A PSYCHOMETRIC INVESTIGATION OF A MATHEMATICS PLACEMENT TEST AT A SCIENCE, TECHNOLOGY, ENGINEERING, AND MATHEMATICS (STEM) GIFTED RESIDENTIAL HIGH SCHOOL

A dissertation submitted to the
Kent State University
College of Education, Health, and Human Services
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

By

Hannah Ruth Anderson

August 2020

A dissertation written by

Hannah Ruth Anderson

B.A., Eastern Illinois University, 2011

M.A., Eastern Illinois University, 2012

Ph.D., Kent State University, 2020

Approved by

_____, Director, Doctoral Dissertation Committee
Aryn C. Karpinski

_____, Member, Doctoral Dissertation Committee
Tricia Niesz

_____, Member, Doctoral Dissertation Committee
Rajeev Rajaram

Accepted by

_____, Director, School of Foundations, Leadership, and Administration
Kimberly S. Schimmel

_____, Dean, College of Education, Health, and Human Services
James C. Hannon

ANDERSON, HANNAH RUTH, PH.D., AUGUST 2020 EVALUATION AND MEASUREMENT

A PSYCHOMETRIC INVESTIGATION OF A MATHEMATICS PLACEMENT TEST AT A SCIENCE, TECHNOLOGY, ENGINEERING, AND MATHEMATICS (STEM) GIFTED RESIDENTIAL HIGH SCHOOL (304 pp.)

Director of Dissertation: Aryn C. Karpinski, Ph.D.

Educational institutions, at all levels, must justify their use of placement testing and confront questions of their impact on students' educational outcomes to assure all stakeholders that students are being enrolled in courses appropriate with their ability in order to maximize their chances of success (Linn, 1994; Mattern & Packman, 2009; McFate & Olmsted III, 1999; Norman, Medhanie, Harwell, Anderson, & Post, 2011; Wiggins, 1989). The aims of this research were to (1) provide evidence of Content Validity, (2) provide evidence of Construct Validity and Internal Consistency Reliability, (3) examine the item characteristics and potential bias of the items between males and females, and (4) provide evidence of Criterion-Related Validity by investigating the ability of the mathematics placement test scores to predict future performance in an initial mathematics course.

Students' admissions portfolios and scores from the mathematics placement test were used to examine the aims of this research. Content Validity was evidenced through the use of a card-sorting task by internal and external subject matter experts. Results from Multidimensional Scaling and Hierarchical Cluster Analysis revealed a congruence of approximately 63 percent between the two group configurations. Next, an Exploratory Factor Analysis was used to investigate the underlying factor structure of the

mathematics placement test. Findings indicated a three factor structure of PreCalculus, Geometry, and Algebra 1, with moderate correlations between factors.

Thirdly, an item analysis was conducted to explore the item parameters (i.e., item difficulty, and item discrimination) and to test for gender biases. Results from the item analysis suggested that the Algebra 1 and Geometry items were generally easy for the population of interest, while the PreCalculus items presented more of a challenge. Furthermore, the mathematics placement test was optimized by removing eleven items from the Algebra 1 factor and two items from the PreCalculus factor. All Internal Consistency Reliability estimates remained strong and ranged from .736 to .950.

Finally, Hierarchical Multiple Linear Regressions were used to examine the relationship between students' total and factor scores from the mathematics placement test with students' performance in their first semester mathematics course. Findings from the four Hierarchical Multiple Linear Regressions demonstrate that the total score students' receive on the mathematics placement test predicts their achievement in their initial mathematics course, above and beyond the contributions of their demographic information and previous academic background. More specifically, the Algebra 1 Factor Score from the mathematics placement test was the strongest predictor of student success among the lower level mathematics courses (i.e., Mathematical Investigations I or II). Similarly, both the Algebra 1 and PreCalculus Factor Scores from the mathematics placement test were significant predictors of students' grades in their first upper level mathematics course (i.e., Mathematical Investigations III or IV), providing evidence of Predictive Validity.

The current mathematics placement test and procedures appear appropriate for the population of interest given the empirical evidence demonstrated in this research study regarding the psychometric properties of the exam. The continued use of the revised mathematics placement test in the course placement decision-making process is advisable.

DEDICATION

I dedicate this work to my loving husband who has made numerous sacrifices so that I may achieve this goal of mine. Thank you for your unwavering support and encouragement along the journey. May we be blessed with many more years to come and countless memories with Levi and Rylee. I also dedicate this work to my family who impressed upon me the importance of education. Thank you for your guidance and prayers. To all my family here today, and to those who have gone before us, but whom we will see again, I hope I have made each one of you proud.

ACKNOWLEDGEMENTS

"You will come to know that what appears today to be a sacrifice will prove instead to be the greatest investment that you will ever make." – Gordon B. Hinckley

I want to thank several individuals for taking the time to invest in me so that I may achieve this academic goal.

First, I would like to acknowledge the high school participating in this study. Without your encouragement, guidance, and collaboration this research would not have been possible. I hope this research provides useful information to assist the institution in moving forward. More specifically, I would like to thank the members of the mathematics team for participating in this research, continually answering my questions and sharing with me your knowledge and expertise, and for being open and honest in asking questions in return. I am appreciative of your partnership and look forward to collaborating more in the future.

Next, I would like to express my appreciation to all of the internal and external subject matter experts that participated in the card-sorting task. To my high school teacher, thank you for always believing in me and pushing me to do my best. To my undergraduate and graduate professors, thank you for your continued support and teaching throughout the years. I will forever cherish your dedication to students and the field of mathematics education. Lastly, for those that willingly responded to a doctoral research study, thank you for dedicating your time so that another student may pursue their educational goals.

Additionally, I would like to acknowledge my dissertation committee members, Dr. Tricia Niesz and Dr. Rajeev Rajaram. Their valuable participation and feedback throughout this process gave me additional clarity and understanding. I appreciate them keeping me grounded in the literature during this research project.

With the sincerest of gratitude, I would like to thank my dissertation director, Dr. Aryn C. Karpinski. There are not enough words to express how thankful I am for the sacrifices and contributions she has made to this work and my life. Thank you for your patience and support during this non-linear journey. Your words of encouragement have continued to resonate with me and have pushed me beyond my comfort zone. Your experience in evaluation and measurement along with your commitment to education is unmeasurable. It has been an honor to learn from you and to be one of your doctoral students, but to become your colleague and a friend is a true life-long blessing.

Most importantly, I wish to acknowledge my entire family for whom I am eternally thankful. To my parents, thank you for your never-ending love and support. You believed in me more than I believed in myself and I thank you for the solid foundation that you laid for our family. Your wisdom and guidance goes beyond measure and I am forever grateful for your continued prayers and words of encouragement. Thank you to my brother for paving the way and for teaching me that anything is possible with hard work and perseverance.

To my devoted husband, thank you for your continued patience, reassurance, love, and support throughout this entire journey. While this process has at times been challenging, frustrating, and time-consuming, I appreciate your understanding and

willingness to entertain Levi and Rylee for hours on end so that I could complete this work. I thank God for joining us together and I am elated to share this achievement with you and our family.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

In educational measurement, constructs such as achievement, interest, and performance are assigned numerical values, through the use of a wide variety of tests and assessments, to infer the abilities and proficiencies of students. The purpose of achievement testing is to measure students' actual knowledge or acquired skills in order to reliably distinguish between students who do and do not have some level of the construct of interest (Slavin, 2007). As one of the primary measures used in educational research, there is an abundance of literature focused on achievement testing.

Beginning at the post-secondary level, numerous articles have been published regarding the use of placement tests for incoming students. Many of these articles mention the continuing decline of academic standards, specifically in the area of mathematics (e.g., Crist, Jacquart, & Shupe, 2002; Hoyt & Sorensen, 2001; Medhanie, Dupuis, LeBeau, Harwell, & Post, 2012; Ngo & Kwon, 2015; Parker, 2005; Schmitz & delMas, 1991). Unsurprisingly, the lowered academic standards in math are said to be related to students' scoring lower on mathematics placement tests. Due to the lower test scores, more students are being assigned to take remedial coursework, which has sparked a conversation about whether or not students are less prepared for college-level work or if the placement tests used are appropriate for this type of decision (Morgan & Michaelides, 2005).

More specifically, nearly one-third of all students entering community colleges take at least one remedial or developmental course in mathematics (e.g., Bailey, 2009;

Hoyt & Sorensen, 2001; Kowski, 2013; Medhanie et al., 2012; Melguizo, Kosiewicz, Prather, & Bos, 2014; Scott-Clayton, 2012). Not only do these remedial courses lower student motivation, but they also add time to student graduation. Furthermore, the additional time students spend taking non-credit courses increases their overall cost to attend and lowers retention rates (Medhanie et al., 2012; Melguizo, Hagedorn, & Cypers, 2008; Ngo & Kwon, 2015; Scott-Clayton, 2012). Some community colleges have even been accused of placing students into these remedial, non-credit courses as a way to increase revenue (Armstrong, 2000). As a result, post-secondary institutions are now being asked to provide evidence of the effectiveness of their placement procedures and measures to ensure that the negative consequences of misplacement are minimized (Armstrong, 2000; Morgan & Michaelides, 2005; Smith & Fey, 2000). Accurately placing students is a necessary, but not sufficient, condition for a placement system as a whole to be effective (Sawyer, 1996).

A similar theme of remediation appears in the K-12 educational literature on achievement testing. In response to the *No Child Left Behind Act* (NCLB), schools and districts are required to demonstrate a yearly increase in their students' academic performance through the use of a standardized assessment. Through this measure of accountability, it is expected that students from traditionally underrepresented populations (i.e., African American, Hispanic, special education, English language learners) would no longer be "academically forgotten" (U.S. Department of Education, Office of the Secretary, & Office of Public Affairs, 2004). As anticipated, school and teacher resources have been directed towards the lower performing groups of individuals

in an effort to simultaneously close the achievement gap and demonstrate adequate yearly progress (Gallagher, 2004).

With teachers' time and attention drawn away from the high-achieving students, the needs of these gifted children have become (somewhat) overlooked. Subotnik, Olszewski-Kubilius, and Worrell (2011) stated that within the areas of research, program funding, policy, and K-12 teacher training, little to no attention is given to the classroom environments and/or needs of high-achieving students. However, the assumption that these academically talented children will thrive on their own is a myth (DeLacy, 2004; Marshall, McGee, McLaren, & Veal, 2011; Mendoza, 2006; Subotnik et al., 2011). Analysts argue that the more recent approach "STEM for all" (i.e., providing all students with as much high quality STEM education as possible) is not working and suggest that a framework called "All STEM for some" be implemented (Atkinson, 2012; Gonzalez & Kuenzi, 2012). In this framework, students who are most interested in STEM and have the potential to do well in STEM are confronted with intensive learning experiences encompassing a challenging curriculum and appropriate assessments (National Academy of Sciences, National Academy of Engineering, & Institute of Medicine, 2007; National Commission on Excellence in Education, 1983). Thus, if excellence, as well as equity, are genuine goals of the American educational system, then there is a dire need for an advanced, differentiated curriculum for gifted and talented students (Gallagher, 2004).

Over the past forty years, specialized Science, Technology, Engineering, and Mathematics (STEM) schools, projects, and programs have been established for gifted children. Within these programs, gifted students are exposed to a challenging college

preparatory curriculum with the expectation of majoring in a STEM field. It has been said that the residential schools provide liberating environments where the students can learn at a pace suited for their talents while being surrounded by like-minded, intellectual peers (Jones, 2009). However, public state-supported residential schools and other STEM programs do not come at a small price. Thus it is expected, as with any new program, that stakeholders (i.e., state legislators and the public) would seek data-driven evidence to establish the positive effects of these schools and programs on students' future educational outcomes (Atkinson & Mayo, 2010; Pfeiffer, Overstreet, & Park, 2010). More recently, research has identified a shortage of valid and reliable instruments to measure the impact and outcomes of these specialized STEM schools and programs (Katzenmeyer & Lawrenz, 2006; Scott, 2012). Some factors contributing to the shortage of reliable indicators are the assessment literacy of the educators within these programs, a lack of formal training in assessment and measurement techniques, and a need to establish partnerships between measurement professionals and K-12 educational institutions.

Assessment literacy can be defined as the ability to design, select, interpret, and use assessment results appropriately for future educational decisions (Quilter & Gallini, 2000). Prior research has indicated that classroom teachers spend up to fifty percent of their instructional time in assessment-related activities such as grading, oral questioning, or administering and interpreting tests (Plake & Impara, 1997; Quilter & Gallini, 2000; Schafer, 1993; Stiggins, 1991). While teachers are largely exposed to assessment practices, few in-service and pre-service teachers have received formal assessment

training (Impara, Plake, & Fager, 1993; Schafer, 1993; Sondergeld, 2014; Stiggins & Bridgeford, 1985). Not only does this gap in training affect teachers' attitudes towards assessment, but it can also affect the students' educational outcomes.

For example, many institutions, like the high school in this study, have favored the development of their own placement tests over the use of commercialized exams such as Compass, Accuplacer, or ALEKS. One of the reasons for choosing to use an in-house exam over other tests is that department-made exams allow faculty to customize the topics and content areas that they judge to be most relevant to making their placement decisions (Bressoud, Mesa, & Rasmussen, 2015; Flores, 2007). However, when asked to validate the scores on their placement measures, many faculty reported feeling unsupported and noted that the policies currently in use were the result of continued experimentation (Ngo & Melguizo, 2016). Due to the time and cost associated with professional development, it is unrealistic to expect all teachers to have extensive training in evaluation and measurement techniques. However, educational assessments, if designed and used properly, can become instruction-enhancing tools. As a result, stakeholders and other critics are seeking data-driven research to evidence the psychometric properties of these placement tests and the effectiveness of their placement policies.

As evident in the literature, a majority of institutions have focused on the latter of these two concerns by examining the predictive validity of their assessments (e.g., Belfield & Crosta, 2012; Denny, Nelson, & Zhao, 2012; Pike & Saupe, 2002; Roth, Crans, Carter, Ariet, & Resnick, 2000; Rueda & Sokolowski, 2004; Schumacher &

Smith, 2008; Siegler et al., 2012). Findings from the research are generally positive and support the use of multiple measures in the placement process, but have neglected to address concerns of item quality, validity, and reliability. For this reason, teacher organizations and researchers can benefit from establishing more partnerships between content experts and assessment professionals. These partnerships can provide opportunities to address issues throughout the test development process and validate the scores on the measure while simultaneously highlighting the importance of measurement and evaluation. In the current study, a comprehensive examination of a mathematics placement test used at a gifted STEM residential high school was conducted. The measurement process and psychometric evidence provided in this study can help this high school and similar institutions be confident in making high-stakes decisions such as course placement.

Rationale

In the era of accountability, placement practices and methods that are rigorous and defensible are critical for educational institutions at varying levels to justify their use and to confront questions of their impact on students' educational outcomes. Frisbie (1988) stated that when the reliability of scores as accurate measures of student achievement are in question, these scores cannot be used to make future educational decisions. Furthermore, one validation study is not sufficient to guarantee the psychometric properties of an assessment throughout its lifetime. Instead, the assessment(s) and policies used, in contexts such as placement testing, need to be continuously reviewed and evaluated to assure that students are being placed into courses

commensurate with their ability in order to maximize the chances of success (Linn, 1994; Mattern & Packman, 2009; McFate & Olmsted III, 1999; Norman et al., 2011; Wiggins, 1989). Overall, when properly constructed and evaluated, assessments can enhance later performance and provide feedback on what has and has not been learned to both the student and other interested stakeholders.

Secondly, the high school in the current study recognized a need to more formally evaluate their mathematics placement exam in an effort to defend the placement policies used and to provide evidence that the decisions from the mathematics placement exam are enhancing later performance. Moreover, when there is more variability in student scores compared to historically consistent data, then a more thorough investigation is warranted. In other words, if the test scores evidence lower reliability, there is an increased likelihood of misrepresenting students' true level of knowledge leading to a decision, which could temporarily or permanently negatively impact, students' educational outcomes (Adedoyin & Mokobi, 2013; Frisbie, 1988; Latterell & Regal, 2003; Linn, 1994; Norman et al., 2011). Finally, previous research regarding placement exams and their psychometric properties have been conducted at the post-secondary level. This study is unique in extending the research to younger grade levels serving a specialized (i.e., gifted) population.

Research Aims

There are four overarching aims of this study: (1) To provide evidence of Content Validity, (2) To provide evidence of Construct Validity and Internal Consistency Reliability, (3) To examine the item characteristics and potential bias of the items

between males and females, and (4) To provide evidence of Criterion-Related Validity by investigating the ability of the mathematics placement test scores to predict future performance in an initial mathematics course. Specifically, this study is comprised of four manuscripts, each addressing one of the following research questions:

Research Question 1 (RQ1): What is the Content Validity of the items on a mathematics placement test for gifted, residential high school students interested in STEM?

Research Question 2 (RQ2): What are the psychometric properties of the scores on a mathematics placement test for gifted, residential high school students interested in STEM?

RQ 2A: What is the Construct Validity of the scores on a mathematics placement test for gifted, residential high school students interested in STEM?

RQ 2B: What is the Internal Consistency Reliability of the item scores on a mathematics placement test for gifted, residential high school students interested in STEM?

Research Question 3 (RQ3): What are the item characteristics (i.e., item parameters and Differential Item Functioning [DIF]) of the mathematics placement test for gifted, residential high school students interested in STEM?

RQ 3A: What are the item parameters (i.e., item difficulty, and item discrimination) of the mathematics placement test for gifted, residential high school students interested in STEM?

RQ 3B: How do the items on a mathematics placement test for gifted, residential high school students interested in STEM differ by sex?

Research Question 4 (RQ4): What is the Criterion-Related Validity of the item scores on a mathematics placement test for gifted, residential high school students interested in STEM?

Significance

The current investigation's findings are anticipated to extend beyond the single setting used in this study and to be applied to a variety of other educational settings. As mentioned previously, the general scope of this study is to examine the psychometric properties of a mathematics placement test used at a gifted, residential high school focused on STEM. The unique contribution is intended to act as a reference for other schools with a STEM and/or gifted education focus so that they may begin the validation process to further extend and improve upon the educational testing practices at other levels of schooling. Moreover, the same validation process could be adapted to examine the identification practices for gifted students across the nation and at varying grade levels.

Finally, this research seeks to draw attention to the nature and quality of teacher-developed assessments within the measurement community so that additional support and/or training can be provided to both pre-service and in-service teachers who wish to improve their classroom assessments. Both those in teacher education and the measurement community agree that assessment of student performance is an important skill for teachers to possess, but little is being done to close the gap. Thus, this research

may serve as a blueprint for teachers, administrators, and/or schools to feel empowered to begin the process of examining their own assessments and practices.

The next section of this document (i.e., Chapter II: Literature Review) provides a review of the literature pertinent to this study including topics such as STEM education, gifted education, and placement testing. Chapter Three (i.e., Methodology) provides an in-depth description of the methods used to address the research questions of this study such as detailed explanations of the measure, variables, and analyses. The next four chapters (i.e., Chapters Four, Five, Six, and Seven) contain manuscripts associated with each overarching research question of this study, as mentioned above. Finally, Chapter Eight (i.e., Conclusion) summarizes the four aims of this research study and provides some final remarks.

CHAPTER II

LITERATURE REVIEW

Strengthening education in the disciplines of science and mathematics has been emphasized in the United States (U.S.) since the early 1980s. The historic piece *A Nation at Risk* (National Commission on Excellence in Education, 1983) highlighted that schools often times focus too much on the foundational skills of reading and computation at the expense of other essential skills such as comprehension, analysis, problem solving, and the ability to draw conclusions. These other essential skills have been deemed critical for technology and science fields and are integral to incorporate in STEM education. Despite the criticisms of the report (Stedman, 1994), STEM education addresses these critical technology and science field skills has the potential to produce students and eventual members of the workforce who are able to solve global challenges such as clean and affordable energy, hunger, health, and national security (President's Council of Advisors on Science and Technology, 2010).

Previous research has argued that, specifically in mathematics, U.S. students are falling behind those in other nations. In 2000, high school students completed the Programme for International Student Assessment (PISA), which measures students' knowledge and skills in areas such as science, mathematics, and reading (Organisation for Economic Co-operation and Development, 2018). Moreover, the international boards of experts that design the assessment framework do so independently to the school curricula of the participating countries to emphasize an adolescent's ability to apply what

they have learned in school to real life situations (Hopfenbeck et al., 2018; Sälzer & Roczen, 2018).

Among the nations participating that year (i.e., 2000), Hong Kong-China, Japan, and Korea had the highest mean scores in mathematical literacy (Organisation for Economic Co-operation and Development, 2018). Twelve years later, the U.S. performed below average on the mathematics portion of the PISA, and was ranked 27th out of the 34 participating countries. However, the PISA assessment is not without critique. Previous research has commented on the PISA's exclusion of students with disabilities from participating in international tests, biasing the sample and impacting future educational policies related to educational equity (Hopfenbeck et al., 2018; Schuelka, 2013). While research has warned policy-makers and researchers to be cautious about using PISA data as a means for valid comparisons, the PISA can provide some descriptive information at the national and international levels (Hopfenbeck et al., 2018).

Similarly, students from around the world participate annually in the International Mathematical Olympiad (IMO). Established in 1959, the IMO is considered the "World Championship Mathematics Competition" for high school students (International Mathematical Olympiad Foundation, 2018). The U.S. has placed first in this competition seven times since their initial participation in 1974, and have accumulated 124 individual gold medals (International Mathematical Olympiad Foundation, 2018). Comparatively, China leads the nations with 19 first place winnings (since 1985), and currently holds 151 individual gold medals. The difference between the U.S.'s seven first place wins and

China's 19 (and the 124 to 151 individual gold medals, respectively) is not alarming at face value. However, as the U.S. has nearly ten years more of IMO participation compared to China, this perspective elucidates that students in the U.S. are falling behind other competitive nations, especially within the field of mathematics.

In an executive report by the President's Council of Advisors on Science and Technology (2010), a statement was made arguing that the U.S. now lags behind other nations in STEM education at both the elementary and secondary levels. However, the report also mentioned that the gap in STEM education is not only a concern of students' proficiency in STEM, but also the lack of interest in STEM among many Americans. For example, a 2007 report found that the U.S. ranked 29th out of a 109 countries in the percentage of 24 year olds with either a mathematics or science degree (Atkinson, Hugo, Lundgren, Shapiro, & Thomas, 2007; Pfeiffer et al., 2010). That same report indicated that between 1985 and 2002, the number of U.S. citizens obtaining STEM graduate degree increased by a mere 14 percent, while the number of graduate STEM degrees awarded to students born outside of the U.S. more than doubled (Atkinson et al., 2007; Atkinson & Mayo, 2010). Previous research has noted, however, that when adolescents with interests an talents in mathematics and science are provided an environment with a challenging curricula, expert instruction, and peer stimulation, they are more likely to pursue STEM at post-secondary institutions (Bloom & Sosniak, 1985; Pyryt, 2000; Subotnik, Duschl, & Selmon, 1993; Tai, Liu, Maltese, & Fan, 2006). Therefore, within the U.S. specifically, political and educational leaders have continued to highlight a dire need to increase support given to the teaching of science and mathematics.

Science, Technology, Engineering, and Mathematics (STEM) Education

From the critical needs outlined in *A Nation at Risk*, the National Science Foundation (NSF) began a movement focusing on Science, Technology, Engineering, and Mathematics (STEM) in order to cultivate a globally-recognized workforce that is diverse, creative, and innovative. Both policymakers and stakeholders agree that widespread literacy in STEM, in addition to specific STEM expertise, is a key component to developing human capital to compete internationally in the 21st century (Breiner, Harkness, Johnson, & Koehler, 2012; Gonzalez & Kuenzi, 2012). Broadly stated, STEM literacy includes both procedural and conceptual skills, abilities, and understandings to equip individuals to encounter and address STEM-related personal, social, and global problems (Bybee, 2010). To solve such large issues, researchers have suggested that literacy in STEM should be integrative across the four complementary components rather than quarantined into individual STEM disciplines (Breiner et al., 2012; Bybee, 2010).

While integrating the four STEM components may be easy to conceptualize, implementing it is not as straightforward. As a result, many schools have launched what is known as the "STEM for All" approach. The intent of "STEM for All" is to provide high-quality STEM education to all K-12 students throughout their schooling (Atkinson, 2012; Basham, Israel, & Maynard, 2010). Applying the "STEM for All" approach requires an increase in K-12 STEM teacher quality, the development and application of consistent and rigorous STEM standards, and a change to existing STEM curricula to better enhance students' awareness of STEM careers, all of which demand a significant

amount of time and money to accomplish (Atkinson, 2012; President's Council of Advisors on Science and Technology, 2010).

Some researchers support a more targeted approach in which STEM teaching and learning is dedicated to students who have an interest in STEM (Atkinson, 2012; Gonzalez & Kuenzi, 2012; National Academy of Sciences et al., 2007; Olszewski-Kubilius, 2009). Within this framework, resources are directed towards specialized STEM schools, such as the 86 member institutions of the National Consortium of Secondary STEM Schools (National Consortium of Secondary STEM Schools, 2018). These types of schools recruit students who are interested in STEM and have demonstrated potential to succeed in the field.

In these specialized STEM schools, students are motivated to "survive" the STEM education pipeline, with a challenging curriculum, expert instruction, and stimulation from their peers (Bloom & Sosniak, 1985; Pyryt, 2000; Subotnik et al., 1993; Tai et al., 2006). Afterwards, students are prepared to contribute to the expanding U.S. economy upon entering the workforce (Atkinson, 2012; Gonzalez & Kuenzi, 2012). However, the overall effectiveness and impact of these institutions on various academic outcomes remains largely unknown. As these public, state-supported, residential academies are expensive, state legislators and the public demand evidence of their impact prior to allocating funds and/or other support (Pfeiffer et al., 2010).

Implemented in 2001, the focus of the *No Child Left Behind* (NCLB) act was to provide all children with a quality education and the opportunity to reach their academic potential (U.S. Department of Education et al., 2004). Whether or not this legislation has

improved or hindered students' educational outcomes remains controversial, as NCLB has concentrated on those students disadvantaged and at risk for academic problems or failures (Gallagher, 2004; U.S. Department of Education et al., 2004). In response to this act and its accountability requirements, teachers began using class time to better prepare students to take state-level "high-stakes" assessments (Gallagher, 2004). However, formal assessments such as these tend to be written at a grade-appropriate level, so that the reading level and complexity of the test is targeted to the population of interest (Clark & Watson, 1995; Gallagher, 2004; Mendoza, 2006; Nitko & Brookhart, 2011). As a result, researchers argue that the needs of gifted students are being overlooked, leaving them to work independently and learn on their own (DeLacy, 2004; Gallagher, 2004; Mendoza, 2006). If excellence and equity are goals in the U.S. education system, and these gifted students are considered the Nation's future thinkers, innovators, and leaders, an advanced, differentiated curriculum for gifted children is necessary (Gallagher, 2004; Grey, 2004; Mendoza, 2006; National Commission on Excellence in Education, 1983).

Gifted Education

Definitions and identification policies and procedures can substantially influence which individuals actually receive gifted services; however, no general consensus exists in describing and classifying these individuals. Prior research has defined giftedness as a "developmental process that is domain specific and malleable" (Subotnik et al., 2011, p. 6). Others emphasize that giftedness is the manifestation of your potential talent through outstanding performance, innovation, and accomplishments in the real world (Erwin & Worrell, 2012; Subotnik et al., 2011).

Similar to these broad definitions, the National Association for Gifted Children states that children are considered gifted when their ability is significantly above the norm for their age (National Association for Gifted Children, 2018). Furthermore, McClain and Pfeiffer (2012) remarked that there can be substantial differences in the definition and identification of giftedness by individual states. Since the high school in the current study is located in the state of Illinois, the following definition of giftedness is applicable:

> "…children and youth with outstanding talent who perform or show the potential for performing at remarkably high levels of accomplishment when compared with other children and youth of their age, experience, and environment. A child shall be considered gifted and talented in any area of aptitude, and, specifically, in language arts and mathematics, by scoring in the top 5% locally in that area of aptitude" (General Assemby of the State of Illinois, 2005).

As evidenced by the definitions above, the concept of giftedness has always included high intelligence and/or exceptional performance. As a result, the identification of gifted students continues to be dominated by the use of achievement and/or IQ test scores (Brown et al., 2005; Ford, 1998; Ford & Grantham, 2003). In fact, 45 of the 50 U.S. states use an achievement or IQ test score such as the SAT or the Stanford-Binet or Wechsler Intelligence scales to screen and identify gifted students (Erwin & Worrell, 2012; Ford, 1998; McClain & Pfeiffer, 2012). More specifically, 33 of these states mandate the use of intelligence or achievement tests to identify gifted students (McClain & Pfeiffer, 2012). While a majority of states use measures of exceptional performance to

identify gifted students, it is unclear whether or not the scores from these assessments are the only piece of information used in the identification process.

The overarching purpose in identifying gifted and talented individuals is to select those students who are excelling academically in addition to those students who have the potential to do well. Therefore, researchers have continued to advocate for the use of multiple measures so that certain populations do not become over- or under-represented in these specialized programs (Brown et al., 2005; Erwin & Worrell, 2012; Ford, 1998; Renzulli & Smith, 1977; Schmeiser, 1995; Subotnik et al., 2011). Furthermore, organizations that publish and develop standardized tests recognize the value of educational assessments, but still convey the importance of using multiple measures to provide complementary or confirmatory information during the decision-making process (Harris, 2003; McClain & Pfeiffer, 2012; Wattenbarger & McLeod, 1989).

Identification processes that use several types and sources of information (i.e., quantitative and qualitative) have the potential to be more rigorous in assessing the observed and expected abilities of individuals from all backgrounds (Erwin & Worrell, 2012; Ford, 1998; Renzulli & Smith, 1977). According to the state of Illinois, schools that plan to serve gifted students through specialized programs must demonstrate the use of at least three assessment measures, including instruments specifically designed to identify gifted students from underrepresented populations (Illinois State Board of Education, 2014). The high school in the current study uses four assessment measures in its application process: (1) Student essays describing their interests in STEM, (2) Two letters of recommendation, (3) Middle school and/or high school transcripts, and (4)

Current SAT (i.e., formerly known as the Scholastic Aptitude Test or the Scholastic Assessment Test) scores (College Board, 2018b). These measures provide those who review the applications with multiple sources of information in order to recommend a student for acceptance into the high school's gifted residential program focused on STEM.

Placement Testing

Although research has not extensively examined placement testing from middle school to high school, a large literature base exists using college and university student populations. In fact, approximately 90% of post-secondary institutions use placement tests (Latterell & Regal, 2003). The near-universal practice of administering placement tests emerged due to the incomparability of unknown factors such as the content and rigor of courses and the grading scales used at different schools (Kossack, 1942; Linn, 1994; Ngo & Kwon, 2015; Noble, Schiel, & Sawyer, 2003). Within the setting of a post-secondary institution, students complete placement tests to determine the appropriate level of beginning coursework. In the same way, once students are accepted into the high school of the current study, they too must complete a series of placement tests to guide their initial course enrollment decisions.

The overarching purpose of placement tests is to match students with a level of instruction that is appropriate given their previous academic preparations (e.g., Akst & Hirsch, 1991; Frisbie, 1982; Marshall & Allen, 2000; Mattern & Packman, 2009; McFate & Olmsted III, 1999; Noble et al., 2003; Sawyer, 1996). Prior research has shown that course placement decisions can have a significant impact on a student's future academic

preparation (McDaniel, Roediger, & McDermott, 2007; Morgan & Michaelides, 2005). For example, students who begin post-secondary mathematics in a course that is appropriate given their background have an increased chance of succeeding in their first course in addition to subsequent mathematics courses (Mattern & Packman, 2009; Norman et al., 2011; Shaw, 1997). For this reason, more research is needed to thoroughly examine placement tests and procedures to ensure that student success is maximized while the consequences of misplacement are minimized. Although these placement tests are typically considered "high-stakes," the psychometric properties of such tests have received relatively little attention (Callahan, 2005; Grubb & Worthen, 1999; Scott-Clayton, 2012). As a result, more research is needed to investigate and evidence the psychometric properties of placement tests.

According to the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2005), test developers are charged with the responsibility to: (1) Provide evidence of what the test measures, its recommended uses, and its strengths and limitations, and (2) Provide evidence that the technical quality (i.e., reliability and validity) of the test meets its intended uses. Additional research has recommended that colleges and universities consider the rigor and defensibility of the policies and methods used to inform placement decisions due to their "high-stakes" classification (Clark & Watson, 1995; Morgan & Michaelides, 2005). Armstrong (1995) stated that both Title V and Federal Civil Rights legislation requires institutions to validate the use of assessment tests in the placement and referral of students. Therefore, regardless of educational level,

future research should continue to identify the psychometric properties of placement tests in order to address questions about the impact of these tests on students and learning.

Within the context of educational measurement and placement decisions, point-to-point theory suggests that the best indicator of future behavior/performance is an individual's past behavior/performance (Belfield & Crosta, 2012; Davis & Shih, 2007; Erwin & Worrell, 2012; Feldhusen & Jarwan, 1995). However, one of the major concerns in previous literature has been the discrepancy between the cognitive behaviors and performances elicited on the placement tests and the cognitive behaviors and performances evaluated in the classroom (Armstrong, 2000; Brown & Niemi, 2007; Madison et al., 2015; Marsh, Roediger, Bjork, & Bjork, 2007; Schmitz & delMas, 1991). For example, if a test forbids the use of a calculator, the score obtained from that test may not accurately predict a student's ability to succeed in a mathematics course that encourages the use of calculators (Akst & Hirsch, 1991). Moreover, point-to-point theory postulates that Predictive (i.e., Criterion-Related) Validity is enhanced when the correspondence between what is measured on a test is congruent with what is needed to succeed in a course (Armstrong, 2000).

Prior research has attempted to examine this relationship by investigating the Predictive Validity of post-secondary placement exams in relation to the course grade received. As previously mentioned, the use of multiple measures is encouraged and provides more accurate course placement decisions compared to test scores alone (e.g., Armstrong, 1995; Erwin & Worrell, 2012; Marwick, 2004; Ngo & Kwon, 2015; Noble et al., 2003). For example, one study showed that combining the Mathematics SAT exam

with either high school GPA (i.e., grade point average) and/or class rank was a better predictor of college achievement over test scores alone (Schumacher & Smith, 2008). However, other studies have cautioned that the usefulness of the Mathematics SAT exam is limited due to the average difference in scores between males and females (Bridgeman & Wendler, 1989, 1991; Davis & Shih, 2007; Gallagher & De Lisi, 1994). More recent research has concluded that the accuracy of placement decisions greatly increases when placement test scores are combined with measures of high school achievement (i.e., high school GPA, high school grades, courses taken; Marwick, 2002; Melguizo et al., 2014; Ngo & Kwon, 2015; Pike, 1991; Scott-Clayton, 2012; Wattenbarger & McLeod, 1989). Although the use of multiple measures have been demonstrated to enhance placement policies and decisions at the post-secondary level, additional research is sought after at the high school level.

Item Bias

Among other types of analyses, Differential Item Functioning (DIF) can be used to detect item bias. DIF occurs when respondents from two groups (i.e., reference and focal group), who are said to be equal on the latent trait, have different probabilities of endorsing an item (Crocker & Algina, 2008; De Ayala, 2009; Hays, Morales, & Reise, 2000). After matching the two groups on their proficiency of the latent trait, the item response function (i.e., item characteristic curve) for each subgroup can be graphed simultaneously to determine if an item is biased. If an item presents with DIF, then there will be a separation between the two curves, as shown in Figure 1 below.

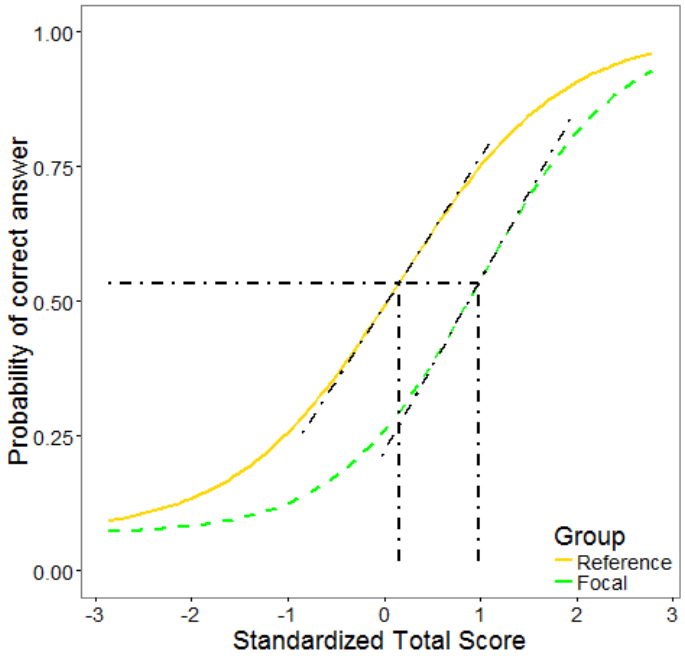

Figure 1. Differential Item Functioning. The above figure is an example of an item exhibiting bias between the reference and focal groups, favoring the reference group (Martinkova, 2016).

