
39
438
66
PSSA_M_Geometry

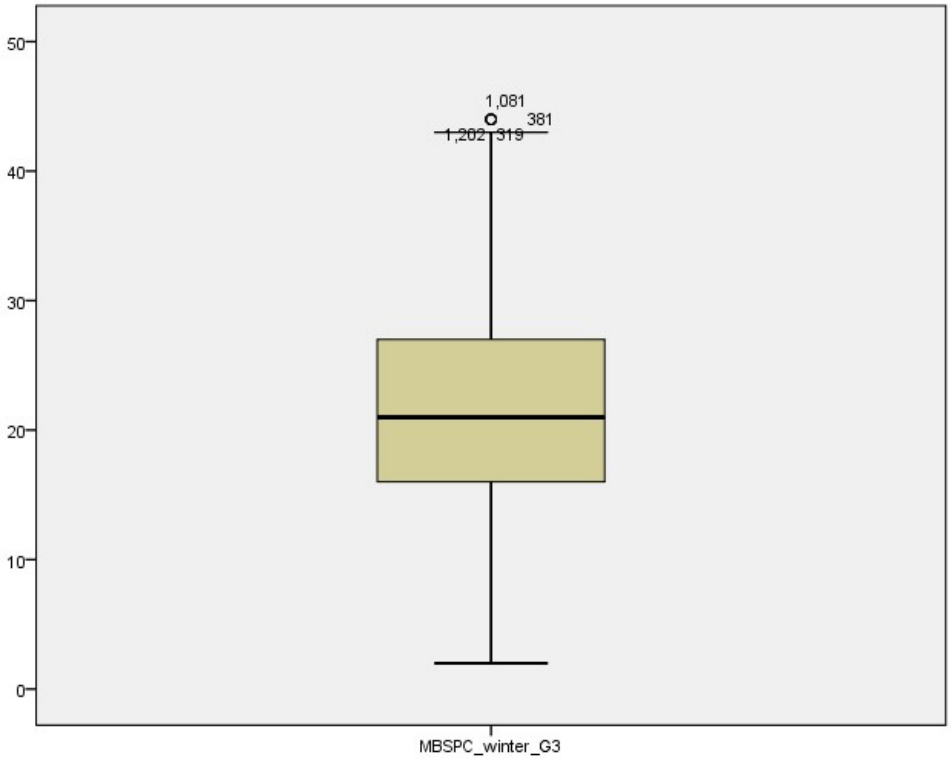

12
10
8
6
4
2
0
825
729
604
794
381
226
709
64
166
202
PSSA_M_Algebraic_Concepts

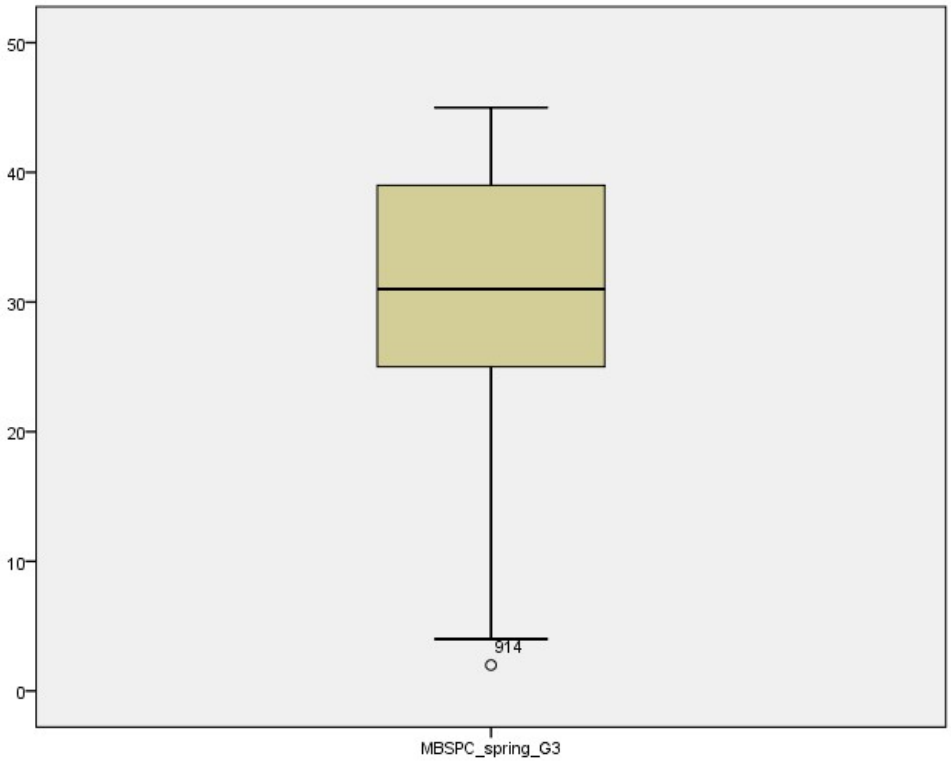

12
10
8
6
4
2
0
597
473
411
269
268
480
202
344
435
PSSA_M_Data_Analysis_Probability