In general, instruments such as placement tests should be free from bias due to characteristics irrelevant to the construct of interest (i.e., sex, race, ethnicity, socio-economic status, age) in addition to producing reliable and valid scores (Schmeiser, 1995). Mattern and Packman (2009) reaffirmed the importance of examining whether placement decisions based on test scores are equally valid for males and females. Historically, the field of mathematics has been dominated by men and since the early 1980s, males have continued to take more advanced mathematics courses in high school compared to females (Catsambis, 1994; Pedro, Wolleat, Fennema, & Becker, 1981). Additionally, research has found that males outperform females on standardized

assessments such as the mathematics subtests of both the SAT and ACT (Bridgeman & Wendler, 1989, 1991; Davis & Shih, 2007; Educational Testing Service, 1989; Gallagher & De Lisi, 1994). However, another study concluded that gender differences in middle school mathematics coursework and performance on exams was minimal (Gallagher & De Lisi, 1994).

Similar to the placement testing literature, a majority of the research regarding item bias has been conducted at institutions of higher education. Further research is needed to examine whether or not there are significant differences in coursework and performance on standardized assessments throughout adolescence for characteristics such as sex, race, ethnicity, and/or socioeconomic status. More specifically, at a gifted residential STEM high school with a strong commitment to gender equity, additional research is needed to examine the presence of item bias on a mathematics placement exam.

Numerous psychometric studies have been conducted to examine individual mathematics placement tests for items exhibiting DIF. If an item presents evidence of DIF, further investigation is needed to warrant discarding the item. On the other hand, if item bias is not evidenced throughout the placement test, the exam and the placement decisions from the scores are equivalent for both subgroups of the population (i.e., males and females). Although previous literature can provide insight, issues related to item bias are specific to the instrument used and the conditions under which it is administered (i.e., participants, location, culture, etc.). For this reason, additional examination of the

mathematics placement test used in the current study is critical to determining whether or not the items on this particular instrument exhibits DIF.

Summary

This study aims to identify the psychometric properties of a mathematics placement test at a residential high school focused on STEM for gifted students. More specifically, this study seeks to provide evidence of reliability and validity, in addition to examining the characteristics of the item parameters (i.e., item difficulty, and item discrimination) and item bias with regards to sex. In light of these objectives, this chapter reviewed the existing literature related to STEM education, gifted education, and placement testing policies and practices, including item bias.

A brief history of STEM education was presented and summarized to illustrate the origins and more recent movements of the field, which included the development of specialized STEM high schools. In addition, a description of the past and present mathematical achievements of the U.S. were discussed to draw attention to the gap in STEM education and students' interest in STEM. However, by creating enriching environments for students interested and talented in science and mathematics, the leak in the STEM education pipeline can be minimized.

Next, the concept of giftedness and gifted education were introduced to demonstrate the varied definitions and identification processes that are currently used. While previous identification policies were centered about the use of achievement and/or IQ test scores, current practices for identifying gifted students have expanded to

incorporate the use of multiple measures, similar to the admission and placement practices of the high school under study.

Lastly, this Literature Review summarized the purposes for and widespread use of placement testing. Several studies indicated the impact of course placement decisions on the future academic potential of students and the importance of evaluating the psychometric properties of such exams used in the decision-making process. Finally, studies were cited that focus on placement exams at the post-secondary level and established the foundation needed to investigate a mathematics placement test used at the high-school level. The following chapter (i.e., Chapter Three) delineates the methodology in this study.

CHAPTER III

METHODOLOGY

The overarching purpose of this study is to investigate the psychometric properties of a mathematics placement test at a residential high school focused on Science, Technology, Engineering, and Mathematics (STEM) for gifted students. More specifically, the four aims of this study are: (1) To provide evidence of Content Validity, (2) To provide evidence of Construct Validity and Internal Consistency Reliability, (3) To examine the item characteristics and potential bias of the items between males and females and (4) To provide evidence of Criterion-Related Validity by investigating the ability of the mathematics placement test scores to predict future performance in an initial mathematics course. Existing data was used to address the following research questions:

Research Question 1 (RQ1): What is the Content Validity of the items on a mathematics placement test for gifted, residential high school students interested in STEM?

Research Question 2 (RQ2): What are the psychometric properties of the scores on a mathematics placement test for gifted, residential high school students interested in STEM?

RQ 2A: What is the Construct Validity of the scores on a mathematics placement test for gifted, residential high school students interested in STEM?

RQ 2B: What is the Internal Consistency Reliability of the item scores on a mathematics placement test for gifted, residential high school students interested in STEM?

Research Question 3 (RQ3): What are the item characteristics (i.e., item parameters and Differential Item Functioning [DIF]) of the mathematics placement test for gifted, residential high school students interested in STEM?

RQ 3A: What are the item parameters (i.e., item difficulty, and item discrimination) of the mathematics placement test for gifted, residential high school students interested in STEM?

RQ 3B: How do the items on a mathematics placement test for gifted, residential high school students interested in STEM differ by sex?

Research Question 4 (RQ4): What is the Criterion-Related Validity of the item scores on a mathematics placement test for gifted, residential high school students interested in STEM?

The subsequent sections provide background information regarding the context and instrument that were used throughout this study. Following this general information is a detailed discussion regarding the participants, procedures, data, and data analyses, if applicable, that were used to address each specific research aim listed above.

Context

The current study's existing data are from one high school campus for academically gifted students in the state of Illinois. Per the mission statement of this institution, it strives to be a teaching and learning laboratory that enrolls academically

talented Illinois students (i.e., Grades 10 through 12) in its advanced, residential college preparatory program with an emphasis in the fields of science and mathematics. In order to attend, students are required to submit an admissions application which includes an essay describing the student's interest in STEM, two letters of recommendation, middle school and/or high school transcripts, and current SAT (i.e., formerly known as the Scholastic Aptitude Test or the Scholastic Assessment Test) scores. As such, the admissions process is highly competitive as students from around the state of Illinois vie for approximately 250 positions each year.

For those students who are invited to attend, the high school provides a diverse and challenging curriculum designed to prepare students for college. Not only does the curriculum include the core subjects of English, history, social sciences, science, and mathematics, but students can also choose to take a course in the fine arts, wellness, or one of the six world languages offered. Additionally, students are provided the opportunity to conduct original and compelling research with expert scholars and scientists at more than 100 institutions. As a result, students graduating are well-rounded individuals equipped with the personal, social, and academic skills needed to succeed in college and beyond.

Measure

After the admissions review process, students are mailed either an acceptance, waitlist, deferral, or non-acceptance letter. For those students that are accepted or waitlisted, an informational flier is included with their admissions letter detailing when and where the mathematics placement test will be administered. This examination is

typically administered around mid-May with two location options. Students can either register to take the placement test on the high school campus or at a location further south in the state to reduce travel costs for students living further away. In either case, the mathematics placement test is proctored by either a mathematics faculty member or an admissions staff member. Both exam proctors are given a script to read verbatim to students prior to taking the test (see Appendices A and B).

The mathematics placement test was developed by mathematics faculty members of this institution in 1985. The original and continuing purpose of the mathematics placement test is to determine a student's incoming mathematical knowledge for appropriate initial course placement commensurate with ability level. Thus, generally speaking, the placement test assesses mathematical knowledge needed prior to entering into a Calculus sequence. More specifically, the developers of the exam created a two-part test measuring various content areas of mathematics, such as Algebra, PreCalculus, Trigonometry, and Geometry. However, neither these sections nor the test as a whole have been subjected to psychometric evaluation, specifically using more advanced quantitative techniques such as Exploratory Factor Analysis (EFA) or Item Response Theory (IRT).

Part I of the assessment largely measures student's knowledge of Algebra 1 content such as simplifying expressions, functions, and exponents. Students are given 45 minutes to complete 50 short-answer items, without a calculator. Assessing higher-level abilities such as the ability to solve numerical problems and/or to manipulate mathematical symbols and equations necessitates a short-answer question format (Nitko

& Brookhart, 2011). While the short-answer format allows students to show their work, the legibility of students' responses can at times complicate the scoring process.

All responses are graded by the mathematics faculty members using an answer key for dichotomous scoring (i.e., "Correct" or "Incorrect"). If a grader is unsure of a student's written response, other graders are consulted. In the event that a student's response cannot be determined, it is marked as an incorrect response. The possible range of scores on Part I is from 0 to 50. After the allotted time has expired for Part I, exam proctors collect any remaining exams and distribute Part II.

Part II of the assessment measures students' knowledge of PreCalculus, Trigonometry, and Geometry content. For this portion, students have 85 minutes to complete a total of 57 multiple-choice items, again without a calculator. The multiple-choice format used on this portion of the test provides students with the correct answer, three distractor answers, and a fifth response option of "I don't know." Although not explicitly written on the test instructions, mathematics faculty members emphasize the use of the "I don't know" option. By purposefully mentioning this, it is believed that students will not guess, but rather consider using the "I don't know" response option so that they do not accidentally place into a higher course than academically appropriate. A similar argument was made by Prieto and Delgado (1999) who noted that educational standards should not be influenced by desired psychometric properties of a test. Said another way, if students are unsure of an answer, it seems more appropriate for them to omit the item rather than encouraging them to guess. After the exam is complete, the multiple-choice items are scanned into a grading software program using a scantron

reader where all items are scored dichotomously (i.e., "Correct" or "Incorrect"), even if the student selected the "I don't know" option. The possible range of scores is from 0 to 57 on Part II.

Types of Missing Data

Before detailing each statistical technique by research question, the multiple types of missing data in this study are outlined and considered. Specifically, the following paragraphs specify how the missing data were addressed throughout the data analysis procedures. If the issue of missing data is not properly addressed, analysis of these data may become biased leading to inaccurate results, conclusions, and implications (e.g., Bennett, 2001; Chen, Wang, & Chen, 2012; Robitzsch & Rupp, 2009; Rose, Davier, & Xu, 2010).

Generally speaking, missing data are present in educational assessments for a variety of reasons. For example, a respondent may forget to return to a skipped item, be unwilling to guess, or experience testing fatigue (Ludlow & O'Leary, 1999; Widaman, 2006). In this particular study, there are three types of missing data that are discussed – omitted items, non-reached items, and the use of "I don't know" as a response option.

Omitted items. As previously mentioned, the mathematics placement test has two parts, short-answer and multiple-choice. Thus, the classification of omitted items were defined in two distinct, but similar ways. First on the short-answer section, omitted items are interpreted as items that have nothing written in the space provided and are completely blank. Similarly, for the multiple-choice section, omitted items are those that have no response on the scantron sheet (i.e., none of the "bubbles"/circles next to the

response options are filled/marked for a particular item). Typically, these omitted (i.e., blank) items occur within the body of the test and are less likely to occur at the end.

When items at the end of a test are left unanswered, these are typically classified as non-reached items. A non-reached item is one in which a respondent does not have the opportunity to answer an item, usually due to time constraints, as opposed to an omitted item where a respondent skips an item by mistake or consciously decides not to provide an answer (Ludlow & O'Leary, 1999). For this reason, omitted and non-reached items are considered to be independent from one another, yet similar in the way they are approached statistically. Consideration of the statistical controls for omitted items are presented followed by a description of the non-reached items and the use of the "I don't know" response.

One method used to address an omitted item is to score the response as incorrect. Various studies have investigated this possibility and have determined that marking omitted items as incorrect heavily distorts item parameters (Rose et al., 2010) and may negatively bias estimates of ability (Culbertson, 2011). As a result, researchers suggest that omitted items be ignored rather than coded as incorrect (e.g., Culbertson, 2011; Custer, Sharairi, & Swift, 2012; De Ayala, Plake, & Impara, 2001; Robitzsch & Rupp, 2009). In the current study, omitted items as defined above, are scored as incorrect by the mathematics faculty members. Thus, to remain consistent with the scoring procedures used, omitted items were coded as missing "M" and then re-coded as incorrect "0" for the selected statistical analyses. A table detailing the item frequencies

and the percentage of omitted responses per item is presented in the third manuscript regarding item analysis.

Non-reached items. As noted earlier, the main difference between an omitted item and a non-reached item is the location within the test where the non-response occurs. The National Assessment of Educational Progress guideline for non-reached items is as follows: "…if the last two or more items are left blank, then the first item of the string is to be treated as incorrect (presumably the student was working on it when time ran out) and the remaining would be treated as not reached" (Ludlow & O'Leary, 1999). This guideline, however, does not take into account the possibility that the respondent just completed the item they were working on when time ran out, rendering all of the remaining items unreached. With the assumption that the causes of particular response patterns are typically unknown, this study will consider all omitted items at the end of each part of the test as non-reached (coded as "NR"). Similar to omitted items, the mathematics faculty members score non-reached items as incorrect. Thus, although initially these items were coded as non-reached ("NR"), they were then re-coded as incorrect ("0") items throughout the various statistical analyses.

"I don't know" response. The third type of missing data addressed in this study is a result of respondents selecting the "I don't know" option on the multiple-choice section of the mathematics placement test. Since the early 1970s, researchers and statisticians alike have continued to argue the advantages and disadvantages of offering such a response option. Some claim that the "I don't know" response option may be informative and thus should be included within the estimation model (Balcombe &

Fraser, 2011). Others propose that the "I don't know" option is not suitable for tests measuring respondent's optimal performance. This response option is considered unsuitable because when respondents differ in their willingness to guess or to select "I don't know," respondents with identical levels of knowledge will receive different scores (Hanna, 1974; Mondak, 2001). Furthermore, Mondak (2001) cautioned that to either discourage guessing and/or to encourage "I don't know" responses, is to seek reliability at the cost of validity.

On the other hand, test developers and administrators will advocate for the use of the "I don't know" option as a way to reduce guessing. A compromise for this was proposed by Zhang (2013) who noted that if it is the intention of the test to minimize guessing and measure precise knowledge, then the "I don't know" option could be used within a penalty scoring model. Another suggestion to address the use of the "I don't know" option was to eliminate the "I don't know" response on multiple-choice questions by using a post-hoc correction (Kline, 1986; Mondak, 2001). In this post-hoc correction, the "I don't know" responses are randomly assigned to the remaining four choices, essentially entering guesses on behalf of the respondents who would not do so themselves (Mondak, 2001).

When a respondent selects the "I don't know" response option, the mathematics faculty members assume that the student is openly admitting to a lack of knowledge on a particular item. Prieto and Delgado (1999) made a similar argument noting that educational standards should not be compromised due to the desired psychometric properties of a test. In other words, if a student is not confident about a particular answer,

then it seems more appropriate to omit the answer rather than guess. In the description of the measure, the original and continuing purpose of the mathematics placement test is to determine students' incoming mathematical knowledge to make the appropriate course placement. Based on this goal of measuring optimal performance, the post-hoc correction or a penalty scoring model are inappropriate due to the differences in individuals' willingness to guess.

When students vary in their willingness to guess, then two students with the same ability level will receive different scores (Culbertson, 2011; Hanna, 1974; Mondak, 2001; Pohl, Gräfe, & Rose, 2014). In this instance, the test is no longer measuring only knowledge of mathematics, but also students' "test-wiseness." Furthermore, by using the post-hoc correction, the researcher is essentially entering a guess on behalf of those students who would not do so themselves (Mondak, 2001). However, if the intention of the placement test is to measure students' maximum performance in mathematics, then all possible sources of measurement error should be reduced to ensure the proper course placement.

As noted previously, the multiple-choice section is scored with a scantron reader using dichotomous "Correct"/"Incorrect" scoring, regardless of whether or not the respondent chose an incorrect choice or the "I don't know" option. For these reasons, the "I don't know" option was ultimately coded as an incorrect response ("0"). During initial data entry, however, the "I don't know" option was coded as "DK" so that information could be collected regarding the frequency of selecting this option per item.

Treatment of missing data. To summarize, three types of missing data were present within this study, namely omitted items, non-reached items, and the use of the "I don't know" response option. Regardless of the missing data initial classification, each type was re-coded as an incorrect response prior to implementing the various statistical analyses to remain consistent with the scoring procedures used by the mathematics faculty members who graded the placement test. The following paragraphs summarize the research questions and provide a detailed description of each study objective and corresponding statistical technique.

Research Aim 1

The goal of Manuscript 1 was to provide evidence of Content Validity of the mathematics placement exam at a gifted residential high school focused on STEM. Content Validity addresses whether or not items on an instrument (i.e., the words/statements comprising the items) and the meaning of these items measures a performance domain for a construct of interest (Cook & Beckman, 2006; Crocker & Algina, 2008; Ebel, 1956; Grant & Davis, 1997; Haynes, Richard, & Kubany, 1995; Martone & Sireci, 2009; Sireci, 1998a). Content Validity contains three subcomponents related to the domain: (1) Definition, (2) Representation, and (3) Relevance. Domain definition refers to the operational definition of the content domain describing both the content areas of interest and the levels of cognition required (Sireci, 1998a). The second and third subcomponents, Domain representation and Domain relevance, require the subjective evaluation of subject matter experts (SMEs). For Domain representation, SMEs are asked to judge whether or not the test items adequately represent the content

and cognitive specifications (Sireci, 1998a). In a similar way, SMEs appraise the relevance of each test item to the primary content domain when examining domain relevance. Although previous literature incorporates varying terminology, such as content domain sampling, content representation, or content relevance, the related definitions remain the same. Overall, evidence that a test adequately represents the underlying content domain remains a vital component to test development and construction (Sireci & Geisinger, 1992).

Former Content Validation studies have used a variety of methods to evaluate item similarities and relevance. Two of the most recognized techniques are item-pairing and item-sorting tasks. In studies by Sireci and Geisinger (1992, 1995), researchers asked SMEs to rate the similarity of a given item-pair on a scale from "Highly Similar" (Coded 1) to "Highly Dissimilar" (Coded 10). In a similar way, SMEs were asked to rate the degree of each item's relevance to the content areas listed (Sireci & Geisinger, 1992, 1995). One year later, Deville and Prometric (1996) used a similar item-pairing task. While the item-pairing technique can provide a more comprehensive examination of content domain representation, it can quickly become burdensome for SMEs to judge when the number of items become too large. For example, the mathematics placement test in the current study consists of 107 total items. If the item-pairing task was used, SMEs would be asked to rate item-similarities for 5,671 unique item-pairs. Not only is this an unrealistic task for an individual to complete, but it is also detrimental to the recruitment of SMEs. Additionally, prior research has suggested the use of sorting procedures requiring SMEs to sort items into a limited number of categories according to

their similarities (Sireci & Geisinger, 1995). The same study also suggested that item-level data be obtained to determine how Factor Analysis or Multidimensional Scaling (MDS) results compare to the dimensions obtained from the SME similarity ratings.

For these reasons, the current study employed a card-sorting task to gather data on the test's content areas. Adopted from a study by D'Agostino, Karpinski, and Welsh (2011), MDS and Hierarchical Cluster Analysis was used to compare the similarity ratings of external SMEs to the similarity ratings of internal SMEs. Generally, when using MDS in Content Validity studies, similarity ratings from SMEs are compared to the original test specifications (D'Agostino et al., 2011; Li & Sireci, 2013; Sireci & Geisinger, 1992, 1995). One dilemma in the current study was the absence of test specifications. However, prior research has demonstrated the complementary use of MDS and Hierarchical Cluster Analysis in the development of content specifications for professional certification exams (Raymond, 1989; Schaefer, Raymond, & Stamps White, 1992). Thus, the design of the current study made use of internal SME item-similarity ratings to develop the content specifications, which were then compared to external SME item-similarity ratings to provide evidence of Content Validity. A discussion of the procedures and data analysis techniques follow.

Participants

The recruitment and qualifications of SMEs is an important consideration in any Content Validation study. The number of SMEs needed for a content validation study will be driven by the range of representation and experiences desired by the researcher (Grant & Davis, 1997). As described previously, the context of the current study is

unique in that it occurs at a gifted residential high school focused on STEM. With its advanced curriculum and residential component, the high school is often times compared to an institution of higher education. However, because the school serves students in grades 10 through 12, it is categorized as a high school. Therefore, to properly assess the Content Validity of this school's mathematics placement test, SMEs at varying levels were recruited.

More specifically, both internal and external SMEs were needed. The external participants in the Content Validation procedures included high school mathematics teachers, high school mathematics teachers with experience teaching gifted students, and mathematics faculty members from community colleges and four-year institutions from across the state of Illinois. These external SMEs were recruited based on their interests, experiences, and contributions to STEM education. After the list was developed, approximately five to ten individuals from each group was contacted via email to be a prospective SME. This email included substantive details about the purpose of the study, the confidentiality of their responses and of the test items, the responsibilities of the participants (i.e., description of the card-sorting tasks and time required of the participant), and the associated risks and/or benefits (see Appendix C for a copy of the email invitation and Appendix D for a copy of informed consent). Additional follow-up recruitment emails were sent as needed.

Similar to the external SMEs, an email invitation was sent to SMEs within the high school. Since the original test specifications were unknown, judgments from internal SMEs were needed to compare responses with external SMEs. For both external

and internal SMEs, demographic information such as gender, race/ethnicity, highest degree awarded, number of years teaching, and courses commonly taught was collected (Appendix E). These data allowed for a basic description of group similarities between the internal and external SMEs.

Procedures

After consenting to participate, the SMEs were mailed rectangular strips of paper containing one test item per card (i.e., 107 total cards) along with directions describing the item-sorting task. The directions instructed each SME to place the items into meaningful piles or groups based on the similarity of the content of the items. Consistent with the sorting rules described by Trochim (1989), SMEs were advised to: (1) place each item or card into only one pile or group, (2) refrain from creating as many piles or groups as there are items, and (3) create more than one pile. Upon completion of the content card-sorting task, SMEs were then asked to record the item numbers in each pile on a piece of paper and to assign each group of items a group title or name (Appendix E). The SMEs then returned their content area groupings and the provided test items on strips of paper via a prepaid envelope.

Upon completion and return of the card-sorting task, each SME's coding sheet was transformed into an individual item-similarity rating matrix where the test item numbers were listed for both the rows and the columns. An entry of "0" indicated that the SME did not categorize a specific item-pair together, whereas an entry of "1" indicated that the SME did put the item-pair in the same group (D'Agostino et al., 2011).

Item #	1	2	3	4	5
1	1				
2	1	1			
3	0	0	1		
4	0	1	1	1	
5	1	0	1	0	1

Figure 2. Item-Similarity Matrix for a single subject-matter expert. This figure is an example of an item-similarity matrix for a single subject matter expert's response.

In Figure 2 (above), the "0" entry for the item-pair (3,2) indicates that through the card-sorting task, SME 1 places Items 2 and 3 into different groups or piles. In contrast, the "1" entry for item-pair (1,5) indicates that SME 1 placed Items 1 and 5 into the same group or pile. Furthermore, a "1" entry on the diagonal of the matrix indicates that the SME always categorized an item in the same pile or group as itself (D'Agostino et al., 2011).

After each individual item-similarity matrix was created, a group item-similarity matrix was constructed by adding the individual item-similarity matrices together (D'Agostino et al., 2011). Values of the group item-similarity matrix ranged from 0 to *n*, where *n* is the total number of SMEs. A value of "0" implies that none of the SMEs categorized the same item-pair together. The largest value *n*, representing the total number of SMEs, appears on the diagonal of the group item-similarity matrix indicating that all SMEs categorized each item with itself. Thus, a larger matrix cell value

represents a greater consensus of SMEs regarding the similarity of the items (D'Agostino et al., 2011).

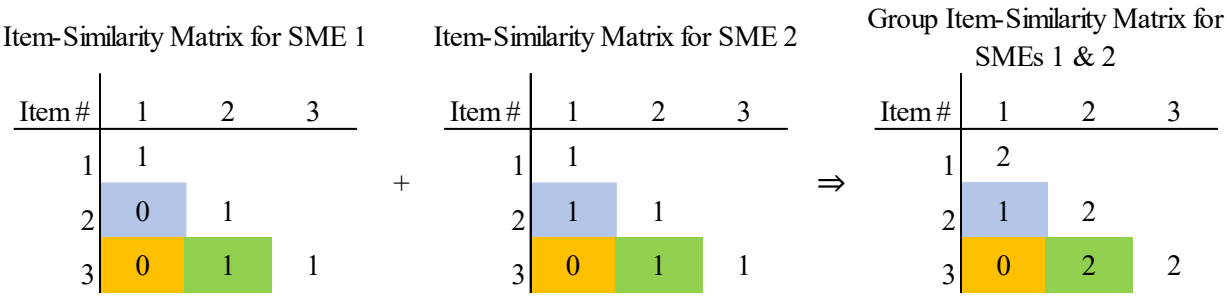

Figure 3. How to create a Group Item-Similarity Matrix. This figure shows how each subject matter expert's item-similarity matrix is combined to create the group item-similarity matrix needed for analysis.

For additional clarification, in the above example for SME 1 and SME 2, the "0" entry for item-pair (1, 3) signifies that neither SME 1 nor SME 2 placed Items 1 and 3 in the same group. An entry of "2" for item-pair (2, 3) demonstrates that both SME 1 and SME 2 placed Items 2 and 3 in the same group. In a similar way, an entry of "1" for item-pair (1, 2) indicates that either SME 1 or SME 2 categorized Items 1 and 2 into the same group, while the other SME did not. Once the group item-similarity matrix was compiled for both the internal and external SMEs, each group matrix was further transformed prior to Multidimensional Scaling and Hierarchical Cluster Analysis.

Since similarity and dissimilarity ratings are inverses of one another, researchers have recommended transforming similarity ratings into dissimilarity ratings prior to data analysis using SPSS (Jaworska & Chupetlovska-Anastasova, 2009; Kruskal & Wish,

1978). For the current study's purposes, the group item-similarity matrices for both internal and external SMEs were converted into group item-similarity ratios. Using a scale from 0 to 1, these ratios were transformed into a group item-dissimilarity matrix using the calculation of 1 - n_{jk} where n is the matrix cell value for the item-pair j and k where $j \neq k$.

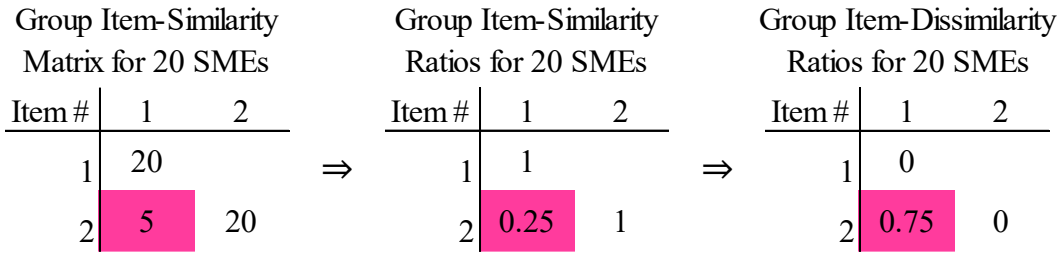

Figure 4. How to create a Group Item-Dissimilarity Matrix. This figure displays the ratio calculation process in order to transform the group item-similarity matrix into a group-item dissimilarity matrix.

As an example, in the above matrices, five out of 20 SMEs categorized Items 1 and 2 together to obtain the group item-similarity ratio, 5/20 = .25 (Figure 4). Finally, the group item-similarity ratio was transformed into a group item-dissimilarity ratio by using a constant, which in this case is 1 (i.e., 1 – .25 = .75). Thus, the final group item-dissimilarity matrix was used in the MDS analysis in SPSS.

Data Analysis

Multidimensional Scaling (MDS) has been used in a variety of fields such as medicine, psychology, psychometrics, and psychophysics due to its ability to accommodate various levels of data without restriction of multivariate normality. MDS

aims to uncover any structure or pattern in data by rescaling a set of similarity or dissimilarity measurements into distances assigned to specific coordinates within a spatial configuration (Agarwal et al., 2007; Jaworska & Chupetlovska-Anastasova, 2009; Mead, 1992; Raymond, 1989). Since MDS strictly relies on judgments of dissimilarity, there are no statistical distribution assumptions that must be met (Wilkinson, 2002). However, one must decide which metric will be used to calculate these distances (i.e., Euclidean, Minkowski's *p*, or City-block). Since the data in the current study were at the interval level, distances were estimated using the traditional Euclidean distance calculation as follows:

$$d_{ij} = \sqrt{\sum_{r=1}^{R}(x_{ir} - x_{jr})^2} \quad [1]$$

where x_{ir} and x_{jr} are the coordinates of points *i* and *j*, respectively, on dimension *r*, in a *R*-dimensional spatial representation (e.g., Arce & Gärling, 1989; Carroll & Arabie, 1980; Davison & Skay, 1991; Giguère, 2006; Jaworska & Chupetlovska-Anastasova, 2009; Steyvers, 2002).

Once the group dissimilarity matrix was analyzed using MDS, the output was interpreted. Interpretation of MDS output includes determining the appropriate number of dimensions to retain. This selection of dimensions is primarily based on three considerations: (1) the values of the fit indices, (2) the amount of change in fit indices from *n* to *n – 1* dimensions, and (3) the interpretability of the dimensions (Whaley & Longoria, 2009). Each of these were examined in the current study to determine the final MDS solution for both internal and external SME responses.

The two fit indices that were used were Kruskal's Stress Function (Kruskal, 1964) and the Squared Correlation Index (R^2). Similar to other goodness-of-fit indices, Kruskal's Stress Function is a calculation of the residual sum of squares (Kruskal, 1964). As such, smaller values indicate a better fit between the data and the MDS solution. For the purposes of this study, the following stress values were used as guidelines: $S = 0$ suggests perfect fit; $0 < S \leq .025$ suggests excellent fit; $.025 < S \leq .05$ suggests good fit; $.05 < S \leq .10$ suggests fair fit; and $S \geq .20$ suggests poor fit (Kruskal, 1964). Secondly, R^2 values are interpreted as the proportion of variance explained by the disparities (Hair Jr, Anderson, Tatham, & Black, 1995; Whaley & Longoria, 2009). In other words, R^2 measures how well the MDS model fits the original data, implying that higher values indicate better fit. In the current study, the MDS solution was considered an acceptable fit if $R^2 \geq .60$ (Hair Jr et al., 1995; Whaley & Longoria, 2009).

Next, to examine the amount of change in fit indices from n to $n - 1$ dimensions, a plot similar to Cattell's Scree Test (Cattell, 1966) was used. The stress values were graphed on the y-axis with the number of dimensions in decreasing order on the x-axis (Hoand, 2008; Jaworska & Chupetlovska-Anastasova, 2009; Whaley & Longoria, 2009). The resulting graph was analyzed for an "elbow" among the data. At this point, the change in stress between one dimension and the next was considered negligible, indicating a possible final MDS solution. Finally, the interpretability of the MDS solution, and its associated number of dimensions, were considered when determining the final solution for both internal and external SME responses.

After the final MDS solutions had been identified, the item coordinates from those solutions were analyzed using Hierarchical Cluster Analysis. MDS and Hierarchical Cluster Analysis are complementary techniques in that MDS graphically displays relationships among items, whereas clustering examines which items group together and why. By imposing Hierarchical Cluster Analysis on the MDS solutions, the domain structure of the internal SMEs and external SMEs can be compared. Additionally, the degree of consensus between the two domain configurations can ultimately be determined (D'Agostino et al., 2011; Sireci & Geisinger, 1992).

Because the purpose of cluster analysis is to group objects (i.e., items or responses) according to particular characteristics they possess, the resulting clusters should have high internal (within-cluster) homogeneity and high external (between-cluster) heterogeneity (Hair Jr et al., 1995). In the current study, a Hierarchical Cluster Analysis was conducted using the agglomerative clustering method. In this method, all objects or items are assigned to their own cluster. Then through an iterative process, the two most similar objects, not already in the same cluster, are combined (Hair Jr et al., 1995; Sarstedt & Mooi, 2014). This process continues until all objects are in one large cluster.

Similar to the MDS analysis, the Euclidean metric was used to calculate the distances between objects within clusters. Smaller distances suggested a greater similarity between objects. Moreover, the average-linkage clustering algorithm was used. This algorithm defines the distances between two objects as the average distance between all pairs of members within the clusters (Hair Jr et al., 1995; Johnson, 1967;

Sarstedt & Mooi, 2014). Thus, this method is less influenced by outliers and the cluster boundaries are determined using all members within a cluster rather than a single cluster member.

Similar to previously mentioned analytics procedures, there is some subjectivity in determining how many clusters to retain and the interpretation of those clusters. Researchers must consider the cluster structure and interpretation in addition to within cluster homogeneity (Hair Jr et al., 1995). Therefore, a dendrogram (i.e., tree graph) was analyzed to explore the changes in the distances between clusters. Additionally, a Scree Plot was created by graphing the number of clusters on the x-axis against the distances at which the clusters are combined on the y-axis. Then, similar to the Scree Plot for eigenvalues, this plot was examined for an "elbow" to indicate the number of clusters to be retained.

Once the final cluster solutions had been determined for both the internal and external SME responses, the two configurations were compared using the Rand and adjusted Rand indices. The Rand index computes the overlap between classification schemes, while the adjusted Rand index controls for overlap by chance due to marginal distributions (D'Agostino et al., 2011). Both indices are reported on a scale from 0 to 1, with higher values indicating a stronger overlap.

Research Aim 2

The goal of Manuscript 2 was to provide evidence of Construct Validity and Internal Consistency Reliability. Construct Validation refers to a process by which a judgment is made regarding whether or not an instrument adequately measures the

intended construct. A construct, also referred to as a latent variable, is not directly observable and has been defined as "some postulated attribute of people, assumed to be reflected in test performance" (Cronbach & Meehl, 1955, p. 283). Commonly studied psychological constructs include anxiety, achievement, and personality. In order to measure a construct of interest, researchers emphasize the need to transform a conceptual definition into an operational definition. The operational definition acts as a bridge to connect the conceptual definition to more concrete observations or indicators. These observations are then assigned numbers to represent how much of the construct an individual possesses.

Aspects of Construct Validation are typically reviewed during the instrument development phase. During this time, the construct of interest and its associated content are manifested into concrete tasks that individuals must complete. In the context of educational assessment, content standards of a course are translated into performance standards which further define "how much of the content standards students must know and be able to do to achieve a particular level of competency" (Morgan & Michaelides, 2005, p. 1). Four widely used approaches to Construct Validation are: (1) the use of correlations between the construct and other variables, (2) differentiation between groups, (3) Factor Analysis, and (4) the Multitrait-Multimethod Matrix (Campbell & Fiske, 1959; Crocker & Algina, 2008). In the current study, evidence of Construct Validity was obtained through an Exploratory Factor Analysis (EFA).

Measure

The mathematics placement test was developed by mathematics faculty members in 1985. The original and continuing purpose of the mathematics placement test is to determine a student's incoming mathematical knowledge for appropriate initial course placement commensurate with ability level. Thus, generally speaking, the placement test assesses mathematical knowledge needed prior to entering into a Calculus sequence.

Part I of the assessment mainly measures student's knowledge of Algebra 1 content such as simplifying expressions, functions, and exponents. Students are given 45 minutes to complete 50 short-answer items, without a calculator. Assessing higher-level abilities such as the ability to solve numerical problems and/or to manipulate mathematical symbols and equations necessitates a short-answer question format (Nitko & Brookhart, 2011). While the short-answer format allows students to show their work, the legibility of students' responses can at times complicate the scoring process.
All responses are graded by the mathematics faculty members using an answer key for dichotomous scoring (i.e., "Correct" or "Incorrect"). If a grader is unsure of a student's written response, other graders are consulted. In the event that a student's response cannot be determined, it is marked as an incorrect response. The possible range of scores on Part I is from 0 to 50. After the allotted time has expired for Part I, exam proctors collect any remaining exams and distribute Part II.

The main focus of Part II of the assessment is to measure students' knowledge of both PreCalculus and Geometry content. For this portion, students have 85 minutes to complete 57 multiple-choice items, again without a calculator. The multiple-choice

format used on this portion of the test provides students with the correct answer, three distractor answers, and a fifth response option of "I don't know." Although not explicitly written on the test instructions, mathematics faculty members and exam proctors emphasize the use of the "I don't know" option. By purposefully mentioning this, it is believed that students will not guess, but rather consider using the "I don't know" response option so that they do not accidentally place into a higher course than academically appropriate. A similar argument was made by Prieto and Delgado (1999) who noted that educational standards should not be influenced by desired psychometric properties of a test. Said another way, if students are unsure of an answer, it seems more appropriate for them to omit the item rather than encouraging them to guess. After the exam is complete, the multiple-choice items are scanned into a grading software program using a scantron reader where all items are scored dichotomously (i.e., "Correct" or "Incorrect"), even if the student selected the "I don't know" option. The possible range of scores is from 0 to 57 on Part II.

Participants and Procedures

Existing data from four cohorts of students was used to examine the research questions in this study. These cohorts consisted of students entering the high school their sophomore year, beginning in the 2014/2015 academic year and ending in the most recent 2017/2018 academic year, for which complete data was available.

Equivalence across the four cohorts was examined for five demographic variables using Chi-Square (χ^2) Tests of Association and One-Way Analyses of Variance (ANOVAs). Chi-Square Tests of Association were conducted across the four cohorts for

the variables of sex and race/ethnicity. There were no significant differences in the proportions between cohort year and either sex or race/ethnicity. For the three remaining variables of socioeconomic status (i.e., median family income), incoming SAT Math (SAT_M) subscores, and incoming SAT Evidence Based Reading and Writing (SAT_ERW) subscores, ANOVAs were used. Again, there were no significant differences between cohort years for each of the three variables. Therefore, all four cohorts were found to be statistically equivalent and were combined into one sample for further analysis.

Previous research has long debated the appropriate sample size to conduct an EFA, with approximately 10 subjects per variable as the general consensus (Comrey & Lee, 1992; Costello & Osborne, 2005; Nunnally & Bernstein, 1978). In the current study, there are 107 items from the mathematics placement test that were factor analyzed. Using the 10:1 subject to variable ratio guideline, 1,070 cases are needed to conduct the EFA. The sample size of the current study was 1,125 which surpassed the recommended 10:1 subject to variable ratio.

Data Analysis

Pett, Lackey, and Sullivan (2003, p. 2) describe factor analysis as "a complex array of structure analyzing procedures used to identify the interrelationships among a large set of observed variables and then, through data reduction, to group a smaller set of these variables into dimensions or factors that have common characteristics." The two broad classifications of factor analysis are Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). Researchers use EFA when the underlying factor

structure of the construct of interest is unknown (Pett et al., 2003; Thompson, 2004). CFA, on the other hand, is used when the researcher has some knowledge or understanding of the underlying factor structure from previous theories of the construct of interest. In the current study, the original factor structure of the mathematics placement test is unknown. Thus, an EFA was conducted using PRELIS and LISREL 9.30.

Assumptions. The main underlying assumption of EFA is that the observed variables are linear combinations of underlying hypothetical/unobservable factors (Kim & Mueller, 1978). The goal in this analysis is to condense the information contained in the original variables into a smaller set of factors with a minimal loss of information (Hair Jr et al., 1995). When discussing and analyzing linear combinations, mathematical theories and assumptions surrounding matrices are used.

Another assumption of EFA is univariate/multivariate normality, which refers to the shape of the distribution of data and its congruence to a normal distribution curve (Hair Jr et al., 1995). However, these assumptions were not considered within this study as the data were dichotomously scored. Similarly, a third consideration for conducting an EFA is the strength of the relationship between two items on an instrument. This information is typically summarized by the Pearson Product-Moment Correlation Coefficient Matrix, sometimes referred to as Pearson's *r* or the correlation matrix (Pett et al., 2003). Because the data are dichotomous, the strength of the relationship between two items on the instrument will be assessed using the Tetrachoric Correlation Matrix. Tetrachoric Correlation Coefficients are used when the latent trait underlying the data is theoretically continuous, but is measured dichotomously (Bonett & Price, 2005; Lorenzo-

Seva & Ferrando, 2012; Uebersax, 2006b). In this study, the underlying latent trait is mathematical knowledge, which is conceptualized as a continuous variable. However, this latent trait is scored dichotomously on the mathematics placement exam (i.e., scoring "Correct" or "Incorrect").

Furthermore, in order to use Tetrachoric Correlations, the following assumptions must be met: (1) the latent trait is normally distributed, (2) rating errors are normally distributed, (3) the variance is homogeneous across all levels of the latent trait, (4) errors are independent between items, and (5) errors are independent between cases (Uebersax, 2006b). The primary limitation of using Tetrachoric Correlations is that these assumptions cannot be mathematically tested.

The goal of factor analysis is to explain the interrelationships among variables, and it is important to have "acceptable" correlation coefficients. Various researchers have differing opinions on what constitutes an "acceptable" correlation coefficient, which is dependent upon the level of measurement of the variables (i.e., nominal, ordinal, interval, or ratio) and how the correlation coefficient is calculated. One generally accepted guideline for interpreting the Pearson Product-Moment Correlation Coefficient is that correlation values should be greater than or equal to .30 (Costello & Osborne, 2005; Pett et al., 2003; Stevens, 2012; Tabachnick & Fidell, 2007). Because the values of Tetrachoric Correlations values are interpreted similarly to Pearson's *r*, the above stated guideline was consulted when examining the Tetrachoric Correlation Matrix in the current study.

Exploratory factor analysis. Exploratory Factor Analysis (EFA) is considered to be "a complex procedure with few absolute guidelines and many options" (Costello & Osborne, 2005, p. 1). The following paragraphs describe the method of factor extraction, rotation, solution refinement, and final interpretation that were used in the current study.

When conducting an EFA, the determinant of the correlation matrix is evaluated to determine if an inverse matrix exists. If the determinant of the correlation matrix is zero, an inverse matrix does not exist, implying that there are no interrelationships between the items (Pett et al., 2003). The correlation matrix would, in this case, not be called an identity matrix. These calculations can all be summarized in what is known as Bartlett's Test of Sphericity (Bartlett, 1950). In a similar way, the Tetrachoric Correlation Matrix calculated with dichotomous data can have a property called non-positive definiteness (Uebersax, 2006a). This occurs when one or more eigenvalues are negative, suggesting that there are linear dependencies among some items (Lorenzo-Seva & Ferrando, 2020). When linear dependencies are present, this indicates that one or more eigenvalues are close to zero, meaning that the matrix is close to being non-invertible (Margalit & Rabinoff, 2018; Pett et al., 2003). Thus, when negative eigenvalues are present and the matrix is close to being singular (i.e., non-invertible), then the extraction methods of Maximum Likelihood (ML) and Generalized Least Squares (GLS) cannot be used because of their reliance on the inverse matrix. Furthermore, ML and GLS extraction methods were not used in this study due to their underlying assumption of multivariate normality. Instead, the factor extraction method of Unweighted Least

Squares (ULS) was used since its calculations do not rely on the inverse matrix or multivariate normality (Uebersax, 2006a).

Regarding the number of factors to be extracted, the two prominent methods used for EFA include the Kaiser-Guttman Rule for eigenvalues (e.g., Comrey & Lee, 1992; Guttman, 1954; Kaiser, 1960; Nunnally & Bernstein, 1994) and the Scree Plot (Cattell, 1966). The Kaiser-Guttman Rule tends to be more objective in that this method extracts those factors whose eigenvalues are greater than 1. On the other hand, examining the Scree Plot requires more of a subjective decision about where the elbow of the plot is located and consequently how many factors should be retained. For these reasons, researchers tend to use a combination of these methods in EFA to guide decisions regarding the number of retained factors. In the current study, the statistical software program PRELIS was used due to its ability to handle dichotomous data and calculate the Tetrachoric Correlation Matrix. One limitation of this program is that the Scree Plot method is unavailable. While PRELIS does allow the researcher to specify the number of factors to retain, there is little previous research and/or theory to support the number of factors to extract in the current study. Therefore, as EFA is an explanatory, theory-building data analytic strategy, this study used PRELIS to automatically determine the number of factors to extract based on the correlation matrix. Once the default number of extracted factors had been established, then additional iterations of the data specified how many factors to extract which were both above and below the defaulted amount.

The next consideration in model specification was whether or not to rotate the extracted factors, which aids in simplifying and clarifying the underlying data structure.

The two common approaches in data rotation are orthogonal and oblique, each having different underlying assumptions. An orthogonal rotation assumes that the underlying factors are uncorrelated, whereas an oblique rotation assumes the opposite (e.g., Costello & Osborne, 2005; Gorsuch, 1983; Pett et al., 2003; Thompson, 2004). Since the underlying latent trait is mathematical knowledge, it was expected that a relationship would be present among the underlying factors necessitating an oblique rotation. Of the possible oblique rotation methods (i.e., Direct Oblimin, Promax, Orthoblique), the Promax rotation was used in the current study. One advantage of the Promax rotation is that it begins with an orthogonal rotation, allowing for the possibility that the underlying factors are in fact uncorrelated (Pett et al., 2003). Additionally, Gorsuch (1983) argued that the Promax rotation ultimately results in stronger correlations between factors and achieves a more simple structure. Accordingly, the oblique rotation method Promax was used.

Using information from the above stated model specifications, the default factor extraction solution was examined for its representativeness and overall fit to the data. Again, since this was an EFA and the underlying factor structure was unknown, additional factor extraction solutions were explored and compared to the initial solution. In doing so, the final interpretation of the factor structure was supported through evidence from the collection of models, including but not limited to the amount of variance explained, the factor loadings, and the correlations between factors.

Internal consistency reliability. As noted earlier, reliability refers to the degree to which data collection, data analysis, and data interpretations are consistent provided

the surrounding conditions remain constant (Wiersma & Jurs, 2009). As such, Internal Consistency Reliability provides evidence of accuracy of results when the same measure is used. Moreover, "internal consistency" would suggest that the items within a measure correlate strongly with one another (Henson, 2001; Kimberlin & Winetrstein, 2008). In selecting the Internal Consistency Reliability method to use, Guttman Split-Half (Guttman, 1945), Coefficient Alpha (Cronbach, 1951), or the Kuder-Richardson Formulas (Kuder & Richardson, 1937), one consideration is how the items on the single test administration are divided. The following paragraphs provide a brief explanation of each reliability estimation method for a single test administration in addition to the rationale for the selected method in the current study.

The first class of methods for estimating the reliability coefficient is generally referred to as the Split-Half Methods. When using this method, the test is divided into two subtests of equal length (Crocker & Algina, 2008). Splitting a test into two equal parts can occur a number of ways such as grouping the items by their even or odd number, separating the first half from the second half, or by rank ordering the items by their difficulties and then assigning matching or similar items to the two halves. Regardless of the type of division, the purpose is to create two parallel tests which can then be scored individually per examinee. Afterwards, a correlation of equivalence can be calculated to provide an estimate of the reliability coefficient for the full-length test (Crocker & Algina, 2008).

One limitation of the Split-Half Method, however, is that the correlation coefficient obtained is usually underestimated as longer tests tend to be more reliable

than shorter tests (Crocker & Algina, 2008). In response to this issue, the Spearman-Brown Prophecy Formula (Brown, 1910; Spearman, 1910) is used to achieve the corrected reliability coefficient estimate of the full-length test. In a similar way, the Guttman Split-Half Method (Guttman, 1945) can be used to estimate the reliability coefficient of the full-length test by calculating the score differences between each half-test. Overall, the most noteworthy shortcoming of the Split-Half Methods is the non-unique reliability coefficient estimates (Crocker & Algina, 2008). There are multiple ways to split a test into two halves, each of which will produce a different reliability estimate.

The other category of methods for estimating reliability coefficients are based on the item covariances. Among this classification are the well-known methods that assess Internal Consistency Reliability – Coefficient (Cronbach's) Alpha and the Kuder-Richardson Formulas (Cronbach, 1951; Kuder & Richardson, 1937). As shown below, previous research has demonstrated the equality of Cronbach's Alpha and the Kuder-Richardson Formulas (e.g., Cliff, 1984; Crocker & Algina, 2008; Feldt, 1969; Onwuegbuzie & Daniel, 2002) in regards to the case of binary data. Cortina (1993) elaborated further by stating that Cronbach's Alpha is a more general version than the Kuder-Richardson estimate. Cronbach's Alpha can be calculated by using the formula

$$\widehat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \hat{\sigma}_i^2}{\hat{\sigma}_X^2}\right) \qquad [2]$$

where k is the number of items on the test, $\hat{\sigma}_i^2$ is the variance of item i, and $\hat{\sigma}_X^2$ is the total test variance. Likewise, with a simple substitution of pq for the variance of item i, the Kuder-Richardson estimate is calculated as follows:

$$KR_{20} = \frac{k}{k-1}\left(1 - \frac{\sum pq}{\hat{\sigma}_X^2}\right) \quad [3]$$

However, when items are dichotomously scored, although equal, the Kuder-Richardson Formula (KR-20) is preferred over Cronbach's Alpha.

Researchers Kuder and Richardson (1937) developed two formulas for estimating internal consistency reliability, namely the KR-20 and the KR-21. While computed similarly, the KR-20 and KR-21 formulas differ in their assumption of item difficulties. If each item is assumed to have the same level of difficulty, then the KR-21 formula can be used (Crocker & Algina, 2008; Kuder & Richardson, 1937; Onwuegbuzie & Daniel, 2002). The mathematics placement test in the current study was constructed to broadly measure the content areas of Algebra 1, PreCalculus, Trigonometry, and Geometry. Moreover, regardless of the factor structure results obtained in the EFA, Algebra 1 is generally viewed as prerequisite knowledge to PreCalculus. Thus, the current study assumed that the item difficulties vary, which necessitates calculating KR-20 as the estimate of internal consistency reliability.

Considerable attention has been given to the "acceptable" value range for Cronbach's Alpha or KR-20 indices. While an internal consistency reliability estimate of .70 may be "acceptable" in some contexts of exploratory research (Nunnally & Bernstein, 1978), L. Ding and Beichner (2009) suggested that the value of KR-20 be greater than or equal to .80. For Coefficient (Cronbach's) Alpha, researchers have continually emphasized the need for higher reliability estimates in educational settings. More specifically, when a particular test score is used for important clinical and/or educational decisions (e.g., course placement), the estimates of internal consistency reliability should

have a minimum value of .90, with .95 considered desirable (e.g., Henson, 2001; Hopkins, 1998; Nunnally & Bernstein, 1994; Oosterhof, 2001; Rossi, Lipsey, & Freeman, 2003). That is, when circumstances require a higher degree of confidence in the accuracy of interpretations, more evidence will be needed to demonstrate the internal consistency of a measure (Cook & Beckman, 2006). Since Cronbach's Alpha is equal to KR-20 with binary data, the abovementioned guidelines for "acceptable" values were used in this study. Therefore, a minimum internal consistency reliability estimate of .90 was considered the standard for the Mathematics Placement Test in the current study.

Finally, the term internal consistency suggests that items measuring the same construct should to some degree correlate with one another (Crocker & Algina, 2008; L. Ding & Beichner, 2009; Henson, 2001; Kimberlin & Winetrstein, 2008). Clark and Watson (1995) recommend that the average inter-item correlation coefficient be between .15 and .20 for scales measuring broad characteristics and between .40 and .50 for those measuring narrower characteristics. Since the relationships between items are unknown, inter-item correlation coefficients ranging from .15 to .50 were considered acceptable in the current study.

Research Aim 3

The goal of research question 3 was to examine the item characteristics and potential bias of the items between males and females. Item analysis is a general term used to define the investigation of statistical properties of examinees' responses to test items (Crocker & Algina, 2008). While many times used during the instrument development phase, item analysis can provide useful insight about item characteristics to

better understand the quality of the test. More specifically, Item Response Theory (IRT) uses a collection of mathematical equations to analyze item-level data which provides information about the differences among individuals on a given construct or latent variable (De Ayala, 2009; Edelen & Reeve, 2007; Hays et al., 2000; Stone & Zhang, 2003). In order to do so, IRT assumes that the underlying latent trait (e.g., mathematical knowledge) is considered to be continuous in nature and can be represented by assigning numerical values to observed variables.

In the context of this study, item analysis included analyzing item parameters such as difficulties (i.e., the percentage of respondents endorsing a positive response for dichotomously scored items) and item discrimination indices through the use of the Two-Parameter Logistic (2PL) model. In essence, the 2PL model is the ordinary logistic regression of the observed dichotomous responses on the unobservable person location and item characterizations (De Ayala, 2009). This analysis was conducted within the IRTPRO 4.2 for Windows computer program, which makes use of the marginal maximum likelihood estimation method to examine the two parameters described above (Bock & Aitkin, 1981; Cai, Thissen, & duToit, 2011).

Item analysis in this study also included an examination of Differential Item Functioning (DIF). The purpose of DIF is to determine whether or not a particular item is biased. In order examine DIF, respondents are split into groups, each of which are equal on the latent trait (e.g., males versus females). If each group has a different probability of endorsing the item, then that item is exhibiting DIF (Crocker & Algina, 2008; De Ayala, 2009; Hays et al., 2000).

Measure

The mathematics placement test was developed by mathematics faculty members in 1985. The original and continuing purpose of the mathematics placement test is to determine a student's incoming mathematical knowledge for appropriate initial course placement commensurate with ability level. Thus, generally speaking, the placement test assesses mathematical knowledge needed prior to entering into a Calculus sequence. More specifically, the developers of the exam created a two-part test measuring three content areas of mathematics, namely Algebra 1, PreCalculus, and Geometry, as previously determined through an Exploratory Factor Analysis (Manuscript 2).

Part I of the assessment mainly measures student's knowledge of Algebra 1 content such as simplifying expressions, functions, and exponents. Students are given 45 minutes to complete 50 short-answer items, without a calculator. Assessing higher-level abilities such as the ability to solve numerical problems and/or to manipulate mathematical symbols and equations necessitates a short-answer question format (Nitko & Brookhart, 2011). While the short-answer format allows students to show their work, the legibility of students' responses can at times complicate the scoring process.
All responses are graded by the mathematics faculty members using an answer key for dichotomous scoring (i.e., "Correct" or "Incorrect"). If a grader is unsure of a student's written response, other graders are consulted. In the event that a student's response cannot be determined, it is marked as an incorrect response. The possible range of scores on Part I is from 0 to 50. After the allotted time has expired for Part I, exam proctors collect any remaining exams and distribute Part II.

The main focus of Part II of the assessment is to measure students' knowledge of both PreCalculus and Geometry content. For this portion, students have 85 minutes to complete 57 multiple-choice items, again without a calculator. The multiple-choice format used on this portion of the test provides students with the correct answer, three distractor answers, and a fifth response option of "I don't know." Although not explicitly written on the test instructions, mathematics faculty members and exam proctors emphasize the use of the "I don't know" option. By purposefully mentioning this, it is believed that students will not guess, but rather consider using the "I don't know" response option so that they do not accidentally place into a higher course than academically appropriate. A similar argument was made by Prieto and Delgado (1999) who noted that educational standards should not be influenced by desired psychometric properties of a test. Said another way, if students are unsure of an answer, it seems more appropriate for them to omit the item rather than encouraging them to guess. After the exam is complete, the multiple-choice items are scanned into a grading software program using a scantron reader where all items are scored dichotomously (i.e., "Correct" or "Incorrect"), even if the student selected the "I don't know" option. The possible range of scores is from 0 to 57 on Part II.

Participants and Procedure

Existing data from four cohorts of students were used in this study. These cohorts included students entering the high school their sophomore year, beginning in the 2014/2015 academic year and ending in the most recent 2017/2018 academic year for which data were available.

Equivalence across the four cohorts was examined for five demographic variables using Chi-Square (χ^2) Tests of Association and One-Way Analyses of Variance (ANOVAs). Chi-Square Tests of Association were conducted across the four cohorts for the variables of sex and race/ethnicity. There were no significant differences in the proportions between cohort year and either sex or race/ethnicity. For the three remaining variables of socioeconomic status (i.e., median family income), incoming SAT Math (SAT_M) subscores, and incoming SAT Evidence Based Reading and Writing (SAT_ERW) subscores, ANOVAs were used. Again, there were no significant differences between cohort years for each of the three variables. Therefore, all four cohorts were approximately statistically equivalent and were combined into one sample for further analysis.

Both De Ayala (2009) and Ding and Beichner (2009) mention that when calibrating test items of high-stakes assessments, reasonably accurate results are obtained when instruments contain 20 or more items and a sample size of at least 500 participants. With regards to test construction, Nunnally and Bernstein (1978) recommend five times as many subjects as items or at least 200 to 300 subjects, whichever is larger. In the current study, there are a total of 107 items and approximately 300 students in each of the four cohorts. Thus the approximate total population of 1,200 students is greater than the recommendations by De Ayala (2009), L. Ding and Beichner (2009), and Nunnally and Bernstein (1978).

As the multiple-choice section had a fifth response option of "I don't know," the data were coded in such a way as to distinguish between incorrect answers and missing

data. More specifically, the coding format was as follows: "1" for a correct response, "0" for an incorrect response, "DK" for selecting the "I don't know" option on the multiple-choice section, and "M" for a missing response (i.e., an item that was left blank). The response frequencies for each item are displayed in Table 5 in the results section below. Prior to analysis, all responses of "I don't know" were recoded as an incorrect response "0" to align with the grading procedures implemented by the mathematics faculty members.

Data Analysis

The Two-Parameter Logistic (2PL) model suggests that the probability of a correct response is both a function of the distance between the person and the item and the ability of the item to differentiate among individuals with varying levels of the latent trait (De Ayala, 2009; Edelen & Reeve, 2007; Hays et al., 2000).

In order to use the 2PL model, three assumptions must be tenable. First, the data for the 2PL model must be dichotomous. In the current study, the individual responses of the mathematics placement test were dichotomously scored (i.e., "Correct" or "Incorrect"), satisfying the first assumption of the 2PL model. Secondly, the 2PL model assumes unidimensionality. The term unidimensionality implies that the observations obtained from the item responses are a function of only one continuous latent variable (e.g., Crocker & Algina, 2008; De Ayala, 2009; L. Ding & Beichner, 2009; Edelen & Reeve, 2007; Hays et al., 2000; Kirisci, Tarter, & Hsu, 1994). That is, unidimensionality of the mathematics placement test suggests that the scores obtained from the assessment are a direct representation of only students' mathematical knowledge. If the test is

multidimensional, this may indicate that there are factors representing other content domain areas or that both students' mathematical knowledge and reading literacy are being measured. Prior to conducting item analysis, factor analytic procedures were used on the mathematics placement test data. Thus, this assumption was tested, and based on the final factor solution of three factors, each dimension was assessed separately to satisfy the unidimensionality assumption.

The final assumption of the 2PL model is local independence. Local independence is defined as the absence of a relationship between the participant's responses from one item to another, while taking into account the participant's level of the latent trait (Crocker & Algina, 2008; De Ayala, 2009; Edelen & Reeve, 2007; Hays et al., 2000; Kirisci et al., 1994). In other words, the success or failure when answering an item should not be dependent upon the response to another item (Bond & Fox, 2007). This assumption can be violated on both teacher-made and high-stakes assessments. On a mathematics test, a teacher may divide a longer question into multiple parts (e.g., the answer to item 3c is dependent upon the answer calculated in 3a). Likewise, high-stakes assessments often violate this assumption when they ask various questions about a particular reading passage. Again, local independence was upheld in the current study because the mathematics placement test consists of 107 mutually exclusive items.

Model specification. As previously mentioned, the purpose of IRT, and more specifically the 2PL model, is to examine the item-level characteristics to provide additional information regarding the quality of an instrument. Among these

characteristics are item difficulties and item discrimination indices, and an item bias investigation, each of which are discussed below.

Item difficulty is defined as the proportion of examinees who correctly answered the item (Crocker & Algina, 2008). When item responses have been dichotomously scored (i.e., "Correct" or "Incorrect"), then the item difficulty value is the same as the mean item score. Generally denoted as $p_i = \frac{R_i}{T_i}$, where R_i is the number of correct responses for item i and T_i is the total number of responses for item i, the values of the proportion p_i can range from 0 to 1 for each item i (Crocker & Algina, 2008; Quaigrain & Arhin, 2017). Previous research suggests that item difficulty values ranging from .20 to .90 are considered acceptable, with the maximum information being obtained when p_i = .50 (Crocker & Algina, 2008; L. Ding & Beichner, 2009; Quaigrain & Arhin, 2017). Additionally, Quaigrain and Arhin (2017) suggest that difficulty indices less than .20 (i.e., the items are too difficult) or greater than .90 (i.e., the items are too easy) be examined further for item revision or deletion. However, when considering an item for revision or deletion, additional factors should be reviewed in addition to item difficulty.

A second consideration in the 2PL model is the ability of an item to discriminate among individuals with varying levels of the latent trait. That is, the item discrimination index, denoted by D, measures the ability of an item to distinguish between high-achieving and low-achieving individuals for the latent trait of interest (i.e., mathematical knowledge in the current study) (Adedoyin & Mokobi, 2013; Crocker & Algina, 2008; De Ayala, 2009; L. Ding & Beichner, 2009; Ferketich, 1991). Furthermore, the value of

the item discrimination index directly corresponds to the slope of the Item Characteristic Curve (ICC).

An ICC graphically displays the relationship between the probability of answering an item correctly and the underlying latent trait (Crocker & Algina, 2008; De Ayala, 2009; Hays et al., 2000). Moreover, the differences in the item difficulties discussed above are evidenced by the horizontal movement of the ICCs. Items with a higher probability of being endorsed (i.e., easier items such as Item 1 in Figure 5 below) are located further left on the scale of the latent trait whereas items with a lower probability of being endorsed (i.e., harder items such as Item 5 in Figure 5 below) are located further right on the scale of the latent trait.

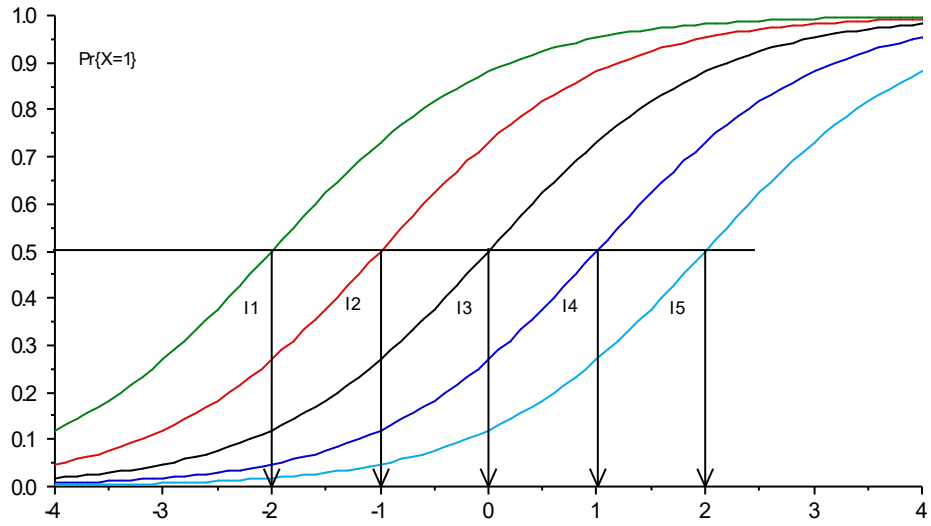

Figure 5. Example of an Item Characteristic Curve. This figure represents an item characteristic curve of five dichotomous items, each with a different level of difficulty (Bradley, 2018).

Generally speaking, the ICC has an S-shaped relationship (i.e., Sigmoid function) indicating that as the respondent's latent trait level increases, so does the probability of answering correctly. From Figure 5 above, the S-shaped function has a steeper slope near the middle of the curve implying that a small change in the latent trait level corresponds to a large change in the chance of endorsing the item (Crocker & Algina, 2008; De Ayala, 2009; Edelen & Reeve, 2007; Hays et al., 2000). This larger slope, or a higher discrimination index value (*D*), provides evidence of item sensitivity, and can detect differences among respondents with varying latent trait levels. Psychometric research provides guidelines for values that are considered "high" or "strong" discrimination indices. The current study used guidelines developed by De Ayala (2009) where the item is determined to be functioning satisfactorily if $.8 \leq D \leq 2.5$.

Other considerations include the direction of the discrimination index. If the discrimination index is negative, the item is performing in a counterintuitive manner (Crocker & Algina, 2008; De Ayala, 2009). In other words, individuals with higher levels of the latent trait are less likely to endorse an easier item compared to individuals with lower levels of the latent trait. In this case, the item with a negative discrimination index should be examined further for possible sentence structure, phrasing of words, and/or a miscoded answer key.

Model fit. Difficulty and discrimination indices can provide useful information at the item level; however, both the individual item fit and the overall model-data fit should be examined. By assessing these fit statistics, the researcher can explore whether or not an individual is responding in such a way that is consistent with the general model.

Considering the placement test in this study, mathematics progresses such that an individual typically should understand Algebra concepts before applying them in a PreCalculus setting. Thus, if an examinee responded correctly to the PreCalculus items towards the end of the exam, then it is expected that he/she responded correctly to the previous Algebra items. If this is not the case, then this examinee's responses do not follow the expected model. A closer look at the examinee's responses may indicate a minor error on the previous Algebra item or possibly a case of academic dishonesty.

In order to assess the item fit and the model-data fit obtained in the 2PL model, this study examined the item-level diagnostic statistics (i.e., $S - \chi^2$) developed by Orlando and Thissen (2000), the M_2 fit statistic developed by Maydeu-Olivares and Joe (2005), and the Root-Mean-Square-Error of Approximation (RMSEA) by Steiger and Lind (1980), each of which are described briefly below. At times, the G^2 statistic, also known as the Likelihood Ratio Statistic, is calculated to examine the model fit. According to Maydeu-Olivares (2013), the G^2 statistic is used when the expected frequencies are greater than five. However, as the number of possible response patterns increases, the expected frequencies decrease and therefore the G^2 statistics is often times not computed due to the sparse observed data, as was the case in the current study.

Again, the Goodness-of-Fit information provide an estimate of how close an individual's predicted response or the model is to the actual observed response or the data (Crocker & Algina, 2008; De Ayala, 2009; Maydeu-Olivares, 2013). That is, the hypothesis is tested that the fitted model is the same as the data-generating model (Maydeu-Olivares, 2013). Thus, if the researcher fails to reject the null hypothesis, then

there is more confidence in the interpretations and inferences drawn from the fitted model.

As mentioned, the first fit statistic examined is the item-level diagnostic statistic S – χ^2 which was developed by Orlando and Thissen (2000). This statistic represents the fit of each individual item to the overall model. When examining these values, an acceptable model-data fit includes no statistically significant differences between the observed and modeled item frequencies.

Similarly, the M_2 fit statistic was used as a measure of overall model-data fit. Developed by Maydeu-Olivares and Joe (2005), the M_n statistic is asymptotically equal to χ^2. This implies that the M_2 fit statistic can be interpreted like χ^2, without the influence of sample size. As previously noted, the χ^2 test null hypothesis states that there are no significant differences between the observed and expected values (Dimitrov, 2013). If the null hypothesis is rejected, then the observed values are significantly different than the expected values, indicating that the model does not represent the data. Thus, for the two goodness-of-fit statistics described above, if the model represents the data, then a larger (i.e., non-significant) *p*-value is desired.

For these analyses and others, an experiment-wise alpha level of .05 was used. In an article by Labovitz (1968), eleven criteria were provided to assist researchers in selecting an appropriate level of significance, some of which include: concerns of practical consequences, conventional levels of the field of research, sample size, and degree of research design control. Furthermore, while the Mathematics Placement Test is a higher-stakes assessment, the exam was developed by faculty members without

knowledge and training in assessment design and advanced quantitative techniques. In considering these criteria in the current study, the conventional .05 level of significance within the field of education was used.

Lastly, the RMSEA fit statistic measures the extent of differences between the observed and expected for each degree of freedom within the model (Browne & Cudeck, 1992; Steiger, 2016). According to previous literature, RMSEA values less than .05 indicate good model fit, and values between .05 and .08 indicate an acceptable model fit (Browne & Cudeck, 1992; Maydeu-Olivares, 2013; Maydeu-Olivares & Joe, 2014; Steiger, 2016). If the RMSEA statistic is greater than or equal to .1, this suggests an unacceptable level of model fit. In this case, it is suggested that alternative models that better represent the data be considered.

Differential item functioning. To identify which items, if any, exhibit DIF, the TSW Likelihood Ratio Test developed by Thissen, Steinberg, and Wainer (1988) was used. The null hypothesis for this test states that there are no group differences in the item parameter estimates (De Ayala, 2009). This calculation follows the $\chi 2$ distribution and is represented by $TSW - \Delta G^2 = G_2^2 - G_1^2$ where G_1^2 and G_2^2 are likelihood ratios. Thus, a significant $TSW - \Delta G^2$ indicates the presence of DIF for that particular item. Similar to before, the significance level was .05.

For the purposes of this study, group comparisons by sex (i.e., male versus female) were conducted. As it was mentioned in the literature review chapter, there is little to no difference in student coursework and performance at the 8th grade level for males and females (Catsambis, 1994). However, males tend to take more advanced

mathematics courses in high school and show a higher achievement in mathematics by age 17 (Catsambis, 1994; Educational Testing Service, 1989; Pedro et al., 1981). It is hypothesized that this lower performance on mathematics exams may cause females to shy away from highly quantitative courses and/or fields of study. In a more recent study by Beede et al. (2011), it was shown that women hold less than a quarter of the jobs in STEM fields nationally. The concerns of women being underrepresented in the STEM fields calls for research to examine why these sex differences in test performance exist so that intervention efforts can be made to change the current trends. Moreover, since the high school of the current study is focused on equal representation of sex (i.e., admittance of approximately fifty percent males and females each year), it is imperative that the mathematics placement test be examined for possible biases and to determine whether or not the placement decisions are equally valid for males and females.

Research Aim 4

The goal of Research Question 4 was to provide evidence of Criterion-Related Validity and to investigate the ability of the test scores to predict future performance in a mathematics course.

Criterion-Related Validity draws an inference from an individual's current exam score to performance on some external criterion of practical importance (Crocker & Algina, 2008; Hambleton, Swaminathan, Algina, & Coulson, 1978). This type of validity can be evidenced either concurrently or predictively. Procedures for concurrent validation are used when the data collected for both the test and the criterion occur at or about the same point in time (Crocker & Algina, 2008; Wiersma & Jurs, 2009). On the

other hand, procedures for predictive validity require a gap in time between when the test was given and when the criterion data are collected (Crocker & Algina, 2008). Additionally, the purpose of predictive validity is to determine whether or not test scores have the ability to predict specified future performance. Thus, the current study sought to evidence Criterion-Related Validity (i.e., Predictive Validity) for the mathematics placement test using Multiple Regression.

More specifically, Multiple Regression was used to investigate the relationship between students' mathematical knowledge, as measured by the mathematics placement test, and students' subsequent performance, as measured by their grade (i.e., a percentage score between zero and 100) in their first semester mathematics course.

Measure

The mathematics placement test was developed by mathematics faculty members in 1985. The original and continuing purpose of the mathematics placement test is to determine a student's incoming mathematical knowledge for appropriate initial course placement commensurate with ability level. Thus, generally speaking, the placement test assesses mathematical knowledge needed prior to entering into a Calculus sequence. More specifically, the developers of the exam created a two-part test measuring three content areas of mathematics, namely Algebra 1, PreCalculus, and Geometry, as previously determined through an Exploratory Factor Analysis (Manuscript 2).

In Manuscript 3, an item analysis was conducted to examine the item parameters (i.e., item difficulties and item discrimination indices) and differential item functioning within each factor. As a result of the study, some items were deleted from the exam. The

Algebra 1 factor had a KR-20 reliability estimate of .895 for 45 items and measured student's knowledge of content such as simplifying expressions, functions, and exponents. The Geometry factor had the lowest reliability estimate (KR-20 = .736) and the fewest number of items (n = 14). These items assessed concepts such as right triangle trigonometry, properties of congruent angles and triangles, and characteristics of a circle. Finally, the PreCalculus factor had a KR-20 reliability estimate of .95 for 35 items and measured student's knowledge of content such as evaluating and graphing quadratic and exponential functions, finding the roots of functions, laws of sines and cosines, and combinatorics. Students' performance on the exam is noted by a raw subscore for each factor (i.e., Algebra 1, Geometry, and PreCalculus) and a total exam score.

Participants and Procedures

Existing data from four cohorts of students were used to examine Research Question 4 in this study. These cohorts consisted of students entering the high school their sophomore year, beginning in the 2014/2015 academic year and ending in the most recent 2017/2018 academic year, for which data was available.

Additionally, group equivalence across the four cohorts was examined and reported for the population information listed above (e.g., gender and race/ethnicity) using Chi-Square Tests of Association. Furthermore, the four cohort means of students' median family incomes (SES), incoming SAT Mathematics scores, and the SAT Evidence-Based Reading and Writing scores were examined for significant differences using the parametric One-Way Analyses of Variance. No significant differences were identified for the five demographic variables and the four cohorts were combined into

one sample for subsequent data analysis. However, due to incomplete and inaccessible data, the final analysis included two of the four cohorts for which the most complete data were available.

Data Analysis

As part of the General Linear Model family of statistical techniques, Multiple Regression is used to explain or predict a criterion (i.e., dependent) variable with more than one predictor (i.e., independent) variable (e.g., Ebel, 1965; Hair Jr et al., 1995; Osborne, 2000; Petrocelli, 2003; Rubio, Berg-Weger, Tebb, Lee, & Rauch, 2003; Stevens, 2012; Wampold & Freund, 1987). There are many types of regression analyses (i.e., Linear, Logistic, Polynomial), which is dependent upon the measurement level of the outcome variable. In the current study, the dependent variables are continuous (i.e., interval level), so a Multiple Linear Regression was used. Although it can be argued that mathematical knowledge may follow a different type of curve, a linear regression model was selected due to the limited time lapse between the start of testing and the completion of their initial mathematics course (i.e., approximately six to eight months).

Furthermore, regression analyses differ in the manner and order in which the independent variables are entered into the model (e.g., simultaneously, stepwise, hierarchically). Hierarchical entry in Multiple Regression allows the researcher to select the order of the entered predictor variables based on previous research and/or theory. When Hierarchical entry is used, the focus is on the change in predictability that is associated with the variables entered later in the analysis, above and beyond the contribution of the previously entered variables (Petrocelli, 2003). Thus, Hierarchical

Multiple Regression was used in the current study to allow the researcher to approximate the reality of placement practices in the high school under study.