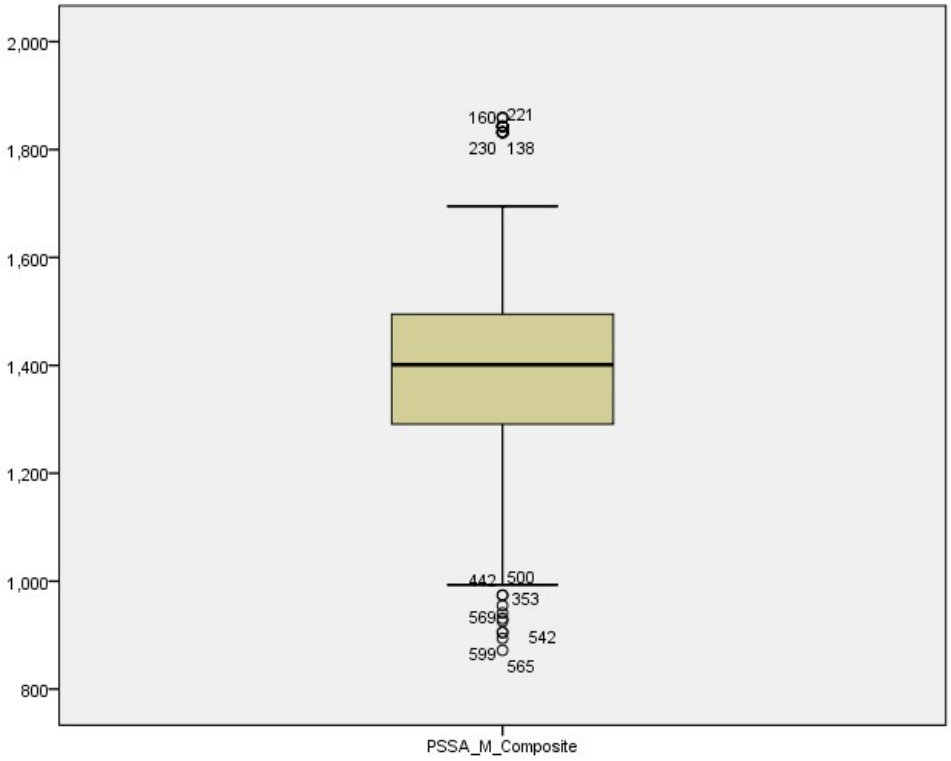
Appendix F

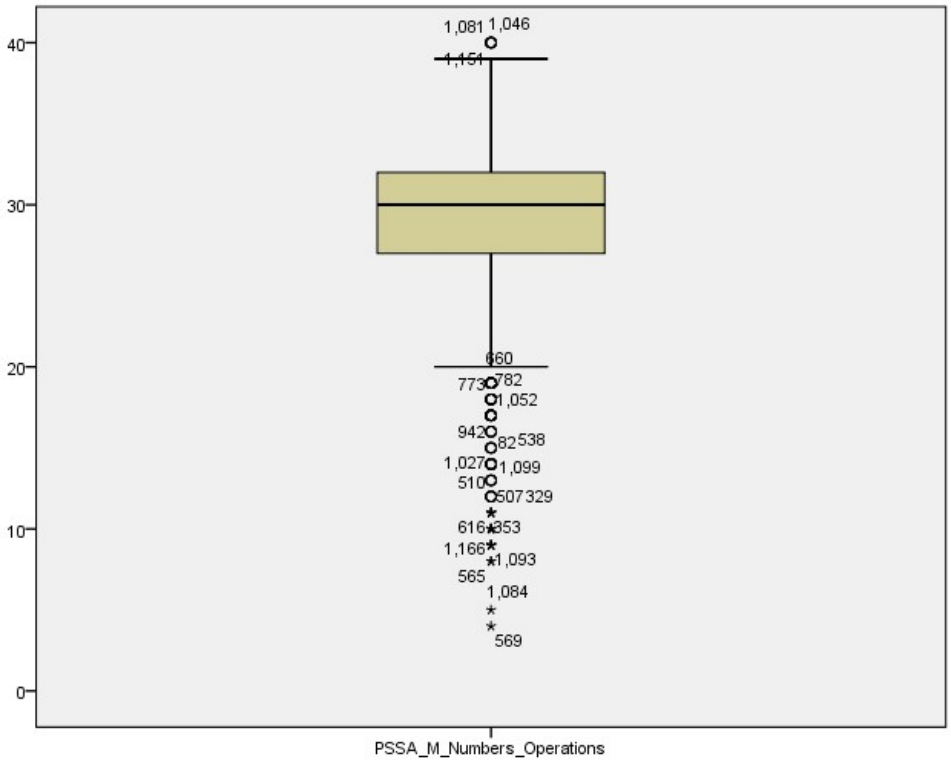
Third Grade Boxplots for the Identification of Outliers