Outlier detection. The Hierarchical Multiple Regression was conducted using SPSS. Before running the regression analyses, the data was examined for potential influential data points, leverage points, and/or outliers. The presence of influential data points can significantly affect the overall analysis. An influential data point is one where if deleted, it would produce a substantial change in the value of at least one regression coefficient (Stevens, 2012). To detect influential data points, Cook's distance (Cook, 1977) and DFBETAS (Hahs-Vaughn, 2016; Stevens, 2012) were used. Cook's distance (Cook, 1977) measures the amount of change in the regression coefficients that would occur if a particular case was omitted. Typically, if Cook's D > 1, it is determined that there is an influential data point. While Cook's D is a composite measure of influence, the DFBETAS indicate which specific coefficients are being most influential by providing information on the change in the predicted value when a specific case is deleted from the model (Hahs-Vaughn, 2016; Stevens, 2012). Thus, when any DFBETA value is outside the range of [-2, 2], this indicates a sizeable change and needs to be examined further.

Next, the predictor variables were investigated for possible outliers using leverage values and Mahalanobis distances. Leverage values are used to quickly identify participants that differ from the rest of the sample on a particular set of predictor variables (Stevens, 2012). The current study used the calculation of $\frac{3p}{n}$, where p is the number of predictors plus 1 and n is the sample size, suggested by Stevens (2012) and

adapted from Hoaglin and Welsch (1978). In this case, if the leverage value > $\frac{3p}{n}$, then this data point was examined further.

Additionally, Mahalanobis distances were used to measure how far each case was from the mean of the independent variable for the remaining cases (Hahs-Vaughn, 2016; Stevens, 2012). To determine whether or not a large enough difference existed, which would indicate a possible outlier, the χ^2 distribution table was used to find the critical value for 11 predictor variables with $\alpha = .001$. If the Mahalanobis distance exceeded the critical value, the case was further investigated.

To find outliers on the criterion variable (y), this study examined the standardized residuals (r_i). Standardized residuals allow the researcher to identify subjects whose predicted score is different from the actual criterion score (Stevens, 2012). Generally speaking, standardized residuals follow a normal distribution with approximately 95% of the standardized residual values falling within two standard deviations of the mean (Stevens, 2012). Thus, if $r_i > |2|$, then that data point was carefully examined (Hair Jr et al., 1995; Stevens, 2012).

Each of the above situations (i.e., influential data points, leverage points, and outliers) were considered in the current study so that the appropriate corrective actions could be made, if needed.

Assumptions. After detecting influential data points, leverage points, and/or outliers, the statistical assumptions of regression must be examined and addressed. These assumptions include: Independence of Errors (i.e., Residuals), Linearity, Normality, and Homoscedasticity (Hahs-Vaughn, 2016; Hair Jr et al., 1995; Stevens, 2012). Although

sometimes not described as an explicit assumption, data used in multiple regression analyses should also be examined for multicollinearity.

Multicollinearity exists when there is a strong correlation between some or all of the independent variables (Hair Jr et al., 1995; Stevens, 2012; Wampold & Freund, 1987). If present, multicollinearity reduces the unique explained variance of each predictor variable while increasing the shared prediction, complicating the interpretation of a predictor variable (Hair Jr et al., 1995; Stevens, 2012). To test multicollinearity, the tolerance, variance inflation factors (VIF), and collinearity diagnostics were examined.

Tolerance is measured as 1 minus the proportion of variance explained in the variable of interest by the other predictor variables (Hair Jr et al., 1995). Thus, a lower tolerance value (i.e., less than .10) suggests that the variable of interest is accounted for by the other variables, suggesting possible multicollinearity problems (Hahs-Vaughn, 2016). By taking the reciprocal of tolerance, the VIF is produced and values greater than 10 are indicative of threats to multicollinearity (Hair Jr et al., 1995).

Lastly, the eigenvalues of the collinearity diagnostics were examined. When multiple eigenvalues are close to zero, this indicates that some independent variables have strong intercorrelations and may present concerns of multicollinearity (Hahs-Vaughn, 2016). In this case, the condition index can be calculated using the square root of the ratio between the largest eigenvalue to each preceding eigenvalue, to ensure that no values exceed 10 (Hahs-Vaughn, 2016). If multicollinearity is suspected in any of the above situations, it is recommended that either one or more of the highly correlated variables be eliminated from the model or consolidated into a single measure.

Revisiting the statistical assumptions of multiple regression, the first assumption regarding Independence of Errors (i.e., residuals) assumes that each participant's responses are not dependent upon the response of another individual (Stevens, 2012). If violated, it is possible to identify variables as statistically significant, when in fact they are not (Keith, 2014). In the current study, each student completed their placement exam under the supervision of an exam proctor, implying that the assumption of independence is tenable. Furthermore, the assumption of independence of errors was examined by plotting the studentized residuals against the unstandardized predicted values.

The second assumption of Linearity describes the degree to which a change in the criterion variable associated with the predictor variable is constant across the range of values for the predictor variable (Hair Jr et al., 1995; Keith, 2014). Using partial regression plots, each predictor variable was examined with the criterion variable for the presence of a linear relationship.

The next assumption, Normality, requires that each continuous variable (i.e., independent and dependent) follow a normal distribution of data (Hair Jr et al., 1995; Stevens, 2012). Normality was checked by creating and examining both a histogram of unstandardized residual values in relation to the normal distribution curve and normal probability plots, generally referred to as Q-Q Plots (Hair Jr et al., 1995; Keith, 2014). The skewness and kurtosis of the unstandardized residuals was also examined.

The final assumption, Homoscedasticity suggests the presence of equal error variances (Hair Jr et al., 1995; Keith, 2014; Stevens, 2012). Similar to previous assumptions, violation of homoscedasticity can affect the standard errors, which in turn

will impact the statistical significance of variables. To test for homoscedasticity, residual plots of the predictor variables against the criterion variable were used to identify whether or not a relatively random display of points was present.

One additional consideration in this multiple regression analysis was the sample size. In the current study, an a priori power analysis was conducted in G*Power 3.1.9.4 for the "Linear Multiple Regression: Fixed Model, R^2 Deviation from Zero" (Faul, Erdfelder, Lang, & Buchner, 2007). For the two regressions involving students' total score on the mathematics placement exam, the software yielded a minimum total sample size of 114 to detect a medium effect given a significant level of .05, power of .80, and nine predictor variables (Cohen, 1988). Likewise, for the two regressions involving students' Algebra 1, Geometry, and PreCalculus subscores, the software tool yielded a minimum total sample size of 123 to detect a medium effect given a significance level of .05, power of .80, and eleven predictor variables.

Correlations. Prior to conducting the multiple regression analysis, correlations were investigated to look at the relationship between the independent and dependent variables. Phi correlations were computed for the relationship between the variables of gender and race/ethnicity, as both are measured on a nominal (i.e., dichotomous) scale. For the case where a nominal variable was correlated with a continuous (i.e., interval level) variable, the Point Biserial correlation was calculated. Finally, the Pearson correlation was calculated to examine the relationship between two continuous (i.e., interval level) variables. The correlation matrix summarizing the information above was reported and included indicators for significant correlational values.

Variables. As stated previously, Hierarchical Multiple Regression was used to explore the relationship between students' mathematical knowledge and their subsequent performance in their first semester mathematics course. In any multivariate analysis, the careful selection of variables is important for statistical conclusion validity. When selecting variables for inclusion, the final decision should be based on either theoretical or conceptual grounds (Hair Jr et al., 1995). The variables considered in this study are provided in Table 1 below.

Table 1

Hierarchical Multiple Linear Regression Model Predictors - Level of Measurement and Coding

Variable Name	Level of Measurement	Code	
(1) Demographic Covariates			
Sex	Nominal (Dichotomous)		
Male		0	
Female		1	
Race	Nominal	Race 1 (r_1)	Race 2 (r_2)
Asian		1	0
White		0	1
Other		0	0
Socioeconomic Status	Interval (Continuous)	-	
Median Family Income			
(2) Incoming Performance Covariates			
SAT Math Score	Interval (Continuous)	-	
SAT Critical Reading Score	Interval (Continuous)	-	
Algebra 1 GPA	Nominal (Dichotomous)		
3.0 or below		0	
4.0		1	
Geometry GPA	Nominal (Dichotomous)		
3.0 or below		0	
4.0		1	
Took an Algebra 2 Course	Nominal (Dichotomous)		
No		0	
Yes		1	
(3) Main Predictor Variables			
Mathematics Placement Test	Interval (Continuous)	-	
Total Score			
Algebra 1 Subscore			
Geometry Subscore			
PreCalculus Subscore			
(4) Criterion Variable			
Grade in 1st Semester Math Course	Interval (Continuous)	-	
Lower Level Math Course			
Upper Level Math Course			

Over the past two decades, numerous articles have detailed the uses, consequences, and challenges of placement exams (e.g., Denny et al., 2012; Farley, 2007; Foley-Peres & Poirier, 2008; Haeck, Yeld, Conradie, Robertson, & Shall, 1997; Rueda & Sokolowski, 2004; Schmitz & delMas, 1991). However, the vast majority of these studies were within the context of a community college or university. Thus, the predictor variables chosen for inclusion in the current study were from similar studies containing varying contexts.

For each of the four regressions conducted in the current study, the first block of the Hierarchical Multiple Regression included student demographic information such as sex, race/ethnicity, and socioeconomic status (SES). A variety of studies have been conducted examining demographic variables and their impact on educational outcomes, specifically math achievement. For example, in a study by Roth et al. (2000), racial differences in mathematics achievement did not exist after controlling for previous coursework in mathematics. Another study mentioned that regardless of racial group, SES was unrelated to gender differences in mathematics achievement or attitudes (Catsambis, 1994). Moreover, Pugh and Lowther (2004) found that regardless of students' race, SES, or type of high school, the greatest indicator of college achievement was the mathematics course(s) taken.

Conversely, additional research has demonstrated SES, especially income, to be an important predictor in mathematics achievement and career decisions, especially for females (Gonzalez & Kuenzi, 2012; Oakes, 1990). Moreover, research has shown that

Black and Hispanic students are less than half as likely to be in gifted education programs compared to White students (Callahan, 2005). The same study also concluded that nine percent of students enrolled in gifted and talented programs were categorized in the bottom quartile of family income (Callahan, 2005). Other studies have concluded that both SES and race/ethnicity strongly correlate with academic performance and account for a significant amount of variance in students' test scores (Sirin, 2005; White et al., 2016). Although the nature of the impact of race/ethnicity and SES on educational achievement is ongoing, these variables have not been considered in the context of a gifted residential high school focused on STEM.

The second block in the regression analyses contained incoming academic information including students' SAT mathematics subscore, SAT Evidence-Based Reading and Writing subscore, students' grades in previous coursework (i.e., GPA of Algebra 1 and Geometry) and whether or not the student had reached an Algebra 2 level course. In a study by Sheel, Vrooman, Renner, and Dawsey (2001), high school GPA, SAT mathematics score, and the student's final grade received in high school Algebra 2 were the most influential predictors of students' college mathematics placement test scores. Similarly, Latterell and Regal (2003) found that other predictors such as high school courses and the grades received in those courses were often stronger predictors of college course success than an incoming placement test score. These variables are similar to others in previous studies, but the context was at the post-secondary level rather than at a high school (Latterell & Regal, 2003; Pugh & Lowther, 2004; Sheel et al., 2001).

The third and final block of the analysis included the high school mathematics placement test scores, one using the total score and another using subscale score of Algebra, Geometry, and PreCalculus. The placement test was positioned last in the Hierarchical Multiple Regression as the amount of variance the placement test explains, over and above the variables in the previous blocks, was central to addressing the fourth research question in this study.

Finally, the criterion (i.e., outcome) variables in this study were students' percentage grades received in their first semester mathematics course, which were divided into lower and upper level courses. Based on the placement exam score, students enter into one of four mathematics courses – Mathematical Investigations I, II, III, or IV. Thus, Mathematical Investigations I and II were categorized as lower level courses with Mathematical Investigations III and IV being categorized as upper level courses. While some students begin the math sequence in either Geometry or BC Calculus I, these decisions are not determined through the use of the placement exam, and thus were not included in the study sample.

Summary

This study aims to identify the psychometric properties of a mathematics placement test at a residential high school focused on STEM for gifted students. More specifically, this study seeks to provide evidence of reliability and validity, in addition to examining the characteristics of the item parameters (i.e., item difficulty, and item discrimination) and item bias with regards to sex. In light of these objectives, this chapter reviewed the research aims of this study and the related methodologies to answer

each of the four research questions. The following chapters (i.e., Four, Five, Six, and Seven) consist of manuscripts for each of the research questions described above. Chapter Eight (i.e., Conclusions) summarizes the four manuscripts and their implications for one another.

CHAPTER IV – MANUSCRIPT 1

CONTENT VALIDITY USING MULTIDIMENSIONAL SCALING AND HIERARCHICAL CLUSTER ANALYSIS: A PRACTICAL APPROACH

Abstract

Educational assessments, when properly constructed, can provide valuable feedback regarding content that has or has not been learned. However, such test results can only be meaningfully interpreted if there is an adequate alignment between the items on the assessment and the local curriculum. For this reason, providing evidence of Content Validity remains an issue of paramount importance throughout the test development process. The current study examined the Content Validity of a mathematics placement test at a Science, Technology, Engineering, and Mathematics (STEM) gifted residential high school. Data were collected from internal and external subject matter experts using a card-sorting technique replicated from a study by D'Agostino et al. (2011) and were analyzed using Multidimensional Scaling and Hierarchical Cluster Analysis. Results demonstrate preliminary evidence of congruence between the two configurations.

Keywords: Content Validity, Multidimensional Scaling, Hierarchical Cluster Analysis, STEM Education

Introduction

Over the past forty years, specialized Science, Technology, Engineering, and Mathematics (STEM) projects and programs have been developed for gifted children. Within these programs, gifted students are exposed to an ambitious college preparatory

curriculum with the expectation of majoring in a STEM field. While students undergo a competitive and challenging application and acceptance process, the effects of these specialized programs remain relatively unknown.

More recently, research has identified a shortage of valid and reliable instruments to measure the impact and outcomes of these specialized STEM programs (Katzenmeyer & Lawrenz, 2006; Scott, 2012). Additionally, in the era of accountability, it is critical that educational institutions at varying levels maintain rigorous and defensible placement practices and methods in order to justify their use and to confront questions of their impact on students' educational outcomes. Frisbie (1988) stated that when the reliability of scores as accurate measures of student achievement are in question, these scores cannot be used to make future educational decisions. Furthermore, one validation study is not sufficient to guarantee the psychometric properties of an assessment throughout its lifetime. Instead, the assessment and policies used, in contexts such as placement testing, need to be continuously reviewed and evaluated to assure that students are being placed into courses commensurate with their ability in order to maximize the chances of success (Linn, 1994; Mattern & Packman, 2009; McFate & Olmsted III, 1999; Norman et al., 2011; Wiggins, 1989). Overall, when properly constructed and evaluated, assessments can provide feedback on what has and has not been learned to both the student and other interested stakeholders.

The purpose of this study was to demonstrate evidence of Content Validity on a mathematics placement test at a Science, Technology, Engineering, and Mathematics (STEM), gifted, residential high school. Previous research on placement exams have

been conducted at the post-secondary level; however, this study extends the research to younger grade levels serving a specific, gifted population. Furthermore, this study sought to replicate an efficient and innovative card-sorting technique by D'Agostino et al. (2011) using the complementary techniques of Multidimensional Scaling (MDS) and Hierarchical Cluster Analysis (HCA) within a new context.

Literature Review

Although prior research has not extensively examined placement testing from middle school to high school, a large literature base exists using college and university student populations. Approximately 90% of post-secondary institutions use placement tests (Latterell & Regal, 2003). The near-universal practice of administering placement tests emerged due to the incomparability of unknown factors such as the content and rigor of courses and the grading scales used at different schools (Kossack, 1942; Linn, 1994; Ngo & Kwon, 2015; Noble et al., 2003). Within the setting of a post-secondary institution, students complete placement tests to determine the appropriate level to begin coursework. In the same way, upon acceptance into the high school in the current study, students must complete a series of placement tests to guide their initial course enrollment.

The overarching purpose of placement tests is to match students with a level of instruction that is appropriate given their previous academic preparations (e.g., Akst & Hirsch, 1991; Frisbie, 1982; Marshall & Allen, 2000; Mattern & Packman, 2009; McFate & Olmsted III, 1999; Noble et al., 2003; Sawyer, 1996). Prior research has shown that course placement decisions can have a significant impact on a student's future academic preparation (McDaniel et al., 2007; Morgan & Michaelides, 2005). For example,

students who begin post-secondary mathematics in a course that is appropriate given their background have an increased chance of succeeding in their first course and subsequent mathematics courses (Mattern & Packman, 2009; Norman et al., 2011; Shaw, 1997). For this reason, more research is needed to thoroughly examine placement tests and procedures to ensure that student success is maximized while the consequences of misplacement are minimized. Although these placement tests are typically considered "high-stakes," the psychometric properties of such tests have received relatively little attention (Callahan, 2005; Grubb & Worthen, 1999; Scott-Clayton, 2012). As a result, more research is needed to investigate and evidence the psychometric properties of placement tests.

Validity is typically defined as the extent to which an instrument measures what it is intended to measure (Wiersma & Jurs, 2009). While this definition is somewhat accurate, it is often times misleading. That is, the instrument itself is not validated, rather the conclusions and interpretations drawn from the scores have validation evidence (Cook & Beckman, 2006; Ebel, 1956; Kimberlin & Winetrstein, 2008; Messick, 1995; Moss, 1992; Schmitz & delMas, 1991). Using these details, validation is defined by Cronbach (1971) as an evidence-collecting process to support the inferences made from the test scores.

Content Validity addresses if the wording/phrasing and meaning measures a set of performance tasks for a construct of interest (Cook & Beckman, 2006; Crocker & Algina, 2008; Ebel, 1956; Grant & Davis, 1997; Haynes et al., 1995; Martone & Sireci, 2009; Sireci, 1998a). Content Validity contains three components related to the domain: (1)

Definition, (2) Representation, and (3) Relevance. The first component, Domain definition, refers to the operational definition of the content domain describing both the content area(s) of interest and the level(s) of cognition required (Sireci, 1998a). This component typically occurs during the design stage before test items have been created or selected.

The second and third components (i.e., Domain representation and Domain relevance) are generally examined after the test's development. Both Domain representation and Domain relevance require the subjective evaluation of subject matter experts (SMEs). For Domain representation, SMEs are asked to judge whether or not the test items adequately represent the content and cognitive specifications (Sireci, 1998a). In a similar way, SMEs appraise the relevance of each test item to the primary content domain when examining Domain relevance. Overall, evidence that a test adequately represents the underlying content domain remains a vital component to test development and construction (Sireci & Geisinger, 1992).

Former Content Validation studies have used a variety of methods to evaluate item similarities and relevance. Two of the most recognized techniques are item-pairing and item-sorting tasks. In studies by Sireci and Geisinger (1992, 1995), researchers asked SMEs to rate the similarity of a given item-pair on a scale from "Highly Similar" (Coded 1) to "Highly Dissimilar" (Coded 10). In a similar way, SMEs were asked to rate the degree of each item's relevance to the content areas listed (Sireci & Geisinger, 1992, 1995). One year later, Deville and Prometric (1996) used a comparable item-pairing task. While the item-pairing technique can provide a more comprehensive examination of

content domain representation, it can quickly become burdensome for SMEs when the number of items become too large. For example, the mathematics placement test in the current study consists of 107 total items. If the item-pairing task was used, SMEs would be asked to rate item-similarities for 5,671 unique item-pairs. Not only is this an unrealistic task for an individual to complete, but it is also detrimental to the recruitment of SMEs. Additionally, prior research has suggested the use of sorting procedures requiring SMEs to sort items into a limited number of categories according to their similarities (Sireci & Geisinger, 1995). The same study also suggested that item-level data be obtained to determine how Factor Analysis or Multidimensional Scaling (MDS) results compare to the dimensions obtained from the SME similarity ratings.

For these reasons, the current study employed a card-sorting task to gather data on the content areas of the exam. Replicated from a study by D'Agostino et al. (2011), MDS and HCA were used to compare the similarity ratings of external SMEs to the similarity ratings of internal SMEs. Generally, when using MDS in Content Validity studies, similarity ratings from SMEs are compared to the original test specifications (D'Agostino et al., 2011; Li & Sireci, 2013; Sireci & Geisinger, 1992, 1995). In the current study, there were no formal test specifications. However, prior research has demonstrated the complementary use of MDS and HCA in the development of content specifications for professional certification exams (Raymond, 1989; Schaefer et al., 1992). Thus, the design of the current study made use of internal SME item-similarity ratings to develop the content specifications, which were then compared to external SME item-similarity ratings to provide evidence of Content Validity.

In educational assessment, evaluating inferences drawn from test scores begins with evaluating the test itself (Sireci, 1998a). Achievement tests, like the mathematics placement exam, should represent the intended domain without the presence of material external to that domain. The current study examined the Content Validity of a mathematics placement test at a STEM gifted residential high school using a card-sorting technique adopted from D'Agostino et al. (2011). Existing research on placement exams has focused on the post-secondary level; however, this study extends the literature base to younger grade levels serving a specific, gifted population.

Methods

The following sections describe the methods used to examine the Content Validity of a mathematics placement test.

Participants

The recruitment and qualifications of SMEs is an important consideration in any Content Validation study. The number of SMEs needed for a content validation study will be driven by the range of representation and experiences desired by the researcher (Grant & Davis, 1997). As described previously, the context of the current study was unique in that it occurred at a gifted residential high school focused on STEM. With its advanced curriculum and residential component, the high school is often times compared to an institution of higher education. However, because the school serves students in grades 10 through 12, it is categorized as a high school. Therefore, to properly assess the Content Validity of this school's mathematics placement test, SMEs at varying levels (i.e., high school, community college, four-year post-secondary institutions) were

recruited from across the state of Illinois. Additionally, the external SMEs were selected for recruitment based on their interests, experiences, and/or contributions to mathematics and STEM education.

Final study participants included nine internal SMEs and eight external SMEs. Of the 17 total participants, seven majored in mathematics education and four majored in mathematics. A summary of the internal and external SME samples for which data were collected is presented in Table 2 below.

Table 2

Subject Matter Expert Demographics

Characteristic	Internal	External
Gender		
Male	5	5
Female	4	3
Education		
Bachelors	0	1
Masters	5	3
Doctorate	4	4
Grade Level Taught		
High School	9	3
Community College	0	2
4-year University	0	3
Average Number of Years Teaching	18.17 (*SD* 11.55)	22.25 (*SD* 10.50)

Measure

Developed in 1985, the continuing purpose of this placement test is to determine a student's incoming mathematical knowledge for appropriate course placement commensurate with ability level. The developers of the exam created a two-part test

measuring mathematical knowledge needed prior to entering into a Calculus sequence. However, like most teacher-made tests, the items were constructed by the mathematics faculty members at the high school without being subjected to formal psychometric evaluation.

Part I of the assessment measures student's knowledge of content such as simplifying expressions, functions, and exponents. Students are given 45 minutes to complete 50 short-answer items, without a calculator, and are encouraged to show their work. The second part of the exam gives students 85 minutes to complete 57 multiple-choice items, without a calculator, related to topics such as functions, graphing, Trigonometry, and Geometry. The multiple-choice items used have the following response options: the correct answer, three distractor answers, and a fifth response option of "I don't know." All responses of the assessment are graded by the mathematics faculty members using an answer key for dichotomous scoring (i.e., "Correct" or "Incorrect"). Thus, the possible range of scores on the mathematics placement test is from 0 to 107.

Procedure

After consenting to participate, the SMEs were mailed a card-sort packet including cards for the 107 items and a response sheet to record their groupings. The cover page of the response sheet asked each individual to report their demographic information such as current employer, grade level(s) taught, highest degree earned, major of the highest degree earned, and total number of years teaching. At the top of the second page, participants were provided the directions for the card-sorting task which instructed

each SME to place the 107 items into meaningful piles or groups based on the similarity of the content of the items. Consistent with the sorting rules described by Trochim (1989), SMEs were advised to: (1) place each item or card into only one pile or group, (2) refrain from creating as many piles or groups as there are items, and (3) create more than one pile. Upon completion of the card-sorting task, SMEs recorded the item numbers in each pile and assigned each group of items a group title or name (Appendix E). All materials were then returned to the Principal Investigator via a prepaid envelope. On average, the task took between 30 to 45 minutes to complete.

Data Analysis

Each SME's coding sheet was transformed into an individual item-similarity rating matrix where the test item numbers were listed for both the rows and the columns. An entry of "0" indicated that the SME did not categorize a specific item-pair together, whereas an entry of "1" indicated that the SME did put the item-pair in the same group (D'Agostino et al., 2011). The diagonal of the square-symmetric matrix contained 1's, representing that an item was always categorized with itself.

After each individual item-similarity matrix was created, two group item-similarity matrices were constructed by adding the individual internal and external item-similarity matrices together, respectively (D'Agostino et al., 2011). Values of the internal group item-similarity matrix range from 0 (no SME chose the item-pair) to 9 (all SMEs placed the two items in the same group). Similarly, values of the external group item-similarity matrix ranged from 0 to 8. Thus, a larger cell value within the matrix represented a greater consensus of SMEs regarding the similarity of the items.

Since similarity and dissimilarity ratings are inverses of one another, researchers have recommended transforming similarity ratings into dissimilarity ratings prior to data analysis (Jaworska & Chupetlovska-Anastasova, 2009; Kruskal & Wish, 1978). For the purpose of the current study, the group item-similarity matrices for both internal and external SMEs were first converted into group item-similarity ratios. Using a scale from 0 to 1, these ratios were then transformed into a group item-dissimilarity matrix using the calculation of $1 - n_{jk}$ where n is the matrix cell value for the item-pair j and k where $j \leq k$.

Using SPSS version 24, each group item-dissimilarity matrix was subjected to multidimensional scaling (MDS) based on the method by Kruskal and Wish (1978). The two fit indices used were Kruskal's Stress Function (Kruskal, 1964) and the Squared Correlation Index also known as Tucker's Coefficient of Congruence (Moroke, 2014). Similar to other goodness-of-fit indices, Kruskal's Stress Function is a calculation of the residual sum of squares (Kruskal, 1964). As such, smaller values indicate a better fit between the data and the MDS solution. For the purposes of this study, the following stress values were used as guidelines: $S = 0$ suggests perfect fit; $0 < S \leq .025$ suggests excellent fit; $.025 < S \leq .05$ suggests good fit; $.05 < S \leq .10$ suggests fair fit; and $S \geq .20$ suggests poor fit (Kruskal, 1964). Secondly, Tucker's Coefficient of Congruence (T) values are interpreted as the proportion of variance explained by the disparities (Hair Jr et al., 1995; Moroke, 2014; Whaley & Longoria, 2009). In other words, T measures how well the MDS model fits the original data, implying that higher values indicate better fit. In the current study, the MDS solution was considered an acceptable fit if $T \geq .60$ (Hair Jr et al., 1995; Whaley & Longoria, 2009). To support interpretation, Euclidean distances

for each item were saved on eight dimensions. The selection of dimensions was primarily based on three considerations: (1) the values of the fit indices, (2) the amount of change in fit indices from *n* to *n – 1* dimensions, and (3) the interpretability of the dimensions (Whaley & Longoria, 2009).

Next, the item scale coordinates for both internal and external SMEs were analyzed using hierarchical cluster analysis (HCA) within SPSS. The goal of HCA is to find the simplest structure possible that still represents homogeneous groupings (Hair Jr et al., 1995). Moreover, by imposing HCA on the MDS solutions, the domain structure of the internal SMEs and external SMEs can be compared and the degree of consensus between the two domain configurations can be determined (D'Agostino et al., 2011; Sireci & Geisinger, 1992). In this study, HCA was conducted using the agglomerative clustering method with Euclidean distances and the average-linkage clustering algorithm (Hair Jr et al., 1995; Johnson, 1967; Sarstedt & Mooi, 2014). Finally, the fit of various cluster solutions were analyzed by exploring the results of several validity indices.

After the final cluster solutions were determined for both the internal and external SME responses, the two configurations were compared using the Rand and adjusted Rand indices. The Rand index (*RI*) computes the overlap between classification schemes, while the adjusted Rand index (*ARI*) controls for overlap by chance due to marginal distributions (Hubert & Arabie, 1985; Rand, 1971). The Rand index was calculated as follows:

$$RI_{ij} = \frac{a+d}{a+b+c+d} \quad [4]$$

where

- a is the number of pairs of items that are placed in the same cluster for both internal and external SMEs;
- b is the number of pairs of items that are placed in the same cluster for the internal SMEs, but not in the same cluster for the external SMEs;
- c is the number of pairs of items that are placed in the same cluster for the external SMEs, but not in the same cluster for the internal SMEs;
- d is the number of pairs of items that are placed in different clusters for both internal and external SMEs (D'Ambrosio, Amodio, Iorio, Pandolfo, & Siciliano, 2020; Rand, 1971; Warrens, 2008).

Using the same definitions as in equation 4, the adjusted Rand index (*ARI*) can be computed as:

$$ARI_{ij} = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)} \quad [5]$$

In equation 5, the *ARI* gives a potential score between -1 and 1, such that a score greater than zero would indicate that the probability of a link being present between the two clusters is greater than random chance (Hoffman, Steinley, & Brusco, 2015). However, in each instance, a higher value closer to 1 indicates a stronger overlap.

Results

The current study used two data analysis techniques to examine the Content Validity of a mathematics placement test at a gifted, STEM residential high school. Results for each data analysis technique used are described below.

Multidimensional Scaling

Upon subjecting each item-dissimilarity matrix to a multidimensional analysis, the stress indices and proportions of variance explained were compared for the configurations of six to nine dimensions. The fit indices for each of the four configurations are in Table 3 below.

Table 3

Fit Indices for Multidimensional Scaling Analysis

Number of Dimensions	Internal SMEs		External SMEs	
	S	T	S	T
6	0.12542	0.99210	0.13417	0.99096
7	0.11681	0.99315	0.11328	0.99356
8	**0.09656**	**0.99533**	**0.09704**	**0.99528**
9	0.08645	0.99626	0.08649	0.99625

Note. S = Kruskal's Stress (Stress-I), T = Tucker's Coefficient of Congruence

Taking into account the interpretability of the dimensions with the above information, the final solution for both Internal and External SMEs was eight dimensions. The coordinates in eight dimensions were saved for each of the final solutions for further analysis using HCA.

Hierarchical Cluster Analysis

To examine the domain structure of the internal and external SME solutions, the final item coordinates for each of the eight dimensional solutions were entered into a HCA. To begin, the number of clusters were allowed to range from a minimum of 1 to a maximum of 106 in each analysis. In order to determine how many clusters to retain, a Scree Plot was created by graphing the number of possible clusters on the x-axis against

the distances at which the clusters were combined on the y-axis. The scree plot was then visually examined for a bend (i.e., departure from parallel to the y-axis to perpendicular) to indicate a possible solution for how many clusters to retain. Similarly, a dendrogram (i.e., tree graph) was analyzed alongside the agglomeration schedule to identify large differences between two subsequent groupings in the analysis. When large distances are present between two cluster groupings, this implies that two non-similar groups are combined, which suggests a possible final solution.

The largest difference of .169 in the external SME analysis occurred between items 98 and 99 suggesting an eight cluster solution. In a similar way, the largest difference in the internal SME analysis was .122 between items 103 and 104, indicating a three cluster solution. Due to the large number of items on the mathematics placement test (107 items), a three-cluster solution was determined to be insufficient. Moreover, one of the goals of using HCA was to compare the two domain structures between internal and external SMEs, implying that each of the final solutions needed to contain the same number of clusters. Next, an eight-cluster solution was examined for the internal SMEs. However, the distance between internal SME items 98 and 99 was small with a difference of .021.

Since a three- and eight-cluster solution were inadequate for both internal and external SMEs, the second largest change in distances was examined. The second largest difference for the internal SME analysis was .094, which occurred between items 100 and 101 suggesting a six-cluster solution. Although the second largest difference did not occur between items 100 and 101 for the external SMEs, there was still a notable change

of .05. Therefore, based on the cluster structure and interpretability, it was determined that a six-cluster solution would be retained for both internal and external SMEs. When possible, the most frequently cited group title was used. Therefore, the final six clusters were: (1) Algebraic Operations, (2) Solving Equations, (3) Graphing Functions, (4) Evaluating Functions, (5) Trigonometry, and (6) Geometry.

Lastly, to quantify the degree of concordance between the internal and external SME configurations, the Rand index (RI_{ij}) and adjusted Rand index (ARI_{ij}) were calculated. These indices are reported on a scale from 0 to 1, with higher values indicating a stronger overlap. Thus, a Rand index of .63 suggests an agreement between the two classifications of approximately 63%. An adjusted Rand index of .13 indicates that there is some congruence between the two domain definitions, providing initial Content Validity evidence.

Discussion

In the process of Content Validation, two readily recognizable techniques for evaluating item similarities and relevance are item-pairing and card-sorting tasks. Item-pairing tasks, while useful for a more comprehensive examination of content domain representation, can be burdensome for the SMEs as the number of test items increase. D'Agostino et al. (2011) proposed a novel approach by combining the methods of Sireci and Geisinger (1992, 1995) and Trochim (1989), which provided an efficient method for exploring domain configurations. This efficiency was further evidenced in the current study as SMEs categorized 107 test items in less than 45 minutes.

The current study further extended these methods by drawing on the research of Raymond (1989) and Schaefer et al. (1992). Through their demonstration of developing content specifications using both MDS and HCA, this study was able to make use of the internal SME ratings to create the content specifications. The resulting models for the internal and external SMEs suggested a virtually unanimous agreement regarding the Trigonometry and Geometry items, but differed in their groupings and the level of detail related to Algebra and other items. The average number of card-sorting groups for the internal SMEs was approximately 16.7, compared to approximately 19.1 for the external SMEs.

While internal and external SMEs often grouped two items similarly, the final cluster solutions differed partly due to the level of detail. For example, one external SME placed items 75, 80, and 93 in one pile and named it "Basic Trig" with items 76 – 79 and 81 – 87 in another pile named "Advanced Trig." Several other SMEs categorized these same items together and provided a similar group name such as "Trigonometry." Another example of the differences in categorization is demonstrated between internal SME #4 and external SME #7. Internal SME #4 labeled one of their larger item groupings as "Exponents and Polynomials." Rather than having one overarching category of polynomials, external SME #7 listed more detailed item groupings such as "Operations with Polynomials," "Factoring Polynomials," and "Polynomial Functions." Due to this discrepancy between the internal and external SMEs, many item pairs were grouped similarly, but ultimately ended up in different clusters.

It is important to note that the two methods used, MDS and HCA, were complementary to one another in this study. The purpose was not to provide alternative ways to view and describe the data, but rather to use HCA as a way to visually represent the MDS configurations. Additionally, the clustering was conducted on the unweighted item coordinates of the MDS solutions, thus assuming that each dimension was considered equally important to the SMEs. Furthermore, by comparing the internal and external SME ratings, these two approaches provided initial evidence of Content Validity by identifying groups of items perceived to be similar by both the internal and external experts.

Implications

Validity is context- and population-specific implying that assessments designed for the general student population can produce biased results without further psychometric scrutiny and documentation (Schmidt & Hunter, 1977). Evidencing the necessary psychometric support for the sample used and the context of the study through rigorous Content Validation procedures is needed to ultimately produce reliable and valid scores resulting in unbiased study results. Data collected from a card-sorting task indicated that the quality and appropriateness of items on the mathematics placement test were perceived similarly by internal and external SMEs. Therefore, faculty members and educational administrators of the high school in the current study can be reassured that the mathematics placement test adequately measures the mathematical domain of interest. Additional research in this area can provide further insight regarding the knowledge and skills measured by the mathematics placement test and how the larger domain of

mathematical knowledge may be further subdivided to provide information that is more specific. Finally, use of the content validation procedures from D'Agostino and colleagues (2011) has implications for researchers in measurement. The application of this technique in a new context, and with a test lacking definitive specifications, can provide researchers with another example and extension to evidence content validity.

Limitations and Future Research

Although the current study supports initial evidence of Content Validity, there were some limitations. Within the final six-cluster solution, the third cluster (i.e., Graphing Functions) had no overlapping items between the internal and external SMEs. Cluster 3 for the internal SMEs included items on sequences and series, combinatorics, and vectors, most of which appeared in Cluster 1 for the external SMEs. Comparatively, Cluster 3 for the external SMEs included items such as linear, exponential, and logarithmic functions and graphs. Upon further examination of the individual SME responses to the card-sorting task, it was determined that both the internal and external SMEs tended to group sequences and series, combinatorics, and vectors into single card piles. Thus, while the two SME groups were in agreement, it is possible that the discrepancy in the average number of card-sorting groups for internal and external SMEs influenced how these items were ultimately clustered. Moreover, when debriefing with the internal SMEs, a few individuals made mention that the current structure of their curriculum directly influenced how they categorized items during the card-sorting task. Future research may consider using both past and present internal SMEs to potentially negate the biasing effects of the current curriculum.

Another limitation of the current study was the small sample size obtained for both the internal and external SMEs. Grant and Davis (1997) stated that the number of SMEs needed for a content validation study is driven by the range of representation and experience desired by the researcher. While a wide range of experience and contribution was sought through the use of email recruitment and subsequent reminders, this study had a response rate of about 65%. Additional research should consider other sampling methods and tools for recruitment to obtain larger sample sizes both internally and externally.

As previously mentioned, Content Validity contains three components related to the domain: (1) Definition, (2) Representation, and (3) Relevance. Moreover, the first component, Domain definition, refers to the operational definition of the content domain describing both the content area(s) of interest and the level(s) of cognition required (Sireci, 1998a). A final limitation of the current study was the absence of an examination regarding the level(s) of cognition required for the various items on the mathematics placement test. Future research may consider extending the current study by asking subject matter experts to rate the level(s) of cognition required for each item using a framework such as Bloom's Taxonomy (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). In doing so, faculty and administrators can examine whether the level(s) of cognition required of students within the mathematics courses is in alignment with the level(s) of cognition being assessed on the mathematics placement test.

Conclusions

Previous research surrounding placement exams and their psychometric properties have been largely conducted at the post-secondary level. However, in an era of accountability, it is recommended that educational institutions be able to defend their placement practices through rigorous examination of the corresponding tests, as these decisions have a significant impact on students' future educational outcomes (Mattern & Packman, 2009; McDaniel et al., 2007; Morgan & Michaelides, 2005; Norman et al., 2011; Shaw, 1997). This study provides a first step in encouraging other schools with a STEM and/or gifted education focus to begin the validation process and extend and improve upon the educational testing practices at other levels of schooling.

Results from the current study supported preliminary evidence of Content Validity for a mathematics placement test at a gifted, residential STEM school using MDS and HCA. Future research should further examine the psychometric properties of this exam including, but not limited to, Construct Validity, Criterion-Related Validity, Reliability, and a more detailed Item Analysis.

CHAPTER V – MANUSCRIPT 2

EXAMINING THE VALIDITY AND RELIABILITY OF A MATHEMATICS PLACEMENT EXAM AT A SCIENCE, TECHNOLOGY, ENGINEERING, AND MATHEMATICS (STEM) GIFTED RESIDENTIAL HIGH SCHOOL

Abstract

Post-secondary institutions administer placement exams due to the incomparability of unknown factors such as the content and rigor of previous courses and the grading scales used at different schools (Kossack, 1942; Linn, 1994; Ngo & Kwon, 2015; Noble et al., 2003). The primary objective of placement testing is to determine a student's incoming knowledge for appropriate course placement commensurate with ability level. Before entering the decision-making process, institutions must provide evidence regarding the psychometric properties of their assessment(s).

The current study examined the Construct Validity and Internal Consistency Reliability of a mathematics placement test at a Science, Technology, Engineering, and Mathematics (STEM) gifted residential high school. Existing data from four cohorts were obtained and analyzed using Exploratory Factor Analysis and the Kuder-Richardson (KR-20) Formula for internal consistency reliability. Results indicated that the mathematics placement test is comprised of three factors, namely PreCalculus, Geometry, and Algebra 1. Strong Internal Consistency Reliabilities suggest that the items in each factor are related to one another and that they are measuring the same construct. Therefore, this study demonstrated evidence of Construct Validity and Internal

Consistency Reliability for the population of interest and can be used in the decision-making process of course placement.

Keywords: Exploratory Factor Analysis, Internal Consistency Reliability, Mathematics Placement Test, STEM Education

Introduction

In educational measurement, constructs such as achievement, interest, and performance are assigned numerical values, through the use of a wide variety of tests and assessments, to infer the abilities and proficiencies of students. The purpose of achievement testing is to measure students' actual knowledge or acquired skills in order to reliably distinguish between students who do and do not have some level of the construct of interest (Slavin, 2007). As one of the primary measures used in educational research, there is an abundance of literature focused on achievement testing.

Beginning at the post-secondary level, numerous articles have been published regarding the use of placement tests for incoming students. Many of these articles mention the continuing decline of academic standards, specifically in the area of mathematics (e.g., Crist et al., 2002; Hoyt & Sorensen, 2001; Medhanie et al., 2012; Ngo & Kwon, 2015; Parker, 2005; Schmitz & delMas, 1991). Unsurprisingly, the lowered academic standards in math are said to be related to students' scoring lower on mathematics placement tests. Due to the lower test scores, more students are being assigned to take remedial coursework, which has sparked a conversation about whether or not students are less prepared for college-level work or if the placement tests used are appropriate for this type of decision (Morgan & Michaelides, 2005).

More specifically, nearly one-third of all students entering community colleges take at least one remedial or developmental course in mathematics (e.g., Bailey, 2009; Hoyt & Sorensen, 2001; Kowski, 2013; Medhanie et al., 2012; Melguizo et al., 2014; Scott-Clayton, 2012). Not only do these remedial courses lower student motivation, but they also add time to student graduation. Furthermore, the additional time students spend taking non-credit courses increases their overall cost to attend and lowers retention rates (Medhanie et al., 2012; Melguizo et al., 2008; Ngo & Kwon, 2015; Scott-Clayton, 2012). Some community colleges have even been accused of placing students into these remedial, non-credit courses as a way to increase revenue (Armstrong, 2000). As a result, post-secondary institutions are now being asked to provide evidence of the effectiveness of their placement procedures and measures to ensure that the negative consequences of misplacement are minimized (Armstrong, 2000; Morgan & Michaelides, 2005; Smith & Fey, 2000). After all, accurately placing students is a necessary, but not sufficient, condition for a placement system as a whole to be effective (Sawyer, 1996).

In the era of accountability, placement practices and methods that are rigorous and defensible are critical for educational institutions at varying levels to justify their use and to confront questions of their impact on students' educational outcomes. Frisbie (1988) stated that when the reliability of scores as accurate measures of student achievement are in question, these scores cannot be used to make future educational decisions. Furthermore, one validation study is not sufficient to guarantee the psychometric properties of an assessment throughout its lifetime. Instead, the assessment(s) and policies used, in contexts such as placement testing, need to be

continuously reviewed and evaluated to assure that students are being placed into courses commensurate with their ability in order to maximize the chances of success (Linn, 1994; Mattern & Packman, 2009; McFate & Olmsted III, 1999; Norman et al., 2011; Wiggins, 1989). Overall, when properly constructed and evaluated, assessments can enhance later performance and provide feedback on what has and has not been learned to both the student and other interested stakeholders.

The purpose of this study was to provide evidence of Construct Validity and Internal Consistency Reliability of a mathematics placement test at a Science, Technology, Engineering, and Mathematics (STEM), gifted, residential high school. Previous research on placement exams have been conducted at the post-secondary level; however, this study extends the research to younger grade levels serving a specific, gifted population.

Literature Review

Although research has not extensively examined placement testing from middle school to high school, a large literature base exists using college and university student populations. In fact, approximately 90% of post-secondary institutions use placement tests (Latterell & Regal, 2003). The near-universal practice of administering placement tests emerged due to the incomparability of unknown factors such as the content and rigor of courses and the grading scales used at different schools (Kossack, 1942; Linn, 1994; Ngo & Kwon, 2015; Noble et al., 2003). Within the setting of a post-secondary institution, students complete placement tests to determine the appropriate level of beginning coursework. In the same way, once students are accepted into the high school

of the current study, they too must complete a series of placement tests to guide their initial course enrollment decisions.

The overarching purpose of placement tests is to match students with a level of instruction that is appropriate given their previous academic preparations (Akst & Hirsch, 1991; Frisbie, 1982; Marshall & Allen, 2000; Mattern & Packman, 2009; McFate & Olmsted III, 1999; Noble et al., 2003; Sawyer, 1996). Prior research has shown that course placement decisions can have a significant impact on a student's future academic preparation (McDaniel et al., 2007; Morgan & Michaelides, 2005). For example, students who begin post-secondary mathematics in a course that is appropriate given their background have an increased chance of succeeding in their first course in addition to subsequent mathematics courses (Mattern & Packman, 2009; Norman et al., 2011; Shaw, 1997). For this reason, more research is needed to thoroughly examine placement tests and procedures to ensure that student success is maximized while the consequences of misplacement are minimized. Although these placement tests are typically considered "high-stakes," the psychometric properties of such tests have received relatively little attention (Callahan, 2005; Grubb & Worthen, 1999; Scott-Clayton, 2012). As a result, more research is needed to investigate and evidence the psychometric properties of placement tests.

Validity

Validity is typically defined as the extent to which an instrument measures what it is intended to measure (Wiersma & Jurs, 2009). While this definition is somewhat accurate, it is often times misleading. That is, the instrument itself is not validated, rather

the conclusions and interpretations drawn from the scores have validation evidence (Cook & Beckman, 2006; Ebel, 1956; Kimberlin & Winetrstein, 2008; Messick, 1995; Moss, 1992; Schmitz & delMas, 1991). Using this specificity, validation is defined by Cronbach (1971) as an evidence collecting process in order to support the inferences being made from the test scores. The three major types of validity are Content Validity, Construct Validity, and Criterion-Related Validity, with Construct Validity being the focus of the current study.

Construct Validation refers to a process by which a judgment is made regarding whether or not an instrument adequately measures the intended construct. A construct, also referred to as a latent variable, is not directly observable and has been defined as "some postulated attribute of people, assumed to be reflected in test performance" (Cronbach & Meehl, 1955, p. 283). Commonly studied psychological constructs include anxiety, achievement, and personality. In order to measure a construct of interest, researchers emphasize the need to transform a conceptual definition into an operational definition. The operational definition acts as a bridge to connect the conceptual definition to more concrete observations or indicators. These observations are then assigned numbers to represent how much of the construct an individual possesses.

Aspects of Construct Validation are typically reviewed during the instrument development phase. During this time, the construct of interest and its associated content are manifested into concrete tasks that individuals must complete. In the context of educational assessment, content standards of a course are translated into performance standards which further define "how much of the content standards students must know

and be able to do to achieve a particular level of competency" (Morgan & Michaelides, 2005, p. 1). Four widely used approaches to Construct Validation are: (1) the use of correlations between the construct and other variables, (2) differentiation between groups, (3) Factor Analysis, and (4) the Multitrait-Multimethod Matrix (Campbell & Fiske, 1959; Crocker & Algina, 2008). In the current study, evidence of Construct Validity was obtained through an Exploratory Factor Analysis (EFA).

Internal Consistency Reliability

Broadly stated, reliability measures the consistency or accuracy of the research and provides evidence to the extent to which the research can be repeated (e.g., Cook & Beckman, 2006; Cronbach, 1951; Nunnally & Bernstein, 1978; Rossi et al., 2003; Wiersma & Jurs, 2009). There are multiple different types of reliability (i.e., Test-Retest, Alternate Forms, and Internal Consistency) each of which have their specific uses. A discussion regarding the various types of reliability is beyond the scope of this study, and readers are encouraged to refer to measurement focused textbooks such as those by Allen and Yen (2001) or Crocker and Algina (2008) for further information.

In the current study, Internal Consistency Reliability was examined, which provides evidence that the items on an instrument are all related and measure the same construct (Cook & Beckman, 2006; Crocker & Algina, 2008; Henson, 2001; Kimberlin & Winetrstein, 2008; Wiersma & Jurs, 2009). This form of reliability only requires a single test administration (i.e., compared to forms of reliability requiring multiple administrations such as Test-Retest or Alternate Forms), which was appropriate to

examine in the current study since the mathematics placement test was only administered once to students (Feldt, Woodruff, & Salih, 1987).

At many institutions, the stated intention of placement testing is to prevent students from enrolling in courses for which they are inadequately prepared and/or unlikely to succeed. However, a common concern is that placement instruments may prevent "able" students from taking courses that they are actually prepared and capable to complete (Flores, 2007). Prior to discussing the effectiveness of the decision-making process, institutions must first provide evidence regarding the psychometric properties of their assessments. The purpose of the current study was to examine the Construct Validity and Internal Consistency Reliability of a mathematics placement test used at a STEM, gifted, residential high school.

Methods

The following sections describe the methods used to examine the Construct Validity and Internal Consistency Reliability of a mathematics placement test.

Participants and Procedures

Existing data from four cohorts of students were used to examine the research questions in this study. These cohorts consisted of students entering the high school their sophomore year, beginning in the 2014/2015 academic year and ending in the most recent 2017/2018 academic year, for which complete data were available.

Equivalence across the four cohorts was examined for five demographic variables using Chi-Square (χ^2) Tests of Association and One-Way Analyses of Variance (ANOVAs). Chi-Square Tests of Association were conducted across the four cohorts for

the variables of sex and race/ethnicity. There were no significant differences in the proportions between cohort year and either sex or race/ethnicity. For the three remaining variables of socioeconomic status (i.e., median family income), incoming SAT Math (SAT_M) subscores, and incoming SAT Evidence Based Reading and Writing (SAT_ERW) subscores, ANOVAs were used. Again, there were no significant differences between cohort years for each of the three variables. Therefore, all four cohorts were found to be statistically equivalent and were combined into one sample for further analysis.

Measure

The mathematics placement test was developed by mathematics faculty members in 1985. The original and continuing purpose of the mathematics placement test is to determine a student's incoming mathematical knowledge for appropriate initial course placement commensurate with ability level. Thus, generally speaking, the two-part placement exam assesses mathematical knowledge needed prior to entering into a Calculus sequence. However, neither of these parts nor the test as a whole have been subjected to psychometric evaluation, specifically using more advanced quantitative techniques such as Exploratory Factor Analysis (EFA).

Part I of the assessment measures student's knowledge of content such as simplifying expressions, functions, and exponents. Students are given 45 minutes to complete 50 short-answer items, without a calculator. All responses are graded by the mathematics faculty members using an answer key for dichotomous scoring (i.e., "Correct" or "Incorrect"). Thus, the possible range of scores on Part I is from 0 to 50.

After the allotted time has expired for Part I, exam proctors collect any remaining exams and distribute Part II.

The second portion of the exam gives students 85 minutes to complete 57 multiple-choice items, without a calculator, covering content such as graphing and evaluating functions, laws of exponents and logarithmic functions, right triangle trigonometry, and law of sines and cosines. The multiple-choice format used on this portion of the test provides students with the correct answer, three distractor answers, and a fifth response option of "I don't know."

Although not explicitly written on the test instructions, exam proctors emphasize the use of the "I don't know" option. By purposefully mentioning this, it is believed that students will not guess, but rather consider using the "I don't know" response option so that they do not accidentally place into a higher course than academically appropriate. A similar argument was made by Prieto and Delgado (1999) who noted that educational standards should not be influenced by desired psychometric properties of a test. Said another way, if students are unsure of an answer, it seems more appropriate for them to omit the item rather than encouraging them to guess. After the exam is complete, the multiple-choice items are scanned into a grading software program using a scantron reader where all items are scored dichotomously (i.e., "Correct" or "Incorrect"), even if the student selected the "I don't know" option. The possible range of scores is from 0 to 57 on Part II.

Data Analysis

In the current study, evidence of Construct Validity was obtained through an Exploratory Factor Analysis (EFA). Pett et al. (2003, p. 2) describe factor analysis as "a complex array of structure analyzing procedures used to identify the interrelationships among a large set of observed variables and then, through data reduction, to group a smaller set of these variables into dimensions or factors that have common characteristics." The two broad classifications of factor analysis are Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). Researchers use EFA when the underlying factor structure of the construct of interest is unknown (Pett et al., 2003; Thompson, 2004). CFA, on the other hand, is used when the researcher has some knowledge or understanding of the underlying factor structure from previous theories of the construct of interest. In the current study, the original factor structure of the mathematics placement test was unknown. Thus, an EFA was conducted using PRELIS and LISREL 9.30.

Moreover, previous research has long debated the appropriate sample size to conduct an EFA, with approximately 10 subjects per variable as the general consensus (Comrey & Lee, 1992; Costello & Osborne, 2005; Nunnally & Bernstein, 1978). In the current study, there were 107 items from the mathematics placement test that were factor analyzed. Using the 10:1 subject to variable ratio guideline, 1,070 cases are needed to conduct the EFA. As previously mentioned, each of the four cohorts contained approximately 280 students, which led to a final sample size of 1,125. Therefore, the

sample size of the current study surpassed the recommended 10:1 subject to variable ratio.

Assumptions. The main underlying assumption of EFA is that the observed variables are linear combinations of underlying hypothetical/unobservable factors (Kim & Mueller, 1978). The goal in this analysis is to condense the information contained in the original variables into a smaller set of factors with a minimal loss of information and simplest method of interpretation (i.e., parsimony; Hair Jr et al., 1995; Harman, 1976). That is, EFA, as an exploratory analytical technique, is used to understand the nature of the relationships between observed variables and factors and to account for the covariation between observed variables (Tucker & MacCallum, 1997). When discussing and analyzing linear combinations, mathematical theories and assumptions surrounding matrices are used.

Another assumption of EFA is univariate/multivariate normality, which refers to the shape of the distribution of data and its congruence to a normal distribution curve (Hair Jr et al., 1995). However, the current study data were dichotomously scored, and thus, this assumption was not examined. Similarly, a third consideration for conducting an EFA is the strength of the relationship between two items on an instrument. This information is typically summarized by the Pearson Product-Moment Correlation Coefficient Matrix, sometimes referred to as Pearson's *r* or the correlation matrix (Pett et al., 2003). Because the data are dichotomous, the strength of the relationship between two items on the instrument was assessed using the Tetrachoric Correlation Matrix. Tetrachoric Correlation Coefficients are used when the latent trait underlying the data is

theoretically continuous, but is measured dichotomously (Bonett & Price, 2005; Lorenzo-Seva & Ferrando, 2012; Uebersax, 2006b). In this study, the underlying latent trait was mathematical knowledge, which can be conceptualized as a continuous variable. However, this latent trait is scored dichotomously on the mathematics placement exam (i.e., scoring "Correct" or "Incorrect").

Furthermore, in order to use Tetrachoric Correlations, the following assumptions must be met: (1) the latent trait is normally distributed, (2) rating errors are normally distributed, (3) the variance is homogeneous across all levels of the latent trait, (4) errors are independent between items, and (5) errors are independent between cases (Uebersax, 2006b). The primary limitation of using Tetrachoric Correlations is that these assumptions cannot be mathematically tested.

The goal of factor analysis is to explain the interrelationships among variables, and it is important to have "acceptable" correlation coefficients. Various researchers have differing opinions on what constitutes an "acceptable" correlation coefficient, which is dependent upon the level of measurement of the variables (i.e., nominal, ordinal, interval, or ratio) and how the correlation coefficient is calculated. One generally accepted guideline for interpreting the Pearson Product-Moment Correlation Coefficient is that correlation values should be greater than or equal to .30 (Costello & Osborne, 2005; Pett et al., 2003; Stevens, 2012; Tabachnick & Fidell, 2007). Because the values of Tetrachoric Correlations values are interpreted similarly to Pearson's *r*, the above stated guideline was consulted when examining the Tetrachoric Correlation Matrix in the current study.

Exploratory factor analysis. Exploratory Factor Analysis (EFA) is considered to be a complex process that has many options and few absolute guidelines (Costello & Osborne, 2005). The following paragraphs describe the methods of factor extraction, rotation, solution refinement, and final interpretation used in the current study.

When conducting an EFA, the determinant of the correlation matrix is evaluated to determine if an inverse matrix exists. If the determinant of the correlation matrix is zero, an inverse matrix does not exist, implying that there are no interrelationships between the items (Pett et al., 2003). The correlation matrix would, in this case, not be called an identity matrix. These calculations can all be summarized in what is known as Bartlett's Test of Sphericity (Bartlett, 1950). In a similar way, the Tetrachoric Correlation Matrix calculated with dichotomous data can have a property called non-positive definiteness (Uebersax, 2006a). This occurs when one or more eigenvalues are negative, suggesting that there are linear dependencies among some items (Lorenzo-Seva & Ferrando, 2020). When linear dependencies are present, this indicates that one or more eigenvalues are close to zero, meaning that the matrix is close to being non-invertible (Margalit & Rabinoff, 2018; Pett et al., 2003). Thus, when negative eigenvalues are present and the matrix is close to being singular (i.e., non-invertible), then the extraction methods of Maximum Likelihood (ML) and Generalized Least Squares (GLS) cannot be used because of their reliance on the inverse matrix. Furthermore, ML and GLS extraction methods were not used in this study due to their underlying assumption of multivariate normality. Instead, the factor extraction method of Minimum Residuals (MINRES), which is equivalent to Unweighted Least Squares (ULS), was used since its

calculations do not rely on the inverse matrix or multivariate normality (Jöreskog, 2003; Uebersax, 2006a).

Regarding the number of factors to be extracted, the two prominent methods used for EFA include the Kaiser-Guttman Rule for eigenvalues (e.g., Comrey & Lee, 1992; Guttman, 1954; Kaiser, 1960; Nunnally & Bernstein, 1994) and the Scree Plot (Cattell, 1966). The Kaiser-Guttman Rule tends to be more objective in that this method extracts those factors whose eigenvalues are greater than 1. On the other hand, examining the Scree Plot requires more of a subjective decision about where the elbow of the plot is located and consequently how many factors should be retained. For these reasons, researchers tend to use a combination of these methods in EFA to guide decisions regarding the number of retained factors.

In the current study, the statistical software program PRELIS was used due to its ability to handle dichotomous data and calculate the Tetrachoric Correlation Matrix. However, Scree Plots are not rendered using this program. Previous research has indicated that results obtained through a Hierarchical Cluster Analysis (HCA) are similar to those obtained through factor analytic procedures (Capra, 2005; Revelle, 1979). For this reason, EFAs were conducted using existing cluster solutions (i.e., examined in Manuscript 1), as a guide for the number of factors to extract. Therefore, as EFA is an explanatory, theory-driven data analytic strategy, additional iterations of the data were conducted with a specific number of factors to extract that were both above and below the previous amounts.

The next consideration when planning an EFA is rotation of the extracted factors, which aids in simplifying and clarifying the underlying data structure (i.e., to obtain simple structure). Simple structure is attained when there are high item loadings on one factor and smaller item loadings on the remaining factors, resulting in a "cleaner" factor solution that is more easily interpreted (Costello & Osborne, 2005; Williams, Onsman, & Brown, 2010). The two common approaches in data rotation are orthogonal and oblique, each having different underlying assumptions.

An orthogonal rotation assumes that the underlying factors are uncorrelated, whereas an oblique rotation assumes the opposite (e.g., Costello & Osborne, 2005; Gorsuch, 1983; Pett et al., 2003; Thompson, 2004). Since the underlying latent trait is mathematical knowledge, a relationship among the underlying factors was expected, necessitating the use of an oblique rotation. Of the possible oblique rotation methods (i.e., Direct Oblimin, Promax, Orthoblique), the Promax rotation was used in the current study. One advantage of the Promax rotation is that it begins with an orthogonal rotation, allowing for the possibility that the underlying factors are in fact uncorrelated (Pett et al., 2003). Additionally, Gorsuch (1983) argued that the Promax rotation ultimately results in stronger correlations between factors and achieves a more simple structure. Accordingly, the oblique rotation method Promax was used.

Using information from the above mentioned model specifications, the default factor extraction solution was examined for its representativeness and overall fit to the data. Again, since this was an EFA and the underlying factor structure was unknown, additional factor extraction solutions were explored and compared to the initial solution.

In doing so, the final interpretation of the factor structure was supported through evidence from the collection of models, including but not limited to the amount of variance explained, the factor loadings, and the correlations between factors.

Internal consistency reliability. Reliability refers to the degree to which data collection, data analysis, and data interpretations are consistent provided the surrounding conditions remain constant (Wiersma & Jurs, 2009). As such, Internal Consistency Reliability provides evidence of accuracy of results when the same measure is used. Moreover, "internal consistency" would suggest that the items within a measure correlate strongly with one another (Henson, 2001; Kimberlin & Winetrstein, 2008).

Two well-known methods that assess Internal Consistency Reliability are Coefficient (Cronbach's) Alpha and the Kuder-Richardson Formulas (Cronbach, 1951; Kuder & Richardson, 1937). As shown below, previous research has demonstrated the equality of Cronbach's Alpha and the Kuder-Richardson Formulas (e.g., Cliff, 1984; Crocker & Algina, 2008; Feldt, 1969; Onwuegbuzie & Daniel, 2002) for binary data. Cortina (1993) elaborated further by stating that Cronbach's Alpha is a more general version than the Kuder-Richardson estimate. Cronbach's Alpha can be calculated by using the formula

$$\widehat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \hat{\sigma}_i^2}{\hat{\sigma}_X^2}\right) \quad [6]$$

where k is the number of items on the test, $\hat{\sigma}_i^2$ is the variance of item i, and $\hat{\sigma}_X^2$ is the total test variance. Likewise, with a simple substitution of pq for the variance of item i, the Kuder-Richardson estimate is calculated as follows:

$$KR_{20} = \frac{k}{k-1}\left(1 - \frac{\sum pq}{\hat{\sigma}_X^2}\right) \quad [7]$$

However, when items are dichotomously scored, although equal, the Kuder-Richardson Formula (KR-20) is preferred over Cronbach's Alpha.

Researchers Kuder and Richardson (1937) developed two formulas for estimating internal consistency reliability, namely the KR-20 and the KR-21. While computed similarly, the KR-20 and KR-21 formulas differ in their assumption of item difficulties. If each item is assumed to have the same level of difficulty, then the KR-21 formula can be used (Crocker & Algina, 2008; Kuder & Richardson, 1937; Onwuegbuzie & Daniel, 2002). However, the current study assumes that the item difficulties vary, which necessitates calculating KR-20 as the estimate of internal consistency reliability.

Considerable attention has been given to the range of generally accepted values for Cronbach's Alpha and KR-20 indices. While an internal consistency reliability estimate of .70 may be advisable in some contexts of exploratory research (Nunnally & Bernstein, 1978), Ding and Beichner (2009) suggested that the value of KR-20 be greater than or equal to .80. More specifically, when a particular test score is used for important clinical and/or educational decisions (e.g., course placement), the estimates of internal consistency reliability should have a minimum value of .90, with .95 considered desirable (e.g., Henson, 2001; Hopkins, 1998; Nunnally & Bernstein, 1994; Oosterhof, 2001; Rossi et al., 2003). Therefore, a minimum internal consistency reliability estimate of .90 was considered the standard for the Mathematics Placement Test in the current study.

Finally, the term internal consistency suggests that items measuring the same construct should to some degree correlate with one another (Crocker & Algina, 2008;

Ding & Beichner, 2009; Henson, 2001; Kimberlin & Winetrstein, 2008). Clark and Watson (1995) recommend that the average inter-item correlation coefficient range between .15 and .20 for scales measuring broad characteristics and .40 and .50 for those measuring narrower characteristics. Since the relationships between items are unknown, inter-item correlation coefficients ranging from .15 to .50 was considered acceptable in the current study.

Results

An Exploratory Factor Analysis (EFA) using Minimum Residuals (MINRES) and oblique (Promax) rotation was conducted to examine the internal structure of the mathematics placement exam. In the final sample (N = 1,125), 472 (42.0%) were Male, 468 (41.6%) were Female, with the remaining 185 (16.4%) not reported at the time of testing. The following race/ethnicities were represented in the EFA sample: Asian (n = 383), Black or African American (n = 69), Hispanic or Latino (n = 80), Two or More Races (n = 53) and White (n = 355). According to the data, students had an average incoming SAT Math subscore of 680.60 (SD = 78.94) and an average incoming SAT Evidence-Based Reading and Writing subscore of 642.46 (SD = 65.31).

The Tetrachoric Correlation Matrix was examined to identify the degree of the relationships between item pairs (available upon request). Interpreted similarly to Pearson's r, if a Tetrachoric correlation coefficient was greater than or equal to .30, it was considered acceptable. Positive correlation coefficients ranged from .002 to .929, while the negative correlation coefficients ranged from -.189 to -.002. Examining the 107 items, fifteen of the items had a weak correlation with a majority of the other items.

However, most item pairs displayed Tetrachoric correlations above .30, suggesting that a factor analysis is appropriate for these data.

Exploratory Factor Analysis

Previous research has indicated that results obtained through a Hierarchical Cluster Analysis (HCA) are similar to those obtained through factor analytic procedures (Capra, 2005; Revelle, 1979). For this reason, EFAs were conducted using the existing cluster solutions (i.e., examined in Manuscript 1) for the number of factors to extract. The factor analysis results for three, eight, and six factors were explored and compared in order to identify the best underlying structure. In both the eight and six factor solutions, Heywood cases were found and removed prior to conducting additional iterations (Lorenzo-Seva & Ferrando, 2020).

The final factor solution revealed the presence of three related components. The correlation between factors ranged from .449 (Factors 1 and 2) to .618 (Factors 1 and 3). Factors 2 and 3 also had a moderate correlation value of .531. Analysis of the Rotated Factor Loading Matrix demonstrated that a majority of the items had a moderate to strong relationship with at least one of the factors and more often than not, values exceeded .400 (see Table 4 below). Factor loadings on the first factor ranged from .141 (FR2) to .888 (MC45). Factor 2 had a minimum factor loading of .270 (FR46) and a maximum factor loading of .855 (MC53). The third factor had the smallest overall factor loading of -.191 (FR11) and a maximum factor loading of .884 (FR28).

Table 4

Promax - Rotated Factor Matrix

Item	PreCalculus	Geometry	Algebra 1	Unique Variance
FR1	0.142	0.186	**0.218**	0.793
FR2	**0.141**	0.028	0.066	0.958
FR3	0.066	0.050	**0.337**	0.831
FR4	0.014	0.033	**0.456**	0.767
FR5	0.113	-0.145	**0.688**	0.518
FR6	-0.034	-0.109	**0.771**	0.511
FR7	0.263	0.076	**0.525**	0.418
FR8	0.123	0.006	**0.599**	0.531
FR9	0.024	0.092	**0.599**	0.554
FR10	0.195	-0.102	**0.719**	0.356
FR11	0.065	0.092	**-0.191**	0.979
FR12	0.107	-0.017	**0.475**	0.710
FR13	0.067	0.087	**0.668**	0.419
FR14	0.123	0.093	**0.299**	0.802
FR15	0.023	0.023	**0.635**	0.561
FR16	-0.182	0.099	**0.707**	0.558
FR17	0.200	-0.005	**0.688**	0.321
FR18	0.182	0.065	**0.683**	0.285
FR19	-0.079	0.120	**0.547**	0.673
FR20	0.029	0.140	**0.586**	0.525
FR21	0.118	0.021	**0.689**	0.392
FR22	-0.030	0.176	**0.572**	0.560
FR23	0.251	0.014	**0.473**	0.555
FR24	0.130	0.001	**0.497**	0.655
FR25	0.052	0.061	**0.447**	0.733
FR26	0.043	-0.007	**0.787**	0.344
FR27	-0.140	-0.027	**0.793**	0.507
FR28	0.042	-0.198	**0.884**	0.326
FR29	0.200	-0.129	**0.761**	0.304
FR30	0.000	-0.067	**0.750**	0.486
FR31	0.167	0.035	**0.399**	0.709
FR32	0.017	0.019	**0.589**	0.628
FR33	0.360	-0.133	**0.732**	0.138
FR34	0.317	-0.050	**0.628**	0.304

FR35	0.116	0.226	**0.349**	0.657
FR36	-0.084	0.127	**0.597**	0.611
FR37	0.076	-0.090	**0.640**	0.583
FR38	0.151	0.037	**0.504**	0.603
FR39	0.275	-0.009	**0.388**	0.647
FR40	0.006	0.085	**0.590**	0.586
FR41	-0.057	0.206	**0.548**	0.583
FR42	0.093	0.122	**0.147**	0.909
FR43	0.136	0.177	**0.478**	0.530
FR44	0.102	-0.035	**0.605**	0.572
FR45	-0.105	0.095	**0.568**	0.682
FR46	-0.129	**0.270**	0.259	0.841
FR47	-0.104	0.074	**0.462**	0.800
FR48	0.133	0.123	**0.262**	0.806
FR49	0.184	-0.016	**0.619**	0.456
FR50	-0.054	0.071	**0.320**	0.891
MC1	**0.516**	-0.223	0.232	0.641
MC2	**0.301**	0.104	0.059	0.838
MC3	0.160	0.221	**0.483**	0.451
MC4	0.425	-0.009	**0.561**	0.219
MC5	0.141	-0.003	**0.655**	0.438
MC6	0.194	-0.108	**0.522**	0.632
MC7	0.232	0.064	**0.580**	0.386
MC8	0.315	-0.009	**0.668**	0.203
MC9	0.134	-0.041	**0.609**	0.541
MC10	**0.680**	-0.116	0.190	0.422
MC11	**0.476**	0.055	0.152	0.625
MC12	**0.489**	0.217	0.265	0.328
MC13	**0.505**	-0.052	0.462	0.289
MC14	**0.359**	0.071	0.259	0.642
MC15	**0.468**	0.035	0.261	0.536
MC16	0.395	-0.089	**0.608**	0.258
MC17	**0.634**	0.001	0.226	0.369
MC18	**0.624**	-0.087	0.359	0.279
MC19	**0.480**	-0.064	0.397	0.428
MC20	**0.667**	-0.194	0.414	0.207
MC21	**0.476**	0.080	0.429	0.260
MC22	**0.475**	0.099	0.318	0.400

MC23	**0.612**	0.133	0.170	0.352
MC24	**0.555**	-0.007	0.174	0.547
MC25	0.204	**0.322**	0.181	0.655
MC26	**0.805**	-0.036	0.069	0.307
MC27	**0.809**	0.125	-0.031	0.273
MC28	**0.722**	0.010	0.094	0.378
MC29	**0.832**	-0.001	0.081	0.219
MC30	**0.809**	-0.009	0.107	0.235
MC31	**0.757**	-0.086	0.141	0.339
MC32	**0.696**	-0.068	0.218	0.334
MC33	**0.690**	-0.043	0.099	0.459
MC34	**0.728**	-0.025	0.091	0.397
MC35	**0.787**	0.103	0.119	0.154
MC36	**0.790**	0.027	0.065	0.286
MC37	**0.784**	0.149	0.051	0.198
MC38	**0.423**	0.220	0.157	0.546
MC39	**0.529**	0.094	0.066	0.613
MC40	**0.558**	-0.002	0.271	0.430
MC41	**0.607**	0.112	0.169	0.384
MC42	**0.664**	0.329	-0.122	0.382
MC43	**0.614**	0.179	-0.090	0.569
MC44	**0.502**	0.125	0.103	0.588
MC45	**0.888**	0.137	-0.320	0.378
MC51	-0.107	**0.804**	0.015	0.408
MC52	-0.096	**0.560**	0.083	0.679
MC53	0.016	**0.855**	-0.112	0.347
MC54	-0.127	**0.469**	0.136	0.752
MC55	-0.027	**0.629**	0.081	0.562
MC56	0.127	**0.467**	-0.167	0.794
MC57	0.178	**0.847**	-0.404	0.405
MC58	0.000	**0.527**	0.089	0.665
MC59	-0.066	**0.773**	-0.192	0.549
MC60	0.050	**0.451**	0.029	0.757
MC61	-0.053	**0.380**	0.277	0.701
MC62	-0.025	**0.343**	0.188	0.791

The naming conventions for each factor were determined by examining the items with the highest factor loadings on each component. The four highest loadings on Factor 1 were .888 (MC45), .832 (MC29), and .809 (MC27 and MC30), which covered content such as polar graphs and trigonometry typically found in an upper level PreCalculus course. Factor 2 had three prominent factor loadings of .855 (MC53), .847 (MC57), and .804 (MC51). The content of these items covered topics generally found in a Geometry course such as congruent triangles, using the properties of angles for two parallel lines cut by a transversal, and proving two angles are congruent. Lastly, some of the highest loadings on Factor 3 were .884 (FR28), .793 (FR27), and .771 (FR6). These three items asked students to manipulate polynomials using their knowledge of the laws of exponents (i.e., multiply, expand, and factor). Based on this information along with the all of the factor loadings displayed above, the final three factors were determined to be PreCalculus, Geometry, and Algebra 1, respectively. Additionally, evidence of simple structure was shown as revealed that several items had a factor loading of .70 or higher on a single factor and only four items had a strong cross-loading between factors (i.e., the factor loading for a single item was greater than or equal to .400 on more than one factor).

Internal Consistency Reliability

Once the final factor structure was determined, the Internal Consistency Reliability estimates were calculated for each factor: PreCalculus KR-20 = .950, Geometry KR-20 = .736, and Algebra 1 KR-20 = .910. The internal consistency within factors was strong, even on the second factor (i.e., Geometry) which only consisted of 14

items. Overall, the information obtained through the EFA suggests that the items on the mathematics placement test can be represented by three underlying factors. Due to the moderate correlations among factors, the instrument adequately measures the larger construct of students' mathematical knowledge, providing preliminary evidence of Construct Validity.

Discussion

The purpose of this study was to provide evidence of Construct Validity and Internal Consistency Reliability of a mathematics placement test at a specialized STEM high school. Using Exploratory Factor Analysis (EFA) and the Kuder-Richardson Formula (KR-20), the psychometric properties of the exam were evidenced.