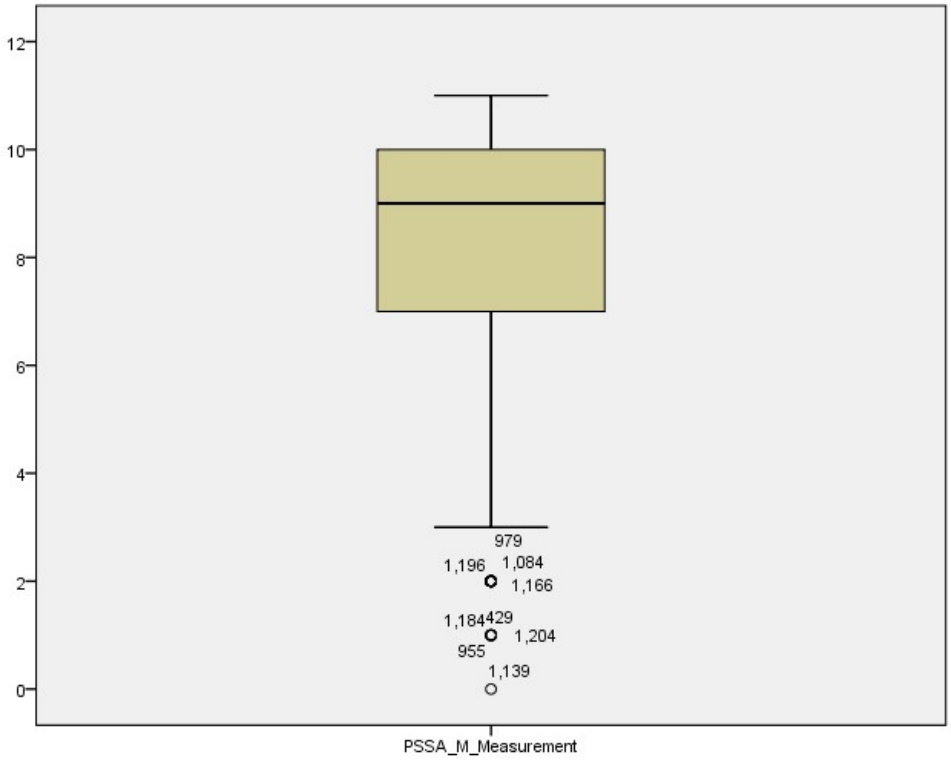


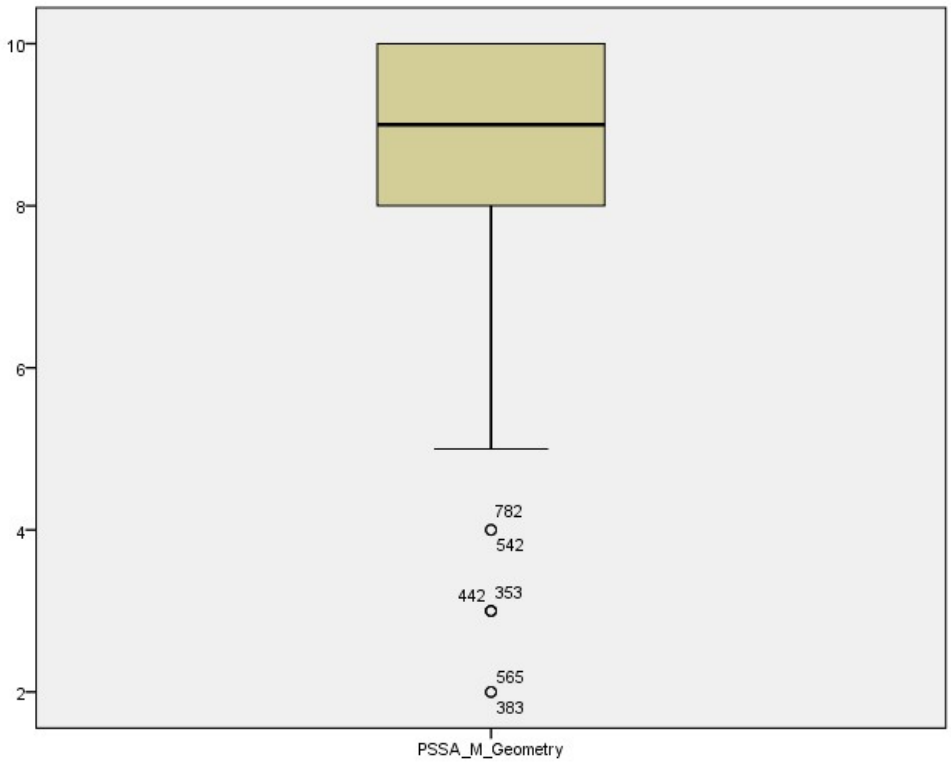

50
40
30
20
10
0
1,081
381
1,202
319
MBSPC_winter_G3

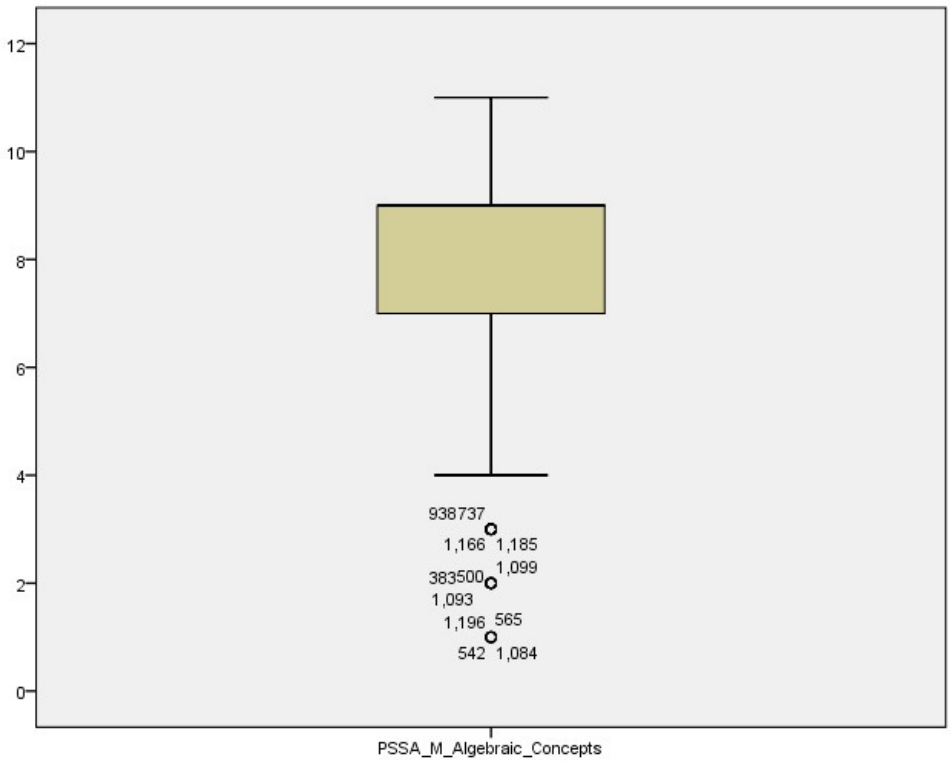

50
40
30
20
10
0
914
MBSPC_spring_G3

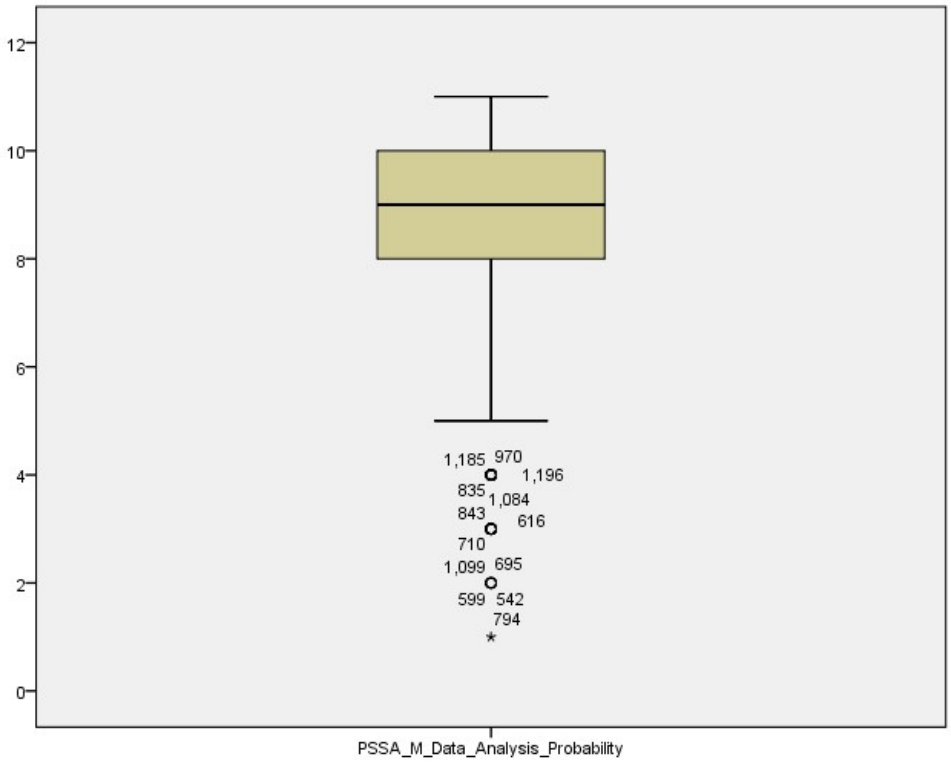

2,000
1,800
1,600
1,400
1,200
1,000
800
160
221
230
138
442
500
353
569
542
599
565
PSSA_M_Composite


40
30
20
10
0
1,081
1,046
1,151
660
773
782
1,052
942
82
538
1,027
1,099
510
507
329
616
353
1,166
1,093
565
1,084
569
PSSA_M_Numbers_Operations


12
10
8
6
4
2
0
979
1,196
1,084
1,166
1,184
429
1,204
955
1,139
PSSA_M_Measurement


10
8
6
4
2
782
542
442
353
565
383
PSSA_M_Geometry


12
10
8
6
4
2
0
938737
1,166
1,185
383500
1,099
1,093
1,196
565
542
1,084
PSSA_M_Algebraic_Concepts


12
10
8
6
4
2
0
1,185
970
1,196
835
1,084
843
616
710
1,099
695
599
542
794
*
PSSA_M_Data_Analysis_Probability