Exploratory Factor Analysis

EFA was used to examine the underlying factor structure of the mathematics placement test based on the students' responses to the 107 items. Using a large sample size (N = 1,125), an EFA with Promax rotation was conducted. The initial number of factors to extract was guided by the results of a Hierarchical Cluster Analysis (HCA) (Capra, 2005; Revelle, 1979). Factor solutions for eight and six factors were analyzed, but due to the presence of Heywood cases and a lack of simple structure, other factor solutions were explored.

The final iteration revealed three distinct factors with 37 items loading on Factor 1, 14 items on Factor 2, and 56 items on Factor 3. After examining the items that loaded highest on each factor, the factor labels were developed using the most prominent content

found within those items. Thus, the three final factor labels were PreCalculus, Geometry, and Algebra 1, respectively.

The labels assigned to each of the three factors were similar to the original content areas of interest as determined by the faculty members who created the exam. Recall that the mathematics placement test is a two-part exam measuring students' mathematical knowledge needed prior to entering into a Calculus sequence. Part I of the assessment consists of 50 short-answer items covering content such as simplifying expressions, functions, and exponents. As can be seen from the EFA results above, the strongest loading for the vast majority of these items (i.e., FR1 – FR50) occurred on Factor 3 which was labeled as Algebra 1. The second part of the exam was developed to measure students' knowledge of topics such as evaluating and graphing functions of higher order, using the properties and laws of sine and cosine, and providing evidence to show the congruence of either two angles or two triangles. As determined through the EFA, there was a distinct division between the Geometry content and the former items encompassing functions and trigonometry, which were more broadly labeled as PreCalculus.

Internal Consistency Reliability

The reliability of each factor was calculated using the Kuder-Richardson (KR-20) Formula (Kuder & Richardson, 1937). Due to the high-stakes nature of this exam and its use in course placement decisions, this study considered a minimum reliability estimate of .90 to be acceptable. Thus, the two factors of PreCalculus (KR-20 = .950) and Algebra 1 (KR-20 = .910) were determined to have acceptable values for reliability while the Geometry factor (KR-20 = .736) was lower than expected. From the literature it is noted

that reliability has a direct relationship with the number of items being examined such that as the number of items increase, so does the reliability estimate (Cortina, 1993; Crocker & Algina, 2008; Wiersma & Jurs, 2009). This was evidenced in the current study as the Geometry factor had the lowest reliability for its 14 items compared to the PreCalculus and Algebra 1 factors, which had acceptable reliability estimates given their 37 and 56 items, respectively.

Overall, the EFA and Internal Consistency Reliability results provide evidence that the mathematics placement test is a valid and reliable measure. More specifically, higher total scores on the mathematics placement test indicates more mathematical knowledge prior to Calculus.

Implications

In the context of large-scale testing (e.g., course placement), psychometric analysis is essential in determining the quality of the test and the information it generates (Adedoyin & Mokobi, 2013). By critically examining the mathematics placement test and its psychometric properties, all stakeholders can be assured that the inferences drawn from the educational assessment are accurate (Harris, 2003; Linn, 1994).

The overarching purpose of placement testing is to enroll students in courses that are commensurate with their ability in order to maximize the chances of success and minimize the unintended, negative consequences (Linn, 1994; Mattern & Packman, 2009; McFate & Olmsted III, 1999; Norman et al., 2011; Wiggins, 1989). By providing evidence of Construct Validity, both students and their parents can be confident in knowing that this assessment measures students' incoming mathematical knowledge

leading up to a Calculus sequence so that proper course placement decisions can be made. Furthermore, demonstrating evidence of strong Internal Consistency Reliability suggests that students' true level of mathematical knowledge is consistently represented by the items, again decreasing the number of inappropriate course placement decisions being made and minimizing the temporary and lasting negative effects on students (Frisbie, 1988).

Secondly, the results of this study have practical benefits for the faculty and educational administrators at the gifted residential high school. Every year, students entering the high school have increased cultural diversity, life experiences, family influences, and their level of preparedness for a challenging college-preparatory curriculum. Thus, by continually demonstrating evidence of validity and reliability, mathematics faculty members can confidently rely on the scores from the mathematics placement test as accurate measures of achievement and can use the scores to make important course placement decisions. Moreover, when faculty become equipped with such diagnostic information, they can better distinguish between students who do or do not need additional academic assistance in their initial mathematics course (Betts, Hahn, & Zau, 2011).

Evidence-based research in education emphasizes evaluating the outcomes of programs and the processes that lead to these outcomes (Slavin, 2007). Additionally, the *Code of Fair Testing Practices in Education* (Nitko & Brookhart, 2011) calls test developers to provide evidence that the technical quality, including validity and reliability, of the test aligns with its intended uses. This study provides an initial step in

demonstrating the psychometric properties of the mathematics placement test to both statewide and local stakeholders. Furthermore, this study emphasizes the importance of educational assessment in the hopes that administrators and faculty alike will use this study as a "template" in additional departments within the high school and similar contexts.

Finally, the implications from this study extend beyond the local context. Placement exams that are valid and reliable are vital to both post-secondary institutions and other gifted STEM residential high schools like the one in the current study. Although the average high school may not have sufficient resources to conduct similar research, there is still a need to have solid and defensible placement tests and practices. The current study can act as a blueprint for similar high schools to begin the assessment validation process at their own institutions.

Limitations and Future Research

This study included data from four cohorts of students applying to a residential STEM high school for gifted children. As such, the content measured on the specific mathematics placement test used in this study, as well as the scores obtained from the assessment, are unique to the school and are not generalizable to other STEM high schools. However, if other similar high schools seek to examine the psychometric properties of their placement exams, the procedures used in this study could be replicated.

Construct Validity was evidenced in the current study using Exploratory Factor Analysis. Since the underlying factor structure was unknown, the number of factors to

extract could not be supported by previous theoretical evidence on the construct of interest. Instead, the current study used results from a Hierarchical Cluster Analysis (Manuscript 1) to determine how many factors to extract in the initial EFA solutions. While this method is supported in the literature, future research should examine the congruencies among HCA and EFA solutions (Capra, 2005; Revelle, 1979).

Comparing the results from the HCA (Manuscript 1) and EFA, the following observations were noted. The Geometry cluster from the HCA had a direct relationship to the Geometry factor of the EFA (i.e., the same items in both). Likewise, all items (i.e., except one) from the HCA Trigonometry cluster loaded the highest on the PreCalculus factor of the EFA. This relationship between the Trigonometry cluster and the PreCalculus factor was expected based on the sequence and design of the high school mathematics courses.

Next, the items in the first two clusters of the HCA (i.e., Algebraic Operations and Solving Equations) were mainly located in the Algebra 1 factor of the EFA. However, the clusters of Graphing and Evaluating Functions were split between the Algebra 1 and the PreCalculus factor. The distinction between the two factors appeared to be related to the placement of the items on the exam. Since mathematical knowledge is hierarchical in nature, meaning that you need to know Algebra first before completing PreCalculus, the majority of the earlier items on the exam loaded on the Algebra 1 factor. Conversely, the items that loaded highest on the PreCalculus factor from clusters three and four were the items involving graphing and evaluating higher order functions. Therefore, there appears to be reasonable evidence to support the similarity of results between the HCA and the

EFA, but a more thorough investigation is needed to further confirm the presence and relationship between Content and Construct Validity.

Another limitation of the current study was the presence of negative variance estimates (i.e., Heywood cases) in the eight and six factor solutions. Heywood cases can appear for a variety of reasons, such as insufficient sample size compared to the number of variables, a large percentage of missing data, or attempting to extract more factors than necessary (Steinberg, 2010). The sample size of the current study was sufficient according to the guidelines of ten subjects per variable for EFA (Comrey & Lee, 1992; Costello & Osborne, 2005; Nunnally & Bernstein, 1978). Additionally, there was only a small percentage of missing data due to the high-stakes nature of the mathematics placement test. Thus, it is possible that extracting eight or six factors were more than what was necessary for the current study. Future research could examine the impact of statistical corrections involving the Heywood cases to determine the appropriate factor solution.

As previously discussed, the final factor structure revealed a three-factor solution of PreCalculus (37 items), Geometry (14 items), and Algebra 1 (56 items). These study results suggest a dramatically imbalanced factor structure, which may warrant further examination. While not all factors need to include the same number of items, it appears from the analysis that Geometry concepts are underrepresented on the mathematics placement test. Additionally, four items from the assessment (i.e., MC4, MC13, MC20, and MC21) cross-loaded between the PreCalculus and Algebra 1 factors, suggesting a possible overlap in content. Future research could include an item analysis to investigate

the item characteristics and potential local dependence between item pairs. By using Item Response Theory techniques, the mathematics placement test can be optimized for future administrations.

Conclusions

This study examined the psychometric properties (i.e., Construct Validity and Internal Consistency Reliability) of the scores on the mathematics placement test used at a gifted residential high school focused on STEM. Mathematics faculty members developed this assessment in 1985 with the intention of measuring students' incoming mathematical knowledge prior to Calculus so that they could properly assign students to their initial mathematics course. Using Exploratory Factor Analysis, it was determined that the mathematics placement test is comprised of three underlying factors, namely PreCalculus, Geometry, and Algebra 1, providing evidence of Construct Validity. Moreover, strong Internal Consistency Reliability, using the Kuder-Richardson (KR-20) Formula, suggest that the items on each factor are related and measuring the same construct.

These results demonstrate that the mathematics placement test is valid and reliable for the population of interest. Therefore, this assessment can be used in the course placement process to measure students' mathematical knowledge leading up to Calculus. Not only is this study important for the educational institution involved, but it is also relevant to other similar STEM high schools for gifted students. In a world of evidence-based practice, this study can act as a catalyst for educational institutions, at all levels, to conduct assessment research and provide evidence regarding the effectiveness

of their placement procedures and measures. In doing so, all stakeholders can be assured that the consequences of misplacement have been minimized while enhancing students' future educational outcomes.

CHAPTER VI – MANUSCRIPT 3

A PSYCHOMETRIC ANALYSIS OF A MATHEMATICS PLACEMENT EXAM: ITEM RESPONSE THEORY AND DIFFERENTIAL ITEM FUNCTIONING

Abstract

The near universal use of placement testing at the post-secondary level arose due to an assortment of unknown factors that could not be directly compared such as the content and rigor of previous courses and the grading scales used at different schools (Kossack, 1942; Linn, 1994; Ngo & Kwon, 2015; Noble et al., 2003). The overarching purpose of placement testing is to determine a student's incoming knowledge for appropriate course placement given their previous coursework. However, to be useful, empirical evidence must come from psychometric analysis of the items to demonstrate that they are well constructed and unambiguous (R. F. Burton, 2005).

The current study examined the item parameters (i.e., item difficulty, and item discrimination) and Differential Item Functioning (DIF) of a mathematics placement test at a Science, Technology, Engineering, and Mathematics (STEM) gifted residential high school. Existing data from four cohorts were obtained and analyzed using Item Response Thoery (IRT), specifically the Two-Parameter Logistic (2PL) Model. Results indicated that the exam was generally "easy" (i.e., the majority of students correctly answered a large number of items on the test) for the population of interest, and may not adequately discriminate among students with varying levels of mathematical knowledge. Items recommended for revision and concerns of item bias are discussed.

Keywords: Item Response Theory, Differential Item Functioning, STEM Education, Mathematics Placement Testing

Introduction

Validity, reliability, comparability, and fairness are just a few of the important elements involved in psychometric appraisal. These terms are not just measurement principles, but are also considered social values that have significant meaning and impact when evaluative judgments and decisions are made (Messick, 1995). As a result, educational institutions using placement exams must address questions about the uses and interpretations of tests and their scoring methods. In order to do so, measurement professionals must first begin with evaluating the test itself to ensure that the items are well constructed, unambiguous, and free of bias (Adedoyin & Mokobi, 2013; R. F. Burton, 2005; Sireci, 1998b). Once the quality of the test has been analyzed and professionals are confident in the characteristics of the test scores, then stakeholders can be assured that the outcomes of the assessment do not lead to uneven or unfair treatment of students, allowing more accurate inferences to be made.

One major limitation, however, is the lack of resources available to examine such characteristics of test scores. While most institutions of higher education have individuals with expertise in assessment, evaluation, and/or measurement, independent schools and schools at the secondary educational level often times do not. As a result, teachers are left to create their own assessments, including placement tests, without having adequate formal training in measurement techniques (Ryan, 2018). For this reason, STEM (i.e., Science, Technology, Engineering, and Mathematics) teacher

organizations and researchers both agree that stronger partnerships between K-12 educational entities and institutions of higher education can be formed to further guide the test development and evaluation process (Sondergeld, 2014).

Using the abovementioned partnership, the current study analyzed the psychometric properties of a mathematics placement test at a gifted, residential STEM high school. More specifically, the purpose of this study was to examine the item parameters (i.e., item difficulty, and item discrimination) and Differential Item Functioning (DIF) of the mathematics placement test using the Two-Parameter Logistic (2PL) Model from Item Response Theory.

Literature Review

The primary objective of achievement testing is to measure students' actual knowledge or acquired skills in order to reliably distinguish between students who do and do not have some level of the construct of interest (McFate & Olmsted III, 1999; Schmitz & delMas, 1991; Slavin, 2007). As such, course placement has become a typical and important use of achievement tests. This is evidenced by the near-universal use of placement tests at the post-secondary level, which emerged due to the difficulty in comparing factors such as the content and rigor of courses and the grading scales used at different schools (Kossack, 1942; Linn, 1994; Ngo & Kwon, 2015; Noble et al., 2003). Environments such as post-secondary education and specialized high schools with varying student experiences and backgrounds can benefit from having a standardized assessment that allows for comparisons to be made among students.

The overarching purpose of placement testing is to match students with an appropriate level of instruction and course material given their previous academic preparations (e.g., Akst & Hirsch, 1991; Frisbie, 1982; Marshall & Allen, 2000; Mattern & Packman, 2009; McFate & Olmsted III, 1999; Noble et al., 2003; Sawyer, 1996). For the process of placement testing administration and score use to be considered successful, it must demonstrate increased accurate placement decisions and a minimal number of inaccurate placement decisions (Harris, 2003; Linn, 1994; Schmitz & delMas, 1991). Undoubtedly, a greater amount of inaccurate placements can be problematic for institutions when underprepared students enroll in, and ultimately fail, a course (McFate & Olmsted III, 1999).

Prior research has shown that course placement decisions can have a significant impact on a student's future academic preparation (McDaniel et al., 2007; Morgan & Michaelides, 2005). For example, students who begin post-secondary mathematics in a course that is appropriate given their background have an increased chance of succeeding in their first course in addition to subsequent mathematics courses (Latterell & Regal, 2003; Mattern & Packman, 2009; Morgan & Michaelides, 2005; Norman et al., 2011; Shaw, 1997). However, when nearly one-third of all students entering community colleges are assigned to take at least one remedial or developmental mathematics course, students experience lower levels of motivation along with increased time and cost to graduation (e.g., Bailey, 2009; Hoyt & Sorensen, 2001; Kowski, 2013; Medhanie et al., 2012; Melguizo et al., 2008; Melguizo et al., 2014; Ngo & Kwon, 2015; Scott-Clayton, 2012). For these reasons, more research is needed to thoroughly examine the

psychometric properties of placement tests to ensure that student success is maximized while the consequences of misplacement are minimized.

Reviewing the psychometric properties of the items and the test also includes an examination of item bias. Instruments such as placement tests should be free from bias related to characteristics irrelevant to the construct of interest (i.e., sex, race, ethnicity, socio-economic status, age; Schmeiser, 1995). Specific to gender differences and item bias, research has revealed the importance of ensuring that placement decisions based on test scores are equally valid for males and females (Mattern & Packman, 2009).

Historically, the field of mathematics has been dominated by men and since the early 1980s, males have continued to take more advanced mathematics courses in high school compared to females (Catsambis, 1994; Pedro et al., 1981). Additionally, research has found that males outperform females on standardized assessments such as the mathematics subtests of both the SAT and ACT (Bridgeman & Wendler, 1989, 1991; Davis & Shih, 2007; Educational Testing Service, 1989; Gallagher & De Lisi, 1994). Even among the high-achieving math students, males have a consistent advantage over females, who are underrepresented in both upper level math courses and subsequent STEM careers.

While that narrative still persists, some research suggests that the gender achievement gap in mathematics may be narrowing. For example, more recent meta-analyses have reported that gender differences in mathematics scores on standardized assessments are minimal and non-significant, concluding that girls have reached parity with boys (Else-Quest, Hyde, & Linn, 2010; Hyde, Lindberg, Linn, Ellis, & Williams,

2008; Lindberg, Hyde, Petersen, & Linn, 2010; Reilly, Neumann, & Andrews, 2015). Other studies demonstrate that girls outperform boys in terms of their grades received in their mathematics courses (Arslan, Canli, & Sabo, 2012; Ding, Song, & Richardson, 2006; Gherasim, Butnaru, & Mairean, 2013; Wang & Degol, 2017). In the majority of studies, these conclusions have been drawn from substantive studies of mean achievement differences for boys versus girls. Fewer psychometric studies exist that address concerns of item bias on these assessments.

As previously mentioned, placement tests should be free from bias with respect to characteristics such as sex, race/ethnicity, and age to ensure that placement decisions and progression through mathematics courses is determined by ability alone (Hope, Adamson, McManus, Chis, & Elder, 2018; Mattern & Packman, 2009; Schmeiser, 1995). When bias is evidenced on a test, respondents with equal underlying abilities receive different scores. Thus, the interpretations made using these test scores are unreliable for the population under study (Bauer, 2017; Hope et al., 2018; Lee & Kim, 2017; O'Neill & McPeek, 1993).

Examining bias is important because the items could actually be valid and reliable questions with scores denoting real, substantive differences between various groups (e.g., males and females). Conversely, the questions may actually be biased relative to various item characteristics, and changes in the question content and/or properties may need to be explored to achieve accurate measurement and eventual equitable outcomes. Differential Item Functioning (DIF) as one analytical strategy can help explain any sex differences when responding to mathematics items so that appropriate psychometric interventions

can be proposed. Specific to the educational institution in this study, the high school admits approximately fifty percent males and females each year. A thorough investigation of the mathematics placement test for potential biases is important to ensure that the exam is fair, and the placement decisions are accurate for both males and females. Therefore, the aim of this study was to examine the item parameters (i.e., item difficulty, and item discrimination) and DIF of the mathematics placement test using Item Response Theory's Two-Parameter Logistic (2PL) Model.

Methods

The following sections describe the methods used to examine the item parameters and DIF of the mathematics placement test.

Context

The data in the current study are from one high school campus for academically gifted students in the state of Illinois. Per the mission statement of this institution, it strives to be a teaching and learning laboratory that enrolls academically talented Illinois students (i.e., Grades 10 through 12) in its advanced, residential college preparatory program with an emphasis in the fields of science and mathematics.

In order to attend, students are required to submit an admissions application which includes an essay describing the student's interest in STEM, two letters of recommendation, middle school and/or high school transcripts, and current SAT (i.e., formerly known as the Scholastic Aptitude Test or the Scholastic Assessment Test) scores. As such, the admissions process is highly competitive as students from around the state of Illinois vie for approximately 250 positions each year.

For those students that are invited to attend, the high school provides a diverse and challenging curriculum designed to prepare students for college. Not only does the curriculum include the core subjects of English, history, social sciences, science, and mathematics, but students can also choose to take a course in the fine arts, wellness, or one of the six world languages offered. Additionally, students are provided the opportunity to conduct original and compelling research with expert scholars and scientists at more than 100 institutions. As a result, students graduating are well-rounded individuals equipped with the personal, social, and academic skills needed to succeed in college and beyond.

Participants and Procedure

Existing data from four cohorts of students were used in this study. These cohorts included students entering the high school their sophomore year, beginning in the 2014/2015 academic year and ending in the most recent 2017/2018 academic year for which data was available.

Equivalence across the four cohorts was examined for five demographic variables using Chi-Square (χ^2) Tests of Association and One-Way Analyses of Variance (ANOVAs). Chi-Square Tests of Association were conducted across the four cohorts for the variables of sex and race/ethnicity. There were no significant differences in the proportions between cohort year and either sex or race/ethnicity. For the three remaining variables of socioeconomic status (i.e., median family income), incoming SAT Math (SAT_M) subscores, and incoming SAT Evidence Based Reading and Writing (SAT_ERW) subscores, ANOVAs were used. Again, there were no significant

differences between cohort years for each of the three variables. Therefore, all four cohorts were found to be statistically equivalent on the demographic variables noted above and were combined into one sample for further analysis.

Measure

Mathematics faculty members developed the mathematics placement test in 1985. The original and continuing purpose of the mathematics placement test is to determine a student's incoming mathematical knowledge for appropriate initial course placement commensurate with ability level. Thus, generally speaking, the placement test assesses mathematical knowledge needed prior to entering into a Calculus sequence. More specifically, the developers of the exam created a two-part test measuring three content areas of mathematics, namely Algebra 1, PreCalculus, and Geometry, as previously determined through an Exploratory Factor Analysis (Manuscript 2).

Part I of the assessment mainly measures student's knowledge of Algebra 1 content such as simplifying expressions, functions, and exponents. Students are given 45 minutes to complete 50 short-answer items, without a calculator. Assessing higher-level abilities such as the ability to solve numerical problems and/or to manipulate mathematical symbols and equations necessitates a short-answer question format (Nitko & Brookhart, 2011). While the short-answer format allows students to show their work, the legibility of students' responses can at times complicate the scoring process.
The mathematics faculty members using an answer key for dichotomous scoring (i.e., "Correct" or "Incorrect") grade all responses. If a grader is unsure of a student's written response, other graders are consulted. In the event that a student's response cannot be

determined, it is marked as an incorrect response. The possible range of scores on Part I is from 0 to 50. After the allotted time has expired for Part I, exam proctors collect any remaining exams and distribute Part II.

The main focus of Part II of the assessment is to measure students' knowledge of both PreCalculus and Geometry content. For this portion, students have 85 minutes to complete 57 multiple-choice items, again without a calculator. The multiple-choice format used on this portion of the test provides students with the correct answer, three distractor answers, and a fifth response option of "I don't know." Although not explicitly written on the test instructions, mathematics faculty members and exam proctors emphasize the use of the "I don't know" option. By purposefully mentioning this, it is believed that students will not guess, but rather consider using the "I don't know" response option so that they do not accidentally place into a higher course than academically appropriate. A similar argument was made by Prieto and Delgado (1999) who noted that educational standards should not be influenced by desired psychometric properties of a test. Said another way, if students are unsure of an answer, it seems more appropriate for them to omit the item rather than encouraging them to guess. After the exam is complete, the multiple-choice items are scanned into a grading software program using a scantron reader where all items are scored dichotomously (i.e., "Correct" or "Incorrect"), even if the student selected the "I don't know" option.

As the multiple-choice section had a fifth response option of "I don't know," the data were coded in such a way as to distinguish between incorrect answers and missing data. More specifically, the coding format was as follows: "1" for a correct response, "0"

for an incorrect response, "DK" for selecting the "I don't know" option on the multiple-choice section, and "M" for a missing response (i.e., an item that was left blank). The response frequencies for each item are displayed in Table 5 in the results section below. Prior to analysis, all responses of "I don't know" were recoded as an incorrect response "0" to align with the grading procedures implemented by the mathematics faculty members. The possible range of scores is from 0 to 57 on Part II.

Data Analysis

Item Response Theory (IRT) uses a collection of mathematical equations to analyze item-level data which provides information about the differences among individuals on a given construct or latent variable (De Ayala, 2009; Edelen & Reeve, 2007; Hays et al., 2000; Stone & Zhang, 2003). In order to do so, IRT assumes that the underlying latent trait (e.g., mathematical knowledge) is considered to be continuous in nature and can be represented by assigning numerical values to observed variables.

Item analysis. Three item analyses using the Birnbaum (1968) Two-Parameter Logistic Model (2PL), which makes use of the marginal maximum likelihood estimation method, were conducted to examine the characteristics of the items on each factor (i.e., Algebra 1, PreCalculus, and Geometry) of the mathematics placement test (Bock & Aitkin, 1981; Cai et al., 2011; Manuscript 2). The 2PL model includes that the probability of a correct response is both a function of the distance between the person and the item and the ability of the item to differentiate among individuals with varying levels of the latent trait (De Ayala, 2009; Edelen & Reeve, 2007; Hays et al., 2000). Thus, the

2PL model is the ordinary logistic regression of the observed dichotomous responses on the unobservable person location and item characterizations (De Ayala, 2009).

Moreover, this model was selected for the additional discrimination parameter (i.e., compared to the 1PL model), which in this study differentiates between various levels of mathematics proficiency. Although the use of the *c* parameter for guessing may apply to these data as well (i.e., as used in the 3PL model), students most likely refrained from guessing by using the optional fifth response of "I don't know" on the multiple-choice items. Thus, it was determined that the 1PL (i.e., Rasch) model was too simplistic and that the 3PL model included an additional parameter that may not be relevant considering the context and response options on the exam in this study.

Difficulty and discrimination indices can provide useful information at the item level; however, both the individual item fit and the overall model-data fit should be examined. In order to assess the item fit and the model-data fit obtained in the 2PL model, this study examined the item-level diagnostic statistics (i.e., $S - \chi^2$) developed by Orlando and Thissen (2000), the M_2 fit statistic developed by Maydeu-Olivares and Joe (2005), and the Root-Mean-Square-Error of Approximation (RMSEA) by Steiger and Lind (1980).

Additionally, the item and total test information curves were examined. The total test information curve is the sum of the item information curves and specifies how much information an instrument provides to separate two respondents with differing abilities in order to reduce the uncertainty about a person's location (De Ayala, 2009). When the

peak of the total test information curve is centered around zero (i.e., the mean), the test is said to target the average ability of the construct of interest.

Moreover, examining the total test information curve and the location of its peak can help direct the design of an instrument to be able to measure along a wide or narrow range of the continuum by adding (or removing) items located within the range of interest (De Ayala, 2009). For example, if stakeholders are interested in providing a better person ability estimation for respondents below $\theta = .70$, then the operational range of the test could be improved by adding one or more items to the lower end of the continuum, which increases the amount of information about those individuals located at the lower end. In the context of high-stakes assessments, test developers may want to specify that the ideal total test information curve have a peak higher than the mean to assess higher proficiencies of the construct of interest.

Finally, both De Ayala (2009) and Ding and Beichner (2009) mention that when calibrating high-stakes assessments test items, reasonably accurate results are obtained when instruments contain 20 or more items and a sample size of at least 500 participants. With regards to test construction, Nunnally and Bernstein (1978) recommend five times as many subjects as items or at least 200 to 300 subjects, whichever is larger. In the current study, there were a total of 107 items and approximately 300 students in each of the four cohorts. Thus the approximate total population of 1,200 students was greater than the recommendations by De Ayala (2009), Ding and Beichner (2009), and Nunnally and Bernstein (1978).

Differential item functioning. The item analyses also included an examination of Differential Item Functioning (DIF) to determine whether or not a particular item is biased with regards to respondents' reported sex (i.e., males versus females). To identify which items, if any, exhibit DIF, Wald Chi-Square (χ^2) tests with accurate item parameter error variance-covariance matrices were used (Cai, 2008; Cai et al., 2011; Lord, 1977). The null hypothesis for this test states that there are no group differences in the item parameter estimates. Therefore, if an item presents evidence of DIF (i.e., $p < .05$), further investigation is needed to warrant discarding or revising the item.

Results

Based on prior research (i.e., Manuscripts 1 and 2), the mathematics placement test is comprised of three factors – Algebra 1, PreCalculus, and Geometry. Therefore, to meet the unidimensionality assumption of the 2PL model, each factor was examined independently. The results presented below are in the order in which they were conducted.

Between 2014 and 2017, 1,125 total students took the mathematics placement exam (see Table 5). The low frequency of missing data is an indication of the higher-stakes of this assessment where students are motivated to answer all questions.

Table 5

Item Response Frequencies for the Mathematics Placement Exam by Factor

	Incorrect		Correct		I Don't Know		Missing	
Algebra 1	*n*	%	*n*	%	*n*	%	*n*	%
MC3	292	25.96	708	62.93	117	10.40	8	0.71

MC4	279	24.80	678	60.27	158	14.04	10	0.89
MC5	125	11.11	924	82.13	73	6.49	3	0.27
MC6	411	36.53	432	38.40	275	24.44	7	0.62
MC7	228	20.27	686	60.98	208	18.49	3	0.27
MC8	228	20.27	774	68.80	120	10.67	3	0.27
MC9	110	9.78	941	83.64	71	6.31	3	0.27
MC16	86	7.64	860	76.44	177	15.73	2	0.18
FR1	54	4.80	1070	95.11	N/A	N/A	1	0.09
FR3	333	29.60	773	68.71	N/A	N/A	19	1.69
FR4	79	7.02	1044	92.80	N/A	N/A	2	0.18
FR5	205	18.22	916	81.42	N/A	N/A	4	0.36
FR6	92	8.18	1012	89.96	N/A	N/A	21	1.87
FR7	143	12.71	972	86.40	N/A	N/A	10	0.89
FR8	188	16.71	925	82.22	N/A	N/A	12	1.07
FR9	168	14.93	933	82.93	N/A	N/A	24	2.13
FR10	164	14.58	930	82.67	N/A	N/A	31	2.76
FR11	194	17.24	928	82.49	N/A	N/A	3	0.27
FR12	123	10.93	970	86.22	N/A	N/A	32	2.84
FR13	282	25.07	811	72.09	N/A	N/A	32	2.84
FR14	56	4.98	1069	95.02	N/A	N/A	0	0.00
FR15	132	11.73	947	84.18	N/A	N/A	46	4.09
FR16	208	18.49	888	78.93	N/A	N/A	29	2.58
FR17	251	22.31	868	77.16	N/A	N/A	6	0.53
FR18	349	31.02	767	68.18	N/A	N/A	9	0.80
FR19	243	21.60	877	77.96	N/A	N/A	5	0.44
FR20	111	9.87	1006	89.42	N/A	N/A	8	0.71
FR21	181	16.09	910	80.89	N/A	N/A	34	3.02
FR22	127	11.29	960	85.33	N/A	N/A	38	3.38
FR23	140	12.44	936	83.20	N/A	N/A	49	4.36
FR24	61	5.42	1050	93.33	N/A	N/A	14	1.24
FR25	56	4.98	1062	94.40	N/A	N/A	7	0.62
FR26	149	13.24	970	86.22	N/A	N/A	6	0.53
FR27	89	7.91	1022	90.84	N/A	N/A	14	1.24
FR28	150	13.33	931	82.76	N/A	N/A	44	3.91
FR29	131	11.64	969	86.13	N/A	N/A	25	2.22
FR30	202	17.96	789	70.13	N/A	N/A	134	11.91
FR31	88	7.82	1031	91.64	N/A	N/A	6	0.53
FR32	49	4.36	1070	95.11	N/A	N/A	6	0.53
FR33	333	29.60	748	66.49	N/A	N/A	44	3.91

FR34	387	34.40	633	56.27	N/A	N/A	105	9.33
FR35	148	13.16	952	84.62	N/A	N/A	25	2.22
FR36	144	12.80	947	84.18	N/A	N/A	34	3.02
FR37	127	11.29	980	87.11	N/A	N/A	18	1.60
FR38	137	12.18	932	82.84	N/A	N/A	56	4.98
FR39	131	11.64	981	87.20	N/A	N/A	13	1.16
FR40	157	13.96	935	83.11	N/A	N/A	33	2.93
FR41	173	15.38	901	80.09	N/A	N/A	51	4.53
FR42	49	4.36	1068	94.93	N/A	N/A	8	0.71
FR43	410	36.44	672	59.73	N/A	N/A	43	3.82
FR44	108	9.60	933	82.93	N/A	N/A	84	7.47
FR45	130	11.56	948	84.27	N/A	N/A	47	4.18
FR47	51	4.53	988	87.82	N/A	N/A	86	7.64
FR48	112	9.96	962	85.51	N/A	N/A	51	4.53
FR49	422	37.51	634	56.36	N/A	N/A	69	6.13
FR50	183	16.27	893	79.38	N/A	N/A	49	4.36

	Incorrect		Correct		I Don't Know		Missing	
PreCalculus	*n*	%	*n*	%	*n*	%	*n*	%
MC1	421	37.42	563	50.04	138	12.27	3	0.27
MC2	666	59.20	339	30.13	117	10.40	3	0.27
MC10	422	37.51	530	47.11	169	15.02	4	0.36
MC11	323	28.71	367	32.62	432	38.40	3	0.27
MC12	375	33.33	305	27.11	439	39.02	6	0.53
MC13	169	15.02	629	55.91	316	28.09	11	0.98
MC14	233	20.71	622	55.29	262	23.29	8	0.71
MC15	537	47.73	244	21.69	337	29.96	7	0.62
MC17	214	19.02	324	28.80	565	50.22	22	1.96
MC18	151	13.42	637	56.62	331	29.42	6	0.53
MC19	192	17.07	589	52.36	335	29.78	9	0.80
MC20	148	13.16	673	59.82	302	26.84	2	0.18
MC21	399	35.47	349	31.02	372	33.07	5	0.44
MC22	275	24.44	205	18.22	632	56.18	13	1.16
MC23	156	13.87	466	41.42	497	44.18	6	0.53
MC24	189	16.80	361	32.09	561	49.87	14	1.24
MC26	415	36.89	460	40.89	248	22.04	2	0.18
MC27	322	28.62	344	30.58	453	40.27	6	0.53
MC28	141	12.53	556	49.42	417	37.07	11	0.98
MC29	148	13.16	320	28.44	645	57.33	12	1.07

MC30	145	12.89	339	30.13	629	55.91	12	1.07
MC31	349	31.02	418	37.16	352	31.29	6	0.53
MC32	237	21.07	392	34.84	488	43.38	8	0.71
MC33	482	42.84	214	19.02	422	37.51	7	0.62
MC34	319	28.36	248	22.04	548	48.71	10	0.89
MC35	117	10.40	199	17.69	796	70.76	13	1.16
MC36	269	23.91	226	20.09	614	54.58	16	1.42
MC37	245	21.78	172	15.29	688	61.16	20	1.78
MC38	110	9.78	629	55.91	374	33.24	12	1.07
MC39	353	31.38	382	33.96	365	32.44	25	2.22
MC40	186	16.53	213	18.93	712	63.29	14	1.24
MC41	176	15.64	275	24.44	655	58.22	19	1.69
MC42	82	7.29	179	15.91	848	75.38	16	1.42
MC43	208	18.49	394	35.02	504	44.80	19	1.69
MC44	561	49.87	317	28.18	238	21.16	9	0.80
MC45	139	12.36	193	17.16	774	68.80	19	1.69
FR2	87	7.73	1035	92.00	N/A	N/A	3	0.27

	Incorrect		Correct		I Don't Know		Missing	
Geometry	*n*	%	*n*	%	*n*	%	*n*	%
MC25	160	14.22	843	74.93	119	10.58	3	0.27
MC51	222	19.73	842	74.84	56	4.98	5	0.44
MC52	160	14.22	934	83.02	26	2.31	5	0.44
MC53	154	13.69	901	80.09	65	5.78	5	0.44
MC54	655	58.22	450	40.00	13	1.16	7	0.62
MC55	354	31.47	723	64.27	41	3.64	7	0.62
MC56	146	12.98	946	84.09	25	2.22	8	0.71
MC57	34	3.02	1069	95.02	15	1.33	7	0.62
MC58	310	27.56	589	52.36	202	17.96	24	2.13
MC59	262	23.29	655	58.22	184	16.36	24	2.13
MC60	232	20.62	809	71.91	64	5.69	20	1.78
MC61	367	32.62	689	61.24	55	4.89	14	1.24
MC62	326	28.98	690	61.33	19	1.69	90	8.00
FR46	150	13.33	944	83.91	N/A	N/A	31	2.76

Algebra 1

The following sections and paragraphs include the results for the Algebra 1 factor of the mathematics placement test including: (1) Item Analysis (i.e., assumptions, item difficulty and discrimination, item and model fit, and test information), and (2) Differential Item Functioning (DIF).

Item analysis. The assumption of local dependence (LD) for dichotomous items was analyzed using the Standardized LD χ^2 statistic developed by Chen and Thissen (1997). Overall, there were a total of 30 item-pairs with LD χ^2 values greater than ten. These item-pairs were further inspected for issues with the wording and/or position of the item as well as possible redundancy in the content of the items (Cai et al., 2011).

The Algebra 1 factor from the mathematics placement test has a total of 56 items. The difficulty of an item (i.e., the *b* parameter) is the point on the θ continuum that corresponds to a 50% chance of endorsing an item. The parameter estimates for item difficulty had a range of -4.70 (FR42) to 12.49 (FR11). However, Item FR11 also had a negative discrimination index (i.e., detailed in the following paragraph), and was deleted. Thus, the next largest item difficulty estimate was .50 (MC6).

Extreme and typical examples of item difficulties and their item characteristic curves (ICCs) are in Figure 6. Item FR42 (i.e., the yellow curve located at the far left-hand side) was the easiest item because the probability of a correct response is high for low ability respondents, and approaches 1 for high ability respondents near θ = -1.5. Item FR5, the orange curve, is displayed to provide a visual representation of a "typical" item response function for items on the Algebra 1 section of the exam. Finally, Item MC6

(i.e., the blue curve), located the furthest right on the horizontal axis represents the most difficult item in the Algebra 1 factor. Additionally, MC6 was the only item to have a positive difficulty estimate indicating that the Algebra 1 section on the Mathematics Placement Test is generally easy for the respondents.

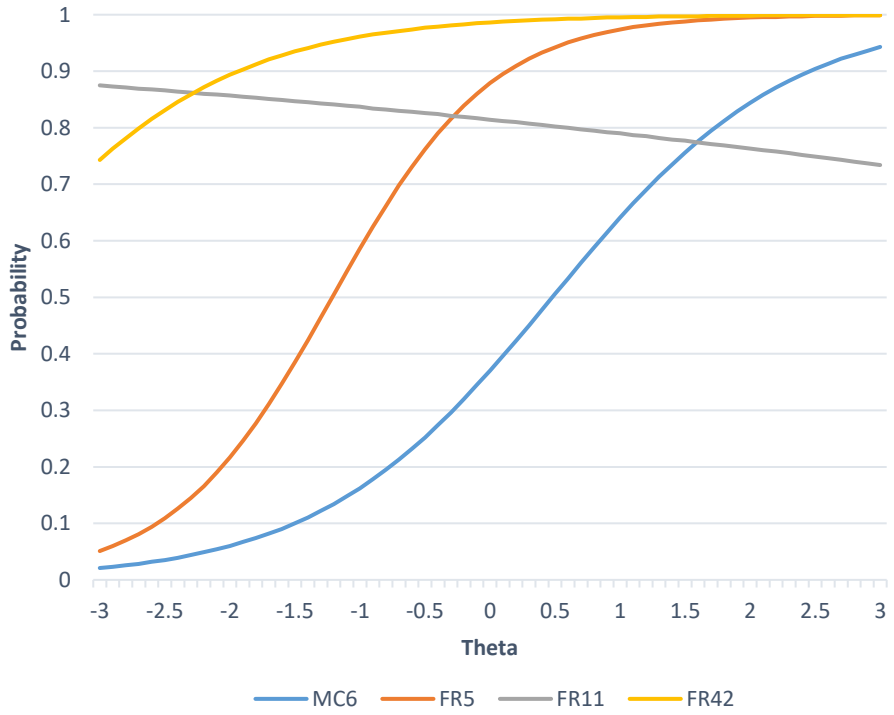

Figure 6. Algebra 1 Item Characteristic Curves. This figure shows the item characteristic curves of four select items from the Algebra 1 section of the Mathematics Placement Test.

Discrimination (i.e., the *a* parameter) is the slope of the item response function assessed at the difficulty of the item. The steeper the slope, the greater the ability of the item to differentiate between individuals with varying abilities. The parameter estimates (see Table 6 below) for the slopes (*a*) ranged from -.13 (FR11) to 4.04 (FR33). The small

negative *a* value on FR11 (i.e., the gray curve) indicated that this item is acting in a counterintuitive manner (see Figure 6 above). Specifically, individuals located further right on the θ continuum (i.e., higher proficiency in Algebra 1) were less likely to answer FR11 correctly compared to those individuals located further left on the θ continuum. Students with a stronger proficiency in Algebra 1 were more likely to answer FR11 incorrectly than those students with a weaker proficiency in Algebra 1. Thus, FR11 was identified for further revision or deletion. FR33, however, had the highest *a* parameter, indicating that item's ability to differentiate between individuals at varying levels of Algebra 1 proficiency. Finally, the slopes of three other items (i.e., FR3, FR42, and FR50) fell below the acceptable range of .8 – 2.5 (De Ayala et al., 2001), warranting a more detailed examination of these items.

Table 6

Item Parameter Estimates and Standard Errors for Algebra 1 Scale (N = 1125)

Item	Label	*a*	(s.e.)	*b*	(s.e.)
1	MC3	1.80	(.12)	-0.47	(.05)
2	MC4	2.79	(.19)	-0.32	(.04)
3	MC5	1.99	(.16)	-1.24	(.07)
4	MC6	1.20	(.10)	**0.50**	(.07)
5	MC7	2.07	(.14)	-0.37	(.05)
6	MC8	3.19	(.23)	-0.58	(.04)
7	MC9	1.69	(.14)	-1.43	(.08)
8	MC16	2.74	(.20)	-0.88	(.05)
9	FR1	1.04	(.17)	-3.32	(.42)
10	FR3	**0.75**	(.08)	-1.25	(.15)
11	FR4	1.12	(.15)	-2.74	(.28)
12	FR5	1.63	(.13)	-1.33	(.08)
13	FR6	1.76	(.18)	-1.91	(.12)
14	FR7	2.11	(.18)	-1.47	(.07)

15	FR8	1.72	(.14)	-1.35	(.08)
16	FR9	1.64	(.14)	-1.44	(.09)
17	FR10	2.38	(.20)	-1.25	(.06)
18	FR11	**-0.13**	(.08)	**12.49**	(8.36)
19	FR12	1.22	(.13)	-2.06	(.17)
20	FR13	2.00	(.15)	-0.82	(.06)
21	FR14	1.02	(.16)	-3.31	(.42)
22	FR15	1.65	(.15)	-1.63	(.10)
23	FR16	1.38	(.12)	-1.34	(.10)
24	FR17	2.56	(.19)	-0.93	(.05)
25	FR18	2.59	(.18)	-0.59	(.05)
26	FR19	1.08	(.10)	-1.45	(.12)
27	FR20	1.78	(.17)	-1.81	(.11)
28	FR21	2.21	(.18)	-1.20	(.06)
29	FR22	1.55	(.15)	-1.73	(.12)
30	FR23	1.75	(.16)	-1.50	(.09)
31	FR24	1.67	(.20)	-2.30	(.18)
32	FR25	1.30	(.18)	-2.77	(.27)
33	FR26	2.43	(.21)	-1.37	(.06)
34	FR27	1.76	(.18)	-1.95	(.13)
35	FR28	2.37	(.20)	-1.28	(.06)
36	FR29	2.95	(.27)	-1.32	(.06)
37	FR30	1.82	(.15)	-0.97	(.07)
38	FR31	1.30	(.15)	-2.37	(.20)
39	FR32	1.61	(.21)	-2.54	(.21)
40	FR33	4.04	(.32)	-0.50	(.04)
41	FR34	2.68	(.20)	-0.24	(.05)
42	FR35	1.32	(.13)	-1.78	(.13)
43	FR36	1.45	(.14)	-1.68	(.12)
44	FR37	1.55	(.15)	-1.77	(.12)
45	FR38	1.60	(.15)	-1.59	(.11)
46	FR39	1.35	(.14)	-1.92	(.14)
47	FR40	1.53	(.14)	-1.57	(.10)
48	FR41	1.44	(.13)	-1.47	(.10)
49	FR42	**0.70**	(.16)	**-4.70**	(.96)
50	FR43	1.56	(.12)	-0.41	(.06)
51	FR44	1.85	(.18)	-1.58	(.10)
52	FR45	1.19	(.13)	-2.01	(.17)

53	FR47	1.01	(.17)	-3.23	(.46)
54	FR48	0.94	(.12)	-2.57	(.28)
55	FR49	1.94	(.14)	-0.26	(.05)
56	FR50	**0.58**	(.09)	-2.86	(.43)

Next, the item-level diagnostics using $S - \chi^2$ (Orlando & Thissen, 2000) were examined to identify items misfitting to the overall model. Six items were statistically significant ($p < .05$ for all) and were further investigated. To measure the overall model-data fit, the M_2 fit statistic was used, which is asymptotically equal to χ^2 (Maydeu-Olivares & Joe, 2005). The value of the M_2 fit statistic suggested that there was not a good fit between the model and the data. However, the RMSEA was .02, which is below the acceptable threshold for good model fit (Browne & Cudeck, 1992; Maydeu-Olivares, 2013; Maydeu-Olivares & Joe, 2014; Steiger, 2016). Therefore, it was determined that the model sufficiently represented the data.

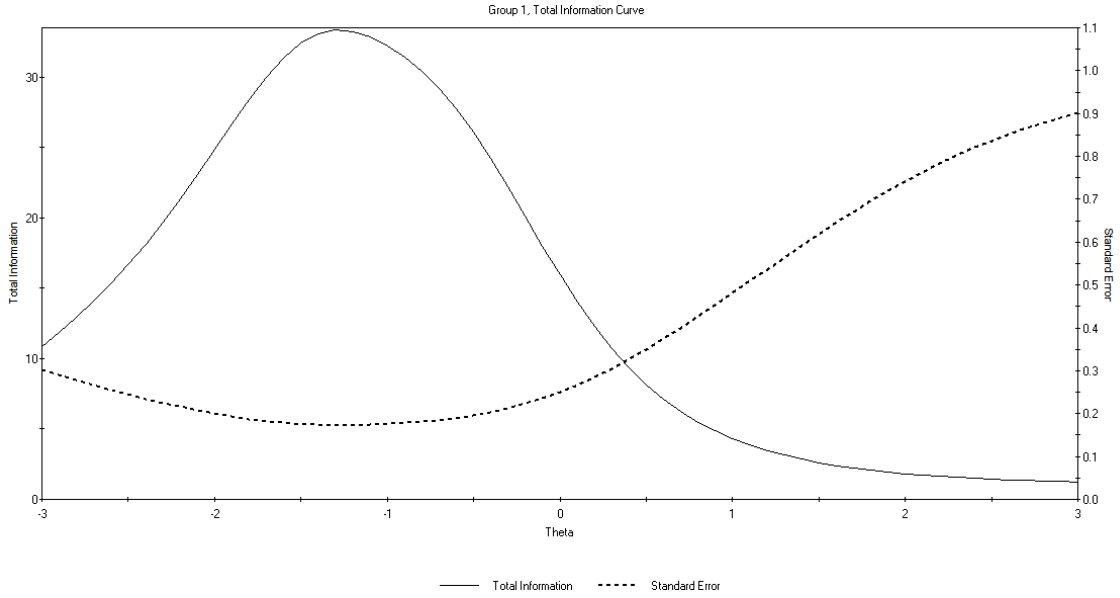

Figure 7. Algebra 1 Total Information Curve. The above figure displays the total test information function which is the sum of the item information functions across all items, which is also graphed with the standard error curve.

Finally, the Total Information Curve above, demonstrated that the maximum information value for the entire Algebra 1 test was 33 (θ = -1.30), which means that more information from the test is below the mean. Therefore, this portion of the test assessed lower levels of the underlying construct, Algebra 1 proficiency, and was not able to distinguish between varying proficiencies along the Algebra 1 continuum.

Differential item functioning. Following the item analysis, DIF was conducted with the 56 Algebra 1 items to identify if any were biased with regards to respondents' reported sex (i.e., males versus females). Nine-hundred thirty-three students had their sex recorded in the data file. Of that total, there were 469 males and 464 females. The range of discrimination and difficulty indices was similar for both males and females. For the

item-level diagnostic statistic ($S - \chi^2$) in each group, the males had three items (FR31, FR42, FR43) that did not fit the model as expected. On the other hand, the females had seven items (MC5, FR12, FR15, FR25, FR31, FR45, FR48) that did not fit the model as expected. Additionally, each group had a few item-pairs potentially violating the local dependence assumption. Overall, using the χ^2 omnibus test (Cai, 2008) and other χ^2 tests for each parameter, it was determined that two items, FR4 and FR14, exhibited DIF ($p <$ .05 for both). Thus, these two items were investigated further for either revision or elimination.

PreCalculus

Similar to the Algebra 1 results above, the following sections and paragraphs include the results of the PreCalculus factor of the mathematics placement test including Item Analysis and Differential Item Functioning (DIF).

Item analysis. The assumption of local dependence (LD) was analyzed for the second factor, PreCalculus, using the Standardized LD χ^2 statistic (Chen & Thissen, 1997). There were a total of 10 item-pairs with LD χ^2 values greater than ten. Further examination of these item-pairs is described in the discussion section below.

The PreCalculus factor from the mathematics placement exam has a total of 37 items. The parameter estimates for item difficulty ranged from -5.86 (FR2) to 1.31 (MC42). Generally speaking, the PreCalculus factor appeared to have a good amount of variability among the item difficulty values (see Table 7 below), representing a moderately difficult section. Additionally, the discrimination parameter estimates ranged from 0.43 (FR2) to 3.90 (MC35). With the exception of two items, FR2 and MC2, all

other items had discrimination indices greater than 1, demonstrating their ability to adequately differentiate between individuals at varying levels of PreCalculus proficiency.

Table 7

Item Parameter Estimates and Standard Errors for PreCalculus Scale (N = 1125)

Item	Label	*a*	(s.e.)	*b*	(s.e.)
1	MC1	1.02	(.14)	-0.02	(.26)
2	MC2	**0.71**	(.10)	1.29	(.22)
3	MC10	1.99	(.26)	0.07	(.26)
4	MC11	1.33	(.20)	0.71	(.19)
5	MC12	2.17	(.36)	0.75	(.16)
6	MC13	2.60	(.34)	-0.21	(.29)
7	MC14	1.22	(.17)	-0.25	(.29)
8	MC15	1.55	(.29)	1.15	(.11)
9	MC17	2.30	(.39)	0.65	(.17)
10	MC18	2.93	(.38)	-0.22	(.29)
11	MC19	1.94	(.28)	-0.11	(.27)
12	MC20	3.45	(.45)	-0.29	(.30)
13	MC21	2.63	(.47)	0.57	(.18)
14	MC22	2.00	(.40)	1.17	(.10)
15	MC23	2.34	(.37)	0.25	(.23)
16	MC24	1.58	(.28)	0.65	(.16)
17	MC26	2.48	(.42)	0.27	(.23)
18	MC27	2.67	(.52)	0.58	(.18)
19	MC28	2.08	(.33)	-0.01	(.26)
20	MC29	3.26	(.73)	0.61	(.16)
21	MC30	3.20	(.69)	0.56	(.17)
22	MC31	2.43	(.46)	0.38	(.21)
23	MC32	2.46	(.48)	0.45	(.20)
24	MC33	1.81	(.42)	1.19	(.08)
25	MC34	2.14	(.49)	0.97	(.11)
26	MC35	**3.90**	(1.17)	0.98	(.10)
27	MC36	2.66	(.68)	0.97	(.10)
28	MC37	3.30	(.95)	1.12	(.07)
29	MC38	1.39	(.23)	-0.28	(.28)

30	MC39	1.34	(.28)	0.61	(.15)
31	MC40	1.95	(.49)	1.14	(.08)
32	MC41	2.15	(.48)	0.85	(.12)
33	MC42	1.92	(.47)	**1.31**	(.07)
34	MC43	1.33	(.28)	0.58	(.15)
35	MC44	1.38	(.30)	0.90	(.11)
36	MC45	1.76	(.46)	1.30	(.08)
37	FR2	**0.43**	(.17)	**-5.86**	(2.18)

Extreme and typical examples of item difficulties and their ICCs are in Figure 8. Item FR2 (i.e., the grey curve located towards the top of the graph) was the easiest item because the probability of a correct response is high for low ability respondents, and approaches 1 for high ability respondents. Item MC35, the blue curve, is displayed to provide a visual representation of the item's ability to discriminate among respondents with varying ability levels, as evidenced by the steepness of the ICC. Lastly, Item MC42 (i.e., the orange curve), located the furthest right on the horizontal axis represents the most difficult item in the PreCalculus factor.

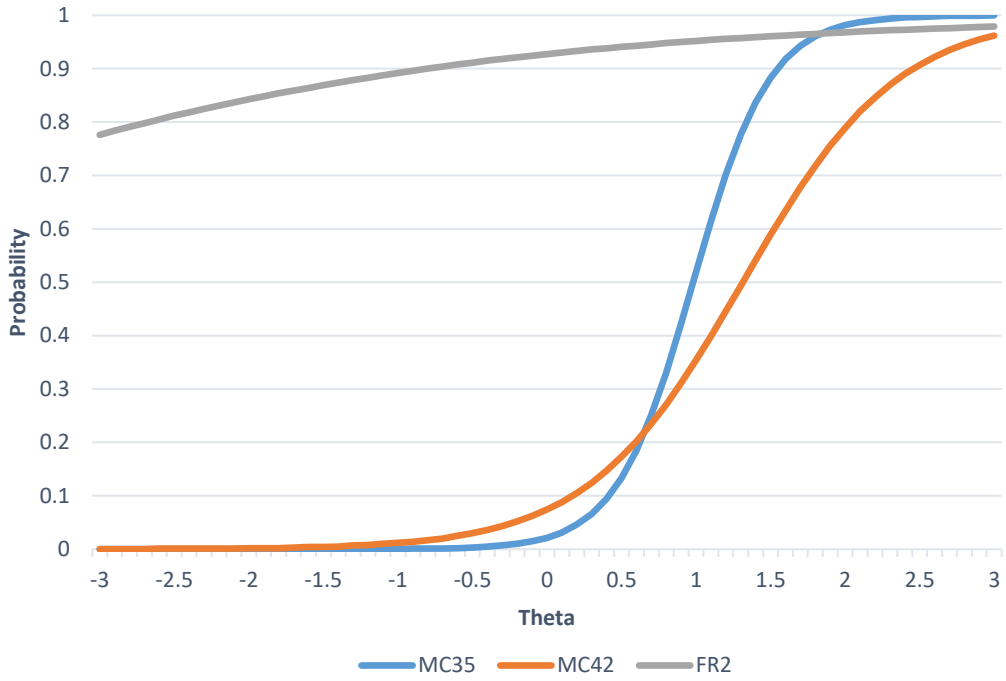

Figure 8. PreCalculus Item Characteristic Curves. This figure shows the item characteristic curves of three select items from the PreCalculus section of the Mathematics Placement Test.

All PreCalculus items were examined for item-model fit using the item-level diagnostic statistic S – χ^2 (Orlando & Thissen, 2000). Six items were statistically significant ($p < .05$) and thus did not fit the overall model as expected. These items were further investigated. In regards to the overall model-data fit, the M_2 fit statistic indicated that there was not a good fit between the model and the data. However, the RMSEA was .05, which is considered to be an acceptable model fit (Browne & Cudeck, 1992; Maydeu-Olivares, 2013; Maydeu-Olivares & Joe, 2014; Steiger, 2016). Thus, it was determined that the model provided a sufficient representation of the model.

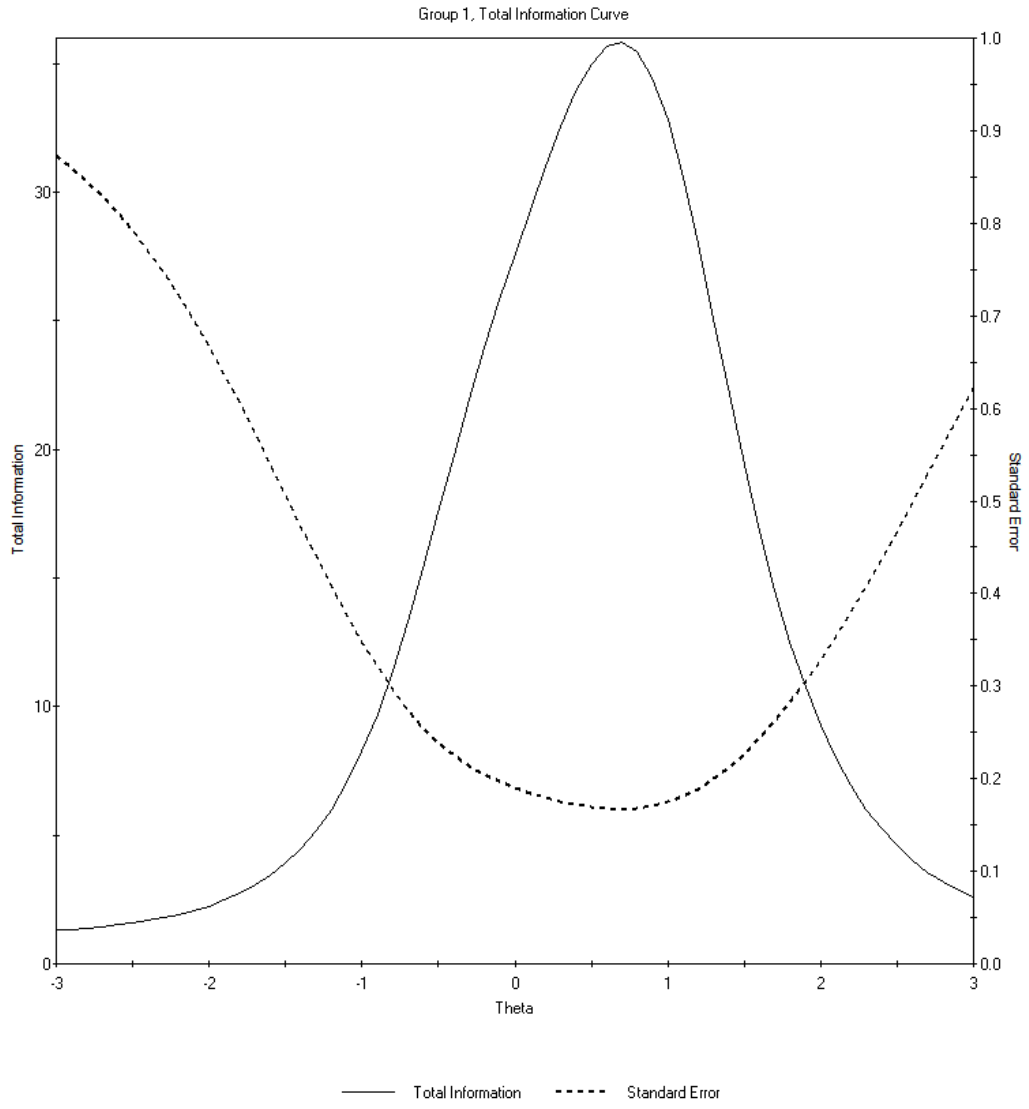

Figure 9. PreCalculus Total Information Curve. The above figure displays the total test information function which is the sum of the item information functions across all items, which is also graphed with the standard error curve.

Lastly, the Total Information Curve above, shows that the maximum information value for the PreCalculus section was approximately 34 ($\theta = 0.60$), meaning that information from the test is a little above average. Therefore, this section of the

mathematics placement test assessed higher levels of PreCalculus proficiency and was sufficiently able to distinguish between varying proficiencies along the PreCalculus continuum.

Differential item functioning. Item biases were explored on the basis of respondents' reported sex for each of the 37 PreCalculus items. The range of difficulty and discrimination indices was similar for both males and females. While the item-level diagnostic statistic ($S - \chi^2$) revealed four misfitting items for males (MC2, MC10, MC31, and FR2), all items demonstrated acceptable model fit for females. Moreover, each group had less than a handful of item-pairs potentially violating the assumption of local dependence. Finally, the χ^2 omnibus test (Cai, 2008) and additional χ^2 tests for each parameter indicated four items that exhibited DIF (MC12, MC23, MC31, and MC36). With regard to item difficulty, items MC12 and MC23 were easier for males than females. Conversely, item MC36 favored females over males. The last item, MC31, discriminated between males and females differently depending on whether or not the individual's ability level was above or below $\theta = 0.20$. In each of these situations, items were further investigated for either revision or elimination.

Geometry

Lastly, the following sections and paragraphs include the results for the Geometry factor of the mathematics placement test including Item Analysis and DIF.

Item analysis. The third factor of the mathematics placement exam, Geometry, has 14 items. The assumption of local dependence was tested and found to be tenable, indicating that each item is measuring a distinct Geometry concept and contributing to

the exam. Next, the parameter estimates for item difficulty and discrimination were analyzed.

Item difficulty values ranged from -2.84 (FR46) to 0.51 (MC54). In Table 8 below, it can be seen that 13 of the 14 total items had a negative difficulty estimate meaning that the Geometry section is generally easy for those completing this exam. Moreover, the parameter estimates for discrimination ranged from .70 (FR46) to 2.67 (MC53). Item FR46 was the only item to fall below the recommended values of discrimination, warranting a more detailed examination of this item (De Ayala et al., 2001).

Table 8

Item Parameter Estimates and Standard Errors for Geometry Scale (N = 1125)

Item	Label	*a*	(s.e.)	*b*	(s.e.)
1	MC25	0.96	(.11)	-1.36	(.14)
2	MC51	2.14	(.20)	-0.86	(.06)
3	MC52	1.29	(.13)	-1.61	(.13)
4	MC53	**2.67**	(.29)	-1.02	(.06)
5	MC54	0.93	(.10)	**0.51**	(.09)
6	MC55	1.55	(.14)	-0.56	(.06)
7	MC56	0.84	(.11)	-2.29	(.26)
8	MC57	1.96	(.26)	-2.33	(.17)
9	MC58	1.14	(.11)	-0.16	(.06)
10	MC59	1.38	(.13)	-0.38	(.06)
11	MC60	0.89	(.10)	-1.32	(.14)
12	MC61	0.91	(.10)	-0.64	(.09)
13	MC62	0.81	(.10)	-0.99	(.13)
14	FR46	**0.70**	(.11)	**-2.84**	(.41)

Extreme and typical examples of item difficulties and their ICCs are in Figure 10. Item FR46 (i.e., the grey curve located at the far left-hand side) was the easiest item because the probability of a correct response is high for low ability respondents, and approaches 1 for high ability respondents above $\theta = 1$. Item MC53, the blue curve, is displayed to provide a visual representation of the item's ability to discriminate among respondents with varying ability levels, as evidenced by the steepness of the ICC. Lastly, Item MC54 (i.e., the orange curve), located the furthest right on the horizontal axis represents the most difficult item in the Geometry factor. Additionally, Item MC54 was the only item to have a positive difficulty estimate, again, indicating that the Geometry section on the Mathematics Placement Test is generally easy for the respondents.

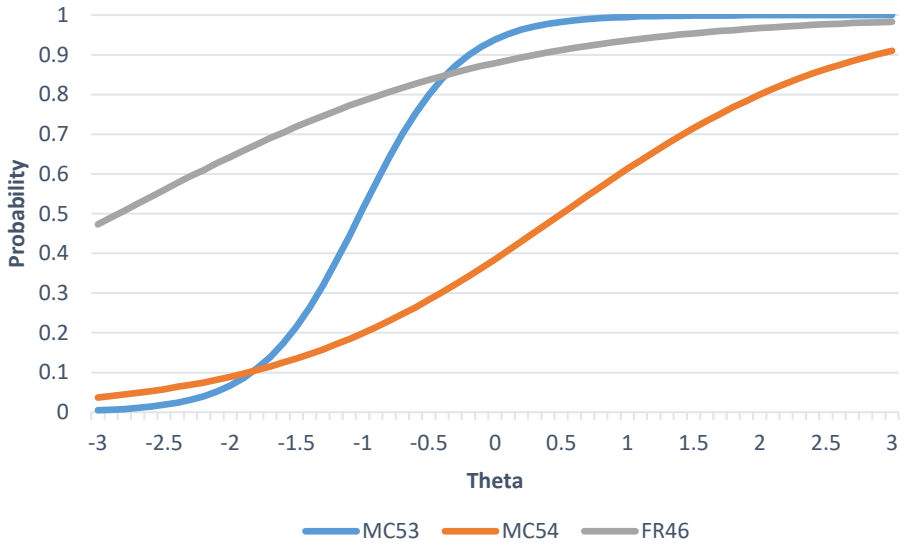

Figure 10. Geometry Item Characteristic Curves. This figure shows the item characteristic curves of three select items from the Geometry section of the Mathematics Placement Test.

Next, each item was examined for model fit using the item-level diagnostic statistic S – χ^2 (Orlando & Thissen, 2000). Only one item, MC57, was found to be statistically significant ($p < .05$) and did not fit the model as expected. A more detailed description of Item MC57 is provided in the discussion section below. Similar to previous factors, the M_2 fit statistic indicated a poor model-data fit. However, the RMSEA was .02, which was well below the acceptable level of good model fit. Therefore, it was determined that the model provided a sufficient representation of the data.

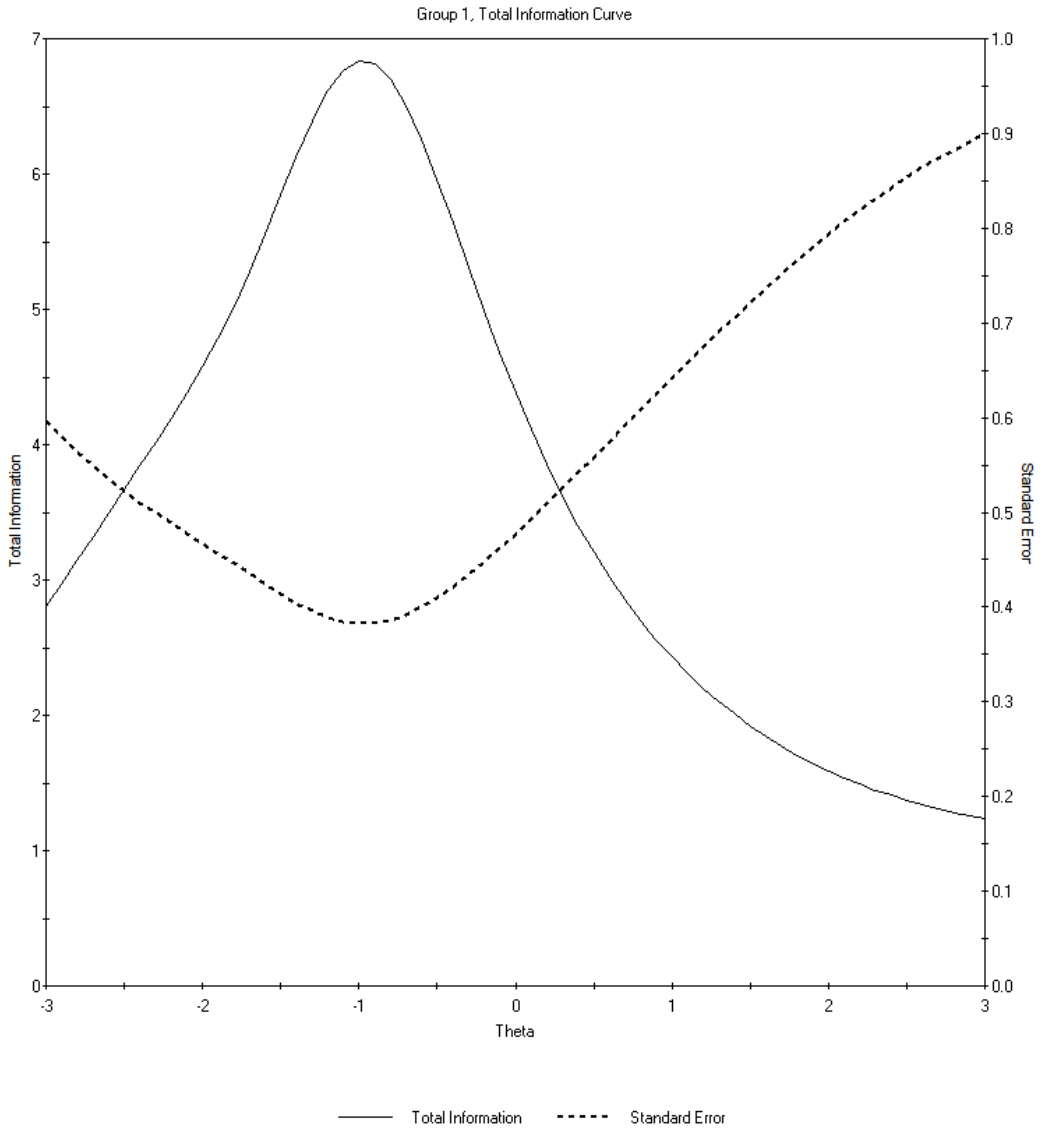

Figure 11. Geometry Total Information Curve. The above figure displays the total test information function which is the sum of the item information functions across all items, which is also graphed with the standard error curve.

Finally, the Total Information Curve (see Figure 11 above) demonstrates a maximum information value of approximately 6.7 (θ = -0.80). This value indicates that

more information from the test is slightly below the mean. Thus, this section of the test assessed lower levels of Geometry proficiency and was not able to distinguish between varying proficiencies along the Geometry continuum.

Differential item functioning. Each of the 14 Geometry items were tested for potential item bias with regards to respondents' reported sex (i.e., males versus females). Both males and females had similar parameter estimates for both difficulty and discrimination and no concerns of violating the local dependence assumption. Using the χ^2 omnibus test (Cai, 2008) and other χ^2 tests for individual parameters, it was determined that two items, MC25 and MC59, exhibited DIF ($p < .05$ for both). Item MC59 demonstrated uniform DIF such that the item was easier for females than males across the θ continuum. Item MC25, on the other hand, differed in its ability to discriminate between males and females depending on whether or not an individual was located above or below $\theta = -0.80$. Both items were investigated further for either revision or elimination.

Discussion

Educational institutions, at all levels, must be prepared to address questions about the uses and interpretations of tests and their scoring methods. To do so, it is imperative that the test itself be evaluated to ensure that the items are well constructed, unambiguous, and free of bias (Adedoyin & Mokobi, 2013; R. F. Burton, 2005; Sireci, 1998b). Once the quality of the test has been analyzed and professionals are confident in the characteristics of the test and scores, then stakeholders can be assured that the outcomes of the assessment do not lead to uneven or unfair treatment of students,

allowing for more accurate inferences to be made. Using IRT, this study examined the item parameters (i.e., item difficulty, and item discrimination) and DIF of the mathematics placement test used at a gifted, STEM, residential high school using the Two-Parameter Logistic (2PL) Model (see Table 9 below). The following sections and paragraphs provide a detailed discussion for each factor (i.e., Algebra 1, PreCalculus, and Geometry) of the mathematics placement test as well as the implications, limitations, and future directions for this study.

Algebra 1

Results from this study indicate that the Algebra 1 section of the mathematics placement test is generally easy for the population of interest suggesting that some revisions be made. As mentioned previously, Item FR11 had a negative discrimination value and was acting in a counterintuitive manner. As such, Item FR11 was recommended for deletion.

Moreover, the 30 item-pairs with possible threats of local dependence were examined further. Based on the value of the Standardized LD χ^2 statistic and the investigation of content similarity among item-pairs, eight additional items (MC8, FR16, FR21, FR26, FR30, FR31, FR33, FR37) were recommended for deletion. An additional two items, FR4 and FR42, may also be considered for deletion. Not only did Item FR4 exhibit DIF, but the (S - χ^2) item-level diagnostic statistic also suggested that FR4 did not fit the model as expected. The second item, FR42 according to the item parameter estimates, was the easiest item (b = -4.70) and also indicated poor item-model fit. After

removing the items listed above, the Algebra 1 factor had an internal consistency reliability (KR-20) of .895 for 45 items compared to the previous .91 for 56 items.

Finally thirteen items (FR1, FR8, FR9, FR12, FR19, FR25, FR32, FR35, FR36, FR38, FR39, FR45, and FR48) are recommended for revision due to their limited contribution of information as determined by their item response functions. By revising or removing items contributing little to no information to the overall Algebra 1 section of the test, the operational range of the exam can be improved. Likewise, to provide a better estimation of ability above -1.30, more items could be added to the higher end of the continuum to expand the operational range of the Algebra 1 section of the mathematics placement test.

PreCalculus

Results regarding the PreCalculus items indicate that this section is moderately challenging for the population of interest. As previously mentioned, items FR2 and MC2 had discrimination indices that fell below the accepted value of .80 (De Ayala et al., 2001). More specifically, Item FR2 was the easiest of the PreCalculus items (b = -5.86) and did not fit the model as expected. Item MC2, although it did not exhibit DIF, the item characteristic curve suggests that this item tends to be easier for males than females. For these reasons, it is recommended that item FR2 be deleted and MC2 be revised for future administrations of this assessment.

Furthermore, the 10 item-pairs with potential threats to the assumption of local dependence were examined along with misfitting items. From these procedures, it was determined that item MC31 did not fit the model as expected and did share similar

content with another item. Thus, item MC31 should be removed. After removing these two items, the new internal consistency reliability estimate (KR-20) remained consistent at .95.

Lastly, it is recommended that 11 additional items be discussed further due to the high frequency of selecting the fifth response option "I don't know." More specifically, item MC35 was previously identified as misfitting the model. Upon additional examination, it was determined that approximately 71% of the respondents answering item MC35 had selected the "I don't know" response option. Use and relevance of this item in placing students in their first mathematics course in the high school should be discussed.

Geometry

The Total Information Curve along with difficulty parameter estimates suggests that the Geometry section of the mathematics placement test is generally easy for the population of interest. Moreover, it is recommended that four items (FR46, MC25, MC57, and MC59) be considered for revision. Item FR46 had a smaller than acceptable discrimination index and appears to be contributing little information to the overall Geometry section according to the item information function. Items MC25 and MC59 exhibited DIF and therefore need to be examined to avoid item bias. As previously stated, item MC57 did not fit the model as expected. After reviewing the item's content, it is believed that one of the distractor options may be contributing additional confusion on item MC57. Thus, it is recommended that item MC57 be discussed further and potentially revised.

One final point to consider is the possibility of removing all 14 Geometry items from the overarching mathematics placement exam. Although it is interesting to know how students perform on Geometry concepts, these items are not used for placement purposes. In order to graduate high school (i.e., in Illinois), each student must complete a high school level Geometry course. However, a vast majority of the gifted students attending the high school of interest complete their required Geometry course prior to acceptance. Therefore, incoming students are only "placed" into Geometry if they have not yet completed the state requirement. As such, it may be advisable to remove the Geometry items from the placement test in exchange for other items that may assist with a more accurate placement of students.

Table 9

Summary of Item Analysis Results

	Algebra 1 (56 items)		PreCalculus (37 items)	Geometry (14 items)
Difficulty	[-4.70, 12.49]	[-4.70, .50]*	[-5.86, 1.31]	[-2.84, .51]
Discrimination	[-.13, 4.04]	[.58, 4.04]*	[.43, 3.90]	[.70, 2.67]
DIF	FR4 and FR14		MC12, MC23, MC31, and MC36	MC25 and MC59
# Items Deleted	11		2	0
# Items Remaining	45		35	14
KR-20	0.895		0.95	0.736

Note. *Item FR11 excluded

Differential Item Functioning

Findings from this study suggest that some gender-based differential item functioning exists on each of the three sections (i.e., Algebra 1, PreCalculus, and Geometry) of the mathematics placement test. While the items with the short-answer format exhibited less DIF than the multiple-choice items, the cause of gender differences in performance on certain items remains unclear.

Previous research has indicated that males have a stronger advantage than females on items using the multiple-choice format (Becker, 1990; Burton, 1996; Garner & Engelhard Jr., 1999). However, results from this study were mixed. Across the PreCalculus and Geometry sections, there were a total of six multiple-choice items that exhibited DIF. Three of those items (i.e., MC12, MC23, and MC25) favored males over females while the remaining three items (i.e., MC31, MC36, and MC59) revealed a distinct advantage for females compared to males. Future research may consider examining the patterns in the choices of distractors made by students who got the item wrong. Such patterns may provide additional insight and explanation of the observed gender differences.

Moreover, the two short-answer items (i.e., FR4 and FR14) that exhibited DIF on the Algebra 1 section of the mathematics placement test demonstrated an advantage for males over females. This result was surprising as previous research has supported the argument that females tend to have an advantage on Algebra items compared to males (Abedalaziz, Leng, & Alahmadi, 2018; Altenhof, 1984; Burton, 1996; Doolittle & Cleary, 1987; Garner & Engelhard Jr., 1999). Additionally, in the current study, the

maximum information value for males on the Algebra 1 section was approximately 38.78 (θ = -0.8) compared to a maximum information value for females of approximately 38.88 (θ = -1.2). Although the amount of information is virtually the same, the location at which the peak occurs is much different. Thus, these findings suggest that the Algebra 1 section of the mathematics placement test was easier for females compared to males, supporting the findings from previous research.

Implications

The purpose of the current study was to examine the item parameters (i.e., item difficulty, and item discrimination) and DIF of the mathematics placement test used at a gifted STEM residential high school. By critically examining the quality of the items on the mathematics placement test, all stakeholders can be assured that the inferences drawn from the educational assessment are accurate and that the assessment outcomes do not lead to unfair or uneven treatment of students (Harris, 2003; Linn, 1994).

Findings have practical implications for the faculty members at the high school in this study as they consider future revisions and administrations of the mathematics placement test. Study results suggested that eleven items should be removed from the Algebra 1 section of the mathematics placement test, with an additional two items recommended for deletion from the PreCalculus section, due to concerns of local dependence, item difficulty, and item discrimination. Additionally, there were a few items that exhibited DIF and should be discussed further to identify why the item was biased on the basis of students' sex. Thus, by equipping faculty members with these

important item-level details, they can more confidently customize the mathematics placement test to accurately place students in their initial mathematics course.

Moreover, this study provides an initial step in demonstrating the need to critically examine the psychometric properties of placement tests at all educational levels. Although the average high school may not have adequate resources to conduct similar research, there is still a need to have solid and defensible placement tests and practices to ensure that decisions are equal and fair for all students. The current study may act as a catalyst for similar high schools to examine the placement tests in use at their institutions.

Limitations and Future Research

One major limitation of this study is the use of the "I don't know" response option on the multiple-choice section of the mathematics placement test. Since the early 1970s, researchers and statisticians alike have continued to argue the advantages and disadvantages of offering such a response option. Some claim that the "I don't know" response option may be informative and thus should be included within the estimation model (Balcombe & Fraser, 2011). Others propose that the "I don't know" option is not suitable for tests measuring respondent's optimal performance and that to either discourage guessing and/or to encourage "I don't know" responses is to seek reliability at the cost of validity (Mondak, 2001).

Some test developers and administrators will advocate for the use of the "I don't know" option as a way to reduce guessing behaviors. A compromise for this was proposed by Zhang (2013) who noted that if it is the intention of the test to minimize guessing and measure precise knowledge, then the "I don't know" option could be used

within a penalty scoring model. Another suggestion to address the use of the "I don't know" option was to eliminate the "I don't know" response on multiple-choice questions by using a post-hoc correction (Kline, 1986; Mondak, 2001). In this post-hoc correction, the "I don't know" responses are randomly assigned to the remaining four choices, essentially entering guesses on behalf of the respondents who would not do so themselves (Mondak, 2001). However, since the goal of the mathematics placement test is to measure optimal performance, the post-hoc correction or a penalty scoring model seem inappropriate due to the differences in individuals' willingness to guess.

When students vary in their willingness to guess, then two students with the same ability level will receive different scores (Culbertson, 2011; Hanna, 1974; Mondak, 2001; Pohl et al., 2014). In this instance, the test is no longer measuring only knowledge of mathematics, but also students' "test-wiseness." Again, if the intention of the placement test is to measure students' maximum performance in mathematics, then all possible sources of measurement error should be reduced to ensure the proper course placement. Future research could examine the various correction models discussed above to determine which, if any, may be best suited for the purposes of this mathematics placement exam.

In the current study, the Two-Parameter Logistic Model (2PL) was used to examine the characteristics of the items on each factor of the mathematics placement test because it was believed that the presence of the "I don't know" response option prevented students from guessing. However, if the "I don't know" response option is removed from the exam, future research could use the 3PL model to re-examine the item parameters

(i.e., item difficulty, item discrimination, and guessing) and DIF of the mathematics placement test.

As noted in the previous section, there are a number of items that have been recommended for revision or deletion. Future research can support these efforts to ensure the use of quality items that adequately measure the construct of interest. Finally, more research is warranted to examine additional factors (i.e., race/ethnicity, socioeconomic status) that may elicit item bias so that stakeholders can be confident that the decisions and interpretations made based off of the scores obtained are equitable across all groups and identities.

Conclusions

While the use of placement tests is a near-universal practice at the post-secondary level, fewer studies have focused their attention on the psychometric properties of these tests. It is imperative that educational institutions at all levels examine their placement testing procedures and assessments to demonstrate their impact on students' future educational outcomes (Mattern & Packman, 2009; McDaniel et al., 2007; Morgan & Michaelides, 2005; Norman et al., 2011). Maintaining a cooperative research partnership between content experts and assessment professionals provides an opportunity to address issues throughout the item development, revision, and piloting process while simultaneously enhancing the visibility of measurement and evaluation. This study encourages similar schools with an emphasis on STEM and/or gifted education to develop relationships with measurement professionals who can provide valuable insight regarding the use and development of placement tests within unique educational settings.

Results from the current study indicate that the mathematics placement test is generally easy for the population of interest. While the PreCalculus items proved to be more challenging, many respondents used the "I don't know" response option for some items. Further discussion should determine whether or not the information obtained from the "I don't know" response is useful in the placement decision-making process. Moreover, it is recommended that the Algebra 1 and Geometry items be reconsidered due to concerns of local dependence, difficulty, and discrimination. Since the Geometry items are not used for placement purposes, future versions of the mathematics placement test may exclude these items in favor of other items that may be of more relevance to placement decisions. Additional conversations are also recommended regarding a few items exhibiting differential item functioning.

Educational assessments, when designed and used properly, can enhance later performance and provide feedback to both the student and other interested stakeholders on what has and has not been learned. Only then can an educational institution provide evidence of maximizing student success while minimizing the consequences of misplacement.

CHAPTER VII – MANUSCRIPT 4

PLACEMENT EXAM SCORES AND FIRST-SEMESTER MATHEMATICS ACHIEVEMENT AT A SCIENCE, TECHNOLOGY, ENGINEERING, AND MATHEMATICS (STEM) GIFTED RESIDENTIAL HIGH SCHOOL

Abstract

According to the literature, the primary purpose of placement testing is to assess students' academic skills and to provide them with instruction that is appropriate for their ability (e.g., Frisbie, 1982; Mattern & Packman, 2009; Morgan & Michaelides, 2005; Noble et al., 2003; Sawyer, 1996). As such, educational institutions, at all levels, must continually review and evaluate their placement tests and policies to ensure that students are enrolled in courses that will increase the probability of success and minimize the unintended negative consequences of misplacement (e.g., Linn, 1994; Mattern & Packman, 2009; McFate & Olmsted III, 1999; Norman et al., 2011; Wiggins, 1989).

To review the placement procedures being used at a Science, Technology, Engineering, and Mathematics (STEM) residential high school for gifted students, the current study sought evidence of the Predictive Validity of the item scores on a mathematics placement test. Existing data from two cohorts were obtained and analyzed using a series of Hierarchical Multiple Linear Regressions. Findings from this study demonstrated the ability of the mathematics placement test total and factor scores to predict students' success in their first semester mathematics course, providing evidence of Criterion-Related Validity for the population of interest.

Keywords: Predictive Validity, Multiple Regression, Mathematics Placement Test, STEM Education

Introduction

In educational measurement, constructs such as achievement, interest, and performance are assigned numerical values, through the use of a wide variety of tests and assessments, to infer the abilities and proficiencies of students. Specific to the current study, the purpose of achievement testing is to measure students' actual knowledge or acquired skills in order to reliably distinguish between students who do and do not have some level of the construct of interest (Slavin, 2007). As one of the primary measures used in educational research, there is an abundance of literature focused on achievement testing as institutions begin to defend their policies and practices surrounding the use of these measures.

At the post-secondary level, numerous articles have been published regarding the use of placement tests for incoming students. Many of these articles mention the continuing decline of academic standards, specifically in the area of mathematics (e.g., Crist et al., 2002; Hoyt & Sorensen, 2001; Medhanie et al., 2012; Ngo & Kwon, 2015; Parker, 2005; Schmitz & delMas, 1991). Unsurprisingly, the lowered academic standards in math are said to be related to students' scoring lower on mathematics placement tests. Due to the lower test scores, more students are being assigned to take remedial coursework, which has sparked a conversation about whether or not students are less prepared for college-level work or if the placement tests used are appropriate for this type of decision (Morgan & Michaelides, 2005).

More specifically, nearly one-third of all students entering community colleges take at least one remedial or developmental course in mathematics (e.g., Bailey, 2009; Hoyt & Sorensen, 2001; Kowski, 2013; Medhanie et al., 2012; Melguizo et al., 2014; Scott-Clayton, 2012). Not only do these remedial courses lower student motivation, but they also add time to student graduation. Furthermore, the additional time students spend taking non-credit courses increases their overall cost to attend and lowers retention rates (Medhanie et al., 2012; Melguizo et al., 2008; Ngo & Kwon, 2015; Scott-Clayton, 2012). Some community colleges have even been accused of placing students into these remedial, non-credit courses as a way to increase revenue (Armstrong, 2000). As a result, post-secondary institutions are now being asked to provide evidence of the effectiveness of their placement procedures and measures to ensure that the negative consequences of misplacement are minimized (Armstrong, 2000; Morgan & Michaelides, 2005; Smith & Fey, 2000). Institutions must remember that accurately placing students is a necessary, but not sufficient, condition for a placement system as a whole to be effective (Sawyer, 1996).

The purpose of this study was to provide evidence of Criterion-Related Validity (i.e., Predictive Validity) of a mathematics placement test at a Science, Technology, Engineering, and Mathematics (STEM), gifted, residential high school. Specifically, this study examined the relationship between the mathematics placement exam and students' performance in their initial mathematics course. Previous research on placement exams have been conducted at the post-secondary level; however, this study extends the research to younger grade levels serving a specific, gifted population.

Literature Review

The overarching purpose of placement tests is to enroll students in courses that are suitably challenging to their current knowledge level (e.g., Akst & Hirsch, 1991; Frisbie, 1982; Marshall & Allen, 2000; Mattern & Packman, 2009; McFate & Olmsted III, 1999; Noble et al., 2003; Sawyer, 1996). When students are not fittingly placed, their courses can either bore or frustrate them, which in turn lowers students' motivation to perform at a normal or higher level (Mattern & Packman, 2009; Morgan & Michaelides, 2005; Noble et al., 2003; Sawyer, 1996).

In addition to impacting student motivation, prior research has shown that course placement decisions can have a significant impact on a student's future academic preparation (McDaniel et al., 2007; Morgan & Michaelides, 2005). For example, students who begin in a post-secondary mathematics course that is appropriate given their background have an increased chance of succeeding in their initial mathematics course and their subsequent mathematics courses (Akst & Hirsch, 1991; Latterell & Regal, 2003; Marshall & Allen, 2000; Mattern & Packman, 2009; Norman et al., 2011; Shaw, 1997). For this reason, more research is needed to thoroughly examine placement tests and procedures to ensure that students are in fact being placed into courses that will maximize the probability of their success (Linn, 1994; Mattern & Packman, 2009; McFate & Olmsted III, 1999; Norman et al., 2011; Wiggins, 1989). Although these placement tests are typically considered "high-stakes," the psychometric properties of such tests have received relatively little attention and need to be evaluated further (Callahan, 2005; Grubb & Worthen, 1999; Scott-Clayton, 2012).

According to the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2005), test developers are charged with the responsibility to: (1) Provide evidence of what the test measures, its recommended uses, and its strengths and limitations, and (2) Provide evidence that the technical quality (i.e., reliability and validity) of the test meets its intended uses. Moreover, previous research has recommended that colleges and universities consider the rigor and defensibility of the policies and methods used to inform placement decisions due to their "high-stakes" classification (Clark & Watson, 1995; Morgan & Michaelides, 2005). Armstrong (1995) stated that both Title V and Federal Civil Rights legislation requires institutions to validate the use of assessment tests in the placement and referral of students. Therefore, regardless of educational level, future research should continue to evaluate and specify the psychometric properties of placement tests in order to address questions about the impact of these tests on students and their learning.

Criterion-Related Validity draws inferences from individuals' exam scores to performance on some external criterion of practical importance (Crocker & Algina, 2008; Hambleton et al., 1978). This type of validity can be evidenced either concurrently or predictively. Procedures for concurrent validation are used when the data collected for both the test and the criterion occur at or about the same point in time (Crocker & Algina, 2008; Wiersma & Jurs, 2009). On the other hand, procedures for predictive validity require a gap in time between when the test was given and when the criterion data are collected (Crocker & Algina, 2008). Additionally, the purpose of predictive validity is to

determine whether or not test scores have the ability to predict specified future performance.

In the context of educational measurement and placement decisions, the best indicator of future behavior/performance is an individual's past behavior/performance (Belfield & Crosta, 2012; Davis & Shih, 2007; Erwin & Worrell, 2012; Feldhusen & Jarwan, 1995). However, one of the major concerns detailed in the existing literature base has been the disparity between the cognition and performance elicited on placement tests and the cognition and performance needed to succeed in the classroom (Armstrong, 2000; Brown & Niemi, 2007; Madison et al., 2015; Marsh et al., 2007; Schmitz & delMas, 1991). For example, if a test forbids the use of a calculator, the score obtained from that test may not accurately predict a student's ability to succeed in a mathematics course that encourages the use of calculators (Akst & Hirsch, 1991). Moreover, Predictive (i.e., Criterion-Related) Validity is enhanced when the correspondence between what is measured on a test is congruent with what is needed to succeed in a course (Armstrong, 2000).

Prior research has attempted to examine this relationship by investigating the Predictive Validity of post-secondary placement exams in relation to course grade received. Within these models, the use of multiple measures is encouraged and provides more accurate course placement decisions compared to test scores alone (e.g., Armstrong, 1995; Erwin & Worrell, 2012; Marwick, 2004; Ngo & Kwon, 2015; Noble et al., 2003). For example, one study showed that combining the SAT Mathematics exam with either high school GPA (i.e., grade point average) and/or class rank was a better predictor of

college achievement over test scores alone (Schumacher & Smith, 2008). However, other studies have cautioned that the usefulness of the SAT Mathematics exam is limited due to the average difference in scores between males and females (Bridgeman & Wendler, 1989, 1991; Davis & Shih, 2007; Gallagher & De Lisi, 1994). More recent research has concluded that the accuracy of placement decisions greatly increases when placement test scores are combined with measures of high school achievement (i.e., high school GPA, high school grades, courses taken; Marwick, 2002; Melguizo et al., 2014; Ngo & Kwon, 2015; Pike, 1991; Scott-Clayton, 2012; Wattenbarger & McLeod, 1989). Although the use of multiple measures have been demonstrated to enhance placement policies and decisions at the post-secondary level, additional research is sought after at the high school level.

Therefore, the current study sought evidence of Criterion-Related Validity (i.e., Predictive Validity) of the scores on a mathematics placement test used at a gifted residential high school for students interested in STEM using a series of Hierarchical Multiple Linear Regressions. These regressions were used to investigate the relationship between students' mathematical knowledge, as measured by the mathematics placement test, and students' subsequent performance, as measured by their grade (i.e., a score represented by a percentage between zero and 100) in their first semester mathematics course. Moreover, the models used included students' demographic information and previous mathematics coursework to mimic the reality of placement practices used at the high school under study and to improve the predictive accuracy of the results.

Methods

The following sections describe the methods used to examine the Criterion-Related Validity of the scores on a mathematics placement test.

Participants and Procedures

Existing data from four cohorts of students were obtained to examine Predictive Validity. These cohorts consisted of students entering the high school their sophomore year, beginning in the 2014/2015 academic year and ending in the most recent 2017/2018 academic year, for which data was available. However, due to incomplete and inaccessible data, the final analysis included two of the four cohorts for which the most complete data were available.

Additionally, group equivalence across the two cohorts was examined and reported for the population information listed above (e.g., gender and race/ethnicity) using Chi-Square Tests of Association. Furthermore, the two cohort means of students' median family incomes (SES), incoming SAT Mathematics scores (SAT_M), and the SAT Evidence-Based Reading and Writing (SAT_ERW) scores were examined for significant differences using Independent Samples t-Tests. No significant differences were identified for four of the five demographic variables (i.e., gender, race/ethnicity, SAT_ERW, and SES). The demographic variable of SAT_M, showed significant differences between the two cohorts ($t_{[539]} = 2.394$, $p < .05$). The cohort from 2014 (n = 257) had a mean SAT_M score of 689.22 (SD = 71.43) compared to the cohort of 2016 (n = 284) which had a mean SAT_M score of 673.45 (SD = 80.85).

To further examine this difference, Cohen's d was calculated as a measure of effect size. An effect size is an indicator of the degree of departure between the null hypothesis (i.e., equivalent means) and the alternate hypothesis (i.e., group means differ), such that a small effect size is .2, medium is .5 and large is .8 (Cohen, 1988). In the current study, the effect size was small (d = .2). Therefore, even though there was a statistically significant difference between the two cohorts on the SAT_M variable, the small effect size justified combining the two cohorts into one sample for subsequent data analysis.

Measure

Mathematics faculty members developed the mathematics placement test in 1985. The original and continuing purpose of the mathematics placement test is to determine a student's incoming mathematical knowledge for appropriate initial course placement commensurate with ability level. Thus, generally speaking, the placement test assesses mathematical knowledge needed prior to entering into a Calculus sequence. More specifically, the developers of the exam created a two-part test measuring three content areas of mathematics, namely Algebra 1, Geometry, and PreCalculus, as previously determined through an Exploratory Factor Analysis (Manuscript 2).

In Manuscript 3, an item analysis was conducted to examine the item parameters (i.e., item difficulties and item discrimination indices) and differential item functioning within each factor. As a result of the study, some items were deleted from the exam. The Algebra 1 factor had a KR-20 reliability estimate of .895 for 45 items and measured student's knowledge of content such as simplifying expressions, functions, and

exponents. The Geometry factor had the lowest reliability estimate (KR-20 = .736) and the fewest number of items (n = 14). These items assessed concepts such as right triangle trigonometry, properties of congruent angles and triangles, and characteristics of a circle. Finally, the PreCalculus factor had a KR-20 reliability estimate of .95 for 35 items and measured student's knowledge of content such as evaluating and graphing quadratic and exponential functions, finding the roots of functions, laws of sines and cosines, and combinatorics. Students' performance on the exam is noted by a raw subscore for each factor (i.e., Algebra 1, Geometry, and PreCalculus) and a total exam score.

Data Analysis

As part of the General Linear Model family of statistical techniques, Multiple Regression is used to explain or predict a criterion (dependent) variable with more than one predictor (independent) variable (e.g., Ebel, 1965; Hair Jr et al., 1995; Osborne, 2000; Petrocelli, 2003; Rubio et al., 2003; Stevens, 2012; Wampold & Freund, 1987). There are many types of regression analyses (i.e., Linear, Logistic, Polynomial), which is dependent upon the measurement level of the outcome variable. In the current study, the dependent variables are continuous (i.e., interval level), so a Multiple Linear Regression was used. Although it can be argued that mathematical knowledge may follow a different type of curve, a linear regression model was selected due to the limited time lapse between the start of testing and the completion of their initial mathematics course (i.e., approximately six to eight months).

Furthermore, regression analyses differ in the manner and order in which the independent variables are entered into the model (e.g., simultaneously, stepwise,

hierarchically). Hierarchical entry in Multiple Regression allows the researcher to select the order of the entered predictor variables based on previous research and/or theory. When Hierarchical entry is used, the focus is on the change in predictability that is associated with the variables entered later in the analysis, above and beyond the contribution of the previously entered control variables (Petrocelli, 2003). Thus, Hierarchical Multiple Linear Regression was used in the current study to control for a series of conceptually-similar variable groupings prior to the main variables of interest – the mathematics placement exam scores for the high school.

Outlier detection. Prior to conducting each multiple regression analysis, data were examined for potential influential data points, leverage points, and/or outliers. The presence of influential data points can significantly affect the overall analysis. An influential data point is one where if deleted, it would produce a substantial change in the value of at least one regression coefficient (Stevens, 2012). To detect influential data points, Cook's distance (Cook, 1977) and DFBETAS (Hahs-Vaughn, 2016; Stevens, 2012) were used. Cook's distance (Cook, 1977) measures the amount of change in the regression coefficients that would occur if a particular case was omitted. Typically, if Cook's D > 1, it is determined that there is an influential data point.

While Cook's D is a composite measure of influence, the DFBETAS indicate which specific coefficients are most influential by providing information on the change in the predicted value when a specific case is deleted from the model (Hahs-Vaughn, 2016; Stevens, 2012). Thus, when any DFBETA value is outside the range of ±2, this indicates a sizeable change and should be examined further.

Next, the predictor variables were investigated for possible outliers using leverage values and Mahalanobis distances. Leverage values are used to quickly identify participants that differ from the rest of the sample on a particular set of predictor variables (Stevens, 2012). The current study used the calculation of $\frac{3p}{n}$, where p is the number of predictors plus 1 and n is the sample size, suggested by Stevens (2012) and adapted from Hoaglin and Welsch (1978). In this case, if a leverage value was greater than or equal to $\frac{3p}{n}$, then the data point was examined further.

Additionally, Mahalanobis distances were used to measure how far each case was from the mean of the independent variable for the remaining cases (Hahs-Vaughn, 2016; Stevens, 2012). To determine whether or not a large enough difference existed, which would indicate a possible outlier, the χ^2 distribution table was used to find the critical value for either 9 or 11 predictor variables ($\alpha = .001$). If the Mahalanobis distance exceeded the critical value, the case was further investigated.

Finally, to find outliers on the criterion variable (y), this study examined the standardized residuals (r_i). Standardized residuals allow the researcher to identify subjects whose predicted score is different from the actual criterion score (Stevens, 2012). Generally speaking, standardized residuals follow a normal distribution with approximately 95% of the standardized residual values falling within two standard deviations of the mean (Stevens, 2012). For the current analysis, all data points were examined to ensure that no more than 5% of the cases fell outside the acceptable range of $r_i < |2|$ and did not need to be further examined (Hair Jr et al., 1995; Stevens, 2012).

Assumptions. After detecting influential data points, leverage points, and/or outliers, the statistical assumptions of regression were examined and addressed. These assumptions included Multicollinearity, Independence of Errors (i.e., Residuals), Linearity, Normality, and Homoscedasticity (Hahs-Vaughn, 2016; Hair Jr et al., 1995; Stevens, 2012).

Multicollinearity exists when there is a strong correlation between some or all of the independent variables (Hair Jr et al., 1995; Stevens, 2012; Wampold & Freund, 1987). If present, multicollinearity reduces the unique explained variance of each predictor variable while increasing the shared prediction, complicating the interpretation of a predictor variable (Hair Jr et al., 1995; Stevens, 2012). To test multicollinearity, the tolerance, variance inflation factors (VIF), and collinearity diagnostics were examined.

Tolerance is measured as 1 minus the proportion of variance explained in the variable of interest by the other predictor variables (Hair Jr et al., 1995). Thus, a lower tolerance value (i.e., less than .10) suggests that the variable of interest is accounted for by the other variables, suggesting possible multicollinearity problems (Hahs-Vaughn, 2016). By taking the reciprocal of tolerance, the VIF is produced and values greater than 10 are indicative of threats to multicollinearity (Hair Jr et al., 1995).

Lastly, the eigenvalues of the collinearity diagnostics were examined. When multiple eigenvalues are close to zero, this indicates that some independent variables have strong intercorrelations and may present concerns of multicollinearity (Hahs-Vaughn, 2016). In this case, the condition index can be calculated using the square root of the ratio between the largest eigenvalue to each preceding eigenvalue, to ensure that no

values exceed 10 (Hahs-Vaughn, 2016). If multicollinearity is suspected in any of the above situations, it is recommended that either one or more of the highly correlated variables be eliminated from the model or consolidated into a single measure.

The next assumption, Independence of Errors (i.e., residuals), assumes that each participant's responses are not dependent upon the response of another individual (Stevens, 2012). If violated, it is possible to identify variables as statistically significant, when in fact they are not (Keith, 2014). In the current study, each student completed their placement exam under the supervision of an exam proctor, implying that the assumption of independence is tenable. Furthermore, the assumption of independence of errors was examined by plotting the studentized residuals against the unstandardized predicted values.

The third assumption of Linearity describes the degree to which a change in the criterion variable associated with the predictor variable is constant across the range of values for the predictor variable (Hair Jr et al., 1995; Keith, 2014). Using partial regression plots, each predictor variable was examined with the criterion variable for the presence of a linear relationship.

The next assumption, Normality, requires that each continuous variable (i.e., independent and dependent) follow a normal distribution of data (Hair Jr et al., 1995; Stevens, 2012). Normality was checked by creating and examining both a histogram of unstandardized residual values in relation to the normal distribution curve and normal probability plots, generally referred to as Q-Q Plots (Hair Jr et al., 1995; Keith, 2014). The skewness and kurtosis of the unstandardized residuals was also examined.

The final assumption, Homoscedasticity suggests the presence of equal error variances (Hair Jr et al., 1995; Keith, 2014; Stevens, 2012). Similar to previous assumptions, violation of homoscedasticity can affect the standard errors, which in turn will impact the statistical significance of variables. To test for homoscedasticity, residual plots of the predictor variables against the criterion variable were used to identify whether or not a relatively random display of points was present.

One additional consideration in this multiple regression analysis was the sample size. In the current study, an a priori power analysis was conducted in G*Power 3.1.9.4 for the "Linear Multiple Regression: Fixed Model, R^2 Deviation from Zero" (Faul et al., 2007). For the two multiple regressions using the total score from the mathematics placement test, the software tool yielded a minimum total sample size of 114 to detect a medium effect given a significance level of .05, power of .80, and nine predictor variables. Similarly, for the two multiple regressions using the three factor subscores from the mathematics placement test, the software tool yielded a minimum total sample size of 123 to detect a medium effect given a significance level of .05, power of .80, and eleven predictor variables (Cohen, 1988).

Correlations. Prior to conducting the multiple regression analyses, correlations were investigated to look at the relationship between the independent and dependent variables. Phi correlations were computed for the relationship between the variables of gender and race/ethnicity, as both are measured on a nominal (i.e., dichotomous) scale. For the case where a nominal variable was correlated with a continuous variable, Point Biserial correlations were calculated. Finally, the Pearson correlations were calculated to

examine the relationship between two continuous variables. The correlation matrix summarizing the information above is reported in the results section and significant correlations at .05, .01, and .001 are identified.

Variables. As stated previously, Hierarchical Multiple Regressions were used to explore the relationships between students' mathematical knowledge and their subsequent performance in their first semester mathematics course. In any multivariate analysis, the careful selection of variables is important for statistical conclusion validity. When selecting variables for inclusion, the final decision should be based on either theoretical or conceptual grounds (Hair Jr et al., 1995). The variables considered in this study are provided in Table 10 below.

Table 10

Hierarchical Multiple Linear Regression Model Predictors - Level of Measurement and Coding

Variable Name	Level of Measurement	Code	
(1) Demographic Covariates			
Sex	Nominal (Dichotomous)		
Male		0	
Female		1	
Race	Nominal	Race 1 (r_1)	Race 2 (r_2)
Asian		1	0
White		0	1
Other		0	0
Socioeconomic Status	Interval (Continuous)	-	
Median Family Income			
(2) Incoming Performance Covariates			
SAT Math Score	Interval (Continuous)	-	
SAT Critical Reading Score	Interval (Continuous)	-	
Algebra 1 GPA	Nominal (Dichotomous)		
3.0 or below		0	
4.0		1	
Geometry GPA	Nominal (Dichotomous)		
3.0 or below		0	
4.0		1	
Took an Algebra 2 Course	Nominal (Dichotomous)		
No		0	
Yes		1	
(3) Main Predictor Variables			
Mathematics Placement Test	Interval (Continuous)	-	
Total Score			
Algebra 1 Subscore			
Geometry Subscore			
PreCalculus Subscore			
(4) Criterion Variable			
Grade in 1st Semester Math Course	Interval (Continuous)	-	
Lower Level Math Course			
Upper Level Math Course			

Over the past two decades, numerous articles have detailed the uses, consequences, and challenges of placement exams (e.g., Denny et al., 2012; Farley, 2007; Foley-Peres & Poirier, 2008; Haeck et al., 1997; Rueda & Sokolowski, 2004; Schmitz & delMas, 1991). However, the vast majority of these studies were within the context of a community college or university. Thus, the predictor variables chosen for inclusion in the current study were from similar studies containing varying contexts.

In the current study, the first block of the Hierarchical Multiple Regression included student demographic information such as sex, race/ethnicity, and socioeconomic status (SES). A variety of studies have been conducted examining demographic variables and their impact on educational outcomes, specifically math achievement. For example, in a study by Roth et al. (2000), racial differences in mathematics achievement did not exist after controlling for previous coursework in mathematics. Another study mentioned that regardless of racial group, SES was unrelated to gender differences in mathematics achievement or attitudes (Catsambis, 1994). Moreover, Pugh and Lowther (2004) found that regardless of students' race, SES, or type of high school, the greatest indicator of college achievement was the mathematics course(s) taken.

Conversely, additional research has demonstrated SES, especially income, to be an important predictor in mathematics achievement and career decisions, especially for females (Gonzalez & Kuenzi, 2012; Oakes, 1990). Moreover, research has shown that Black and Hispanic students are less than half as likely to be in gifted education programs compared to White students (Callahan, 2005). The same study also concluded that nine percent of students enrolled in gifted and talented programs were categorized in the

bottom quartile of family income (Callahan, 2005). Other studies have concluded that both SES and race/ethnicity strongly correlate with academic performance and account for a significant amount of variance in students' test scores (Sirin, 2005; White et al., 2016). Although the nature of the impact of race/ethnicity and SES on educational achievement is ongoing, these variables have not been considered in the context of a gifted residential high school focused on STEM.

The second block in the regression analysis contained incoming academic information including students' SAT mathematics subscore, SAT Evidence-Based Reading and Writing subscore, students' grades in previous coursework (i.e., Algebra 1 and Geometry) and whether or not the student had reached an Algebra 2 level course. In a study by Sheel et al. (2001), high school GPA, SAT mathematics score, and the student's final grade received in high school Algebra 2 were the most influential predictors of students' college mathematics placement test scores. Similarly, Latterell and Regal (2003) found that other predictors such as high school courses and the grades received in those courses were often stronger predictors of college course success than an incoming placement test score. These variables are similar to others in previous studies, but the context was at the post-secondary level rather than at a high school (Latterell & Regal, 2003; Pugh & Lowther, 2004; Sheel et al., 2001).

The third and final block of the analysis included either the total score, or the three subscores of Algebra 1, Geometry, and PreCalculus, from the high school mathematics placement. The placement test was positioned last in the Hierarchical Multiple Linear Regression as the amount of variance the placement test explains, over

and above the variables in the previous blocks, was central to addressing the research question in this study.

Finally, the criterion (i.e., outcome) variables in this study were students' percentage grades received in their first semester mathematics course, which were divided into lower and upper level courses. Based on the placement exam score, students enter into one of four mathematics courses – Mathematical Investigations I, II, III, or IV. Thus, Mathematical Investigations I and II were categorized as lower level courses with Mathematical Investigations III and IV being categorized as upper level courses. While some students begin the math sequence in either Geometry or BC Calculus I, these decisions are not determined through the use of the placement exam, and thus were not included in the study sample.

Results

The main research question in this study was, "What is the Criterion-Related Validity of the item scores on a mathematics placement test for gifted, residential high school students interested in STEM?" More specifically, this study examined the relationship between the predictor and outcome variables of the mathematics placement exam in relation to how students' perform in their initial mathematics course using Hierarchical Multiple Linear Regression. Four regression analyses were conducted, two for the lower level courses and two for the upper level courses.

Multiple Regression for Lower Level Courses

The first two Hierarchical Multiple Linear Regressions were conducted for students completing either Mathematical Investigations I or II. After all outliers and

assumptions were tested for both the total score regression and the factor score regression, it was determined that the two samples were identical. To reduce redundancy, the outlier detection, assumption, and descriptive statistics sections will only be presented once. Following that discussion, the correlation matrix and regression results for the total mathematics placement test score as a predictor is presented first followed by the regression involving the factor subscores as predictors.

Outlier detection. The two lower level mathematics courses had an initial enrollment of 234 students. Through the process of data cleaning and outlier detection, an additional seven cases were removed for a final sample of 227 students. Tables 11 and 12 below provide details regarding the outlier testing that was conducted, the acceptable values for each test, the range of values that were obtained, and the action taken as a result of each outlier check. Five of the seven cases were removed because of missing data present on one or more independent variable. The other two cases were removed as potential outliers due to the Mahalanobis Distances obtained.

Table 11

Multiple Regression Outlier Checking for Lower Level Mathematics Courses - Total Score

Measure	Recommended Value(s)	Case(s)	Obtained Value(s)	Action
Missing Data	No missing data on any IV	5	≥1 on IV(s)	Removed
Cook's Distance	Cook's D < 1	None	[0, .082]	Retain
DFBETAS	DFBETA ≤ |2|	None	[-1.627, 1.397]	Retain
Leverage Values	Leverage < $3p/n$ = .131	2	[.018, .146]	Retain
Mahalanobis Distance	Mahalanobis Distance < χ^2 (i.e., *cv* for 9 IVs and α = .001)	2	Distances > 27.877	Removed
Standardized Residuals	No more than 5% of r_i > |2|	12	Cases ≈5.29% (r_i > |2|)	Retain

Table 12

Multiple Regression Outlier Checking for Lower Level Mathematics Courses - Subscale Scores

Measure	Recommended Value(s)	Case(s)	Obtained Value(s)	Action
Missing Data	No missing data on any IV	5	≥1 on IV(s)	Removed
Cook's Distance	Cook's D < 1	None	[0, .07]	Retain
DFBETAS	DFBETA ≤ |2|	None	[-.499, .720]	Retain
Leverage Values	Leverage < $3p/n$ = .157	None	[.021, .149]	Retain
Mahalanobis Distance	Mahalanobis Distance < χ^2 (i.e., *cv* for 11 IVs and α = .001)	2	Distances > 31.264	Removed
Standardized Residuals	No more than 5% of r_i > |2|	12	Cases ≈5.29% (r_i > |2|)	Retain

Assumptions. Prior to examining the predictive ability of the mathematics placement test scores, the assumptions of multiple regression were examined. Multicollinearity was examined using values of Tolerance, Variance Inflation Factors (VIFs), and collinearity diagnostics. Tolerance values for the total score regression ranged from .372 to .931 and had VIFs between 1.075 and 2.690 indicating that all values were within acceptable limits for all predictors. Similarly, the tolerance values for the factor score regression fell between .346 and .939 with VIFs ranging from 1.065 to 2.891, again indicating that all values were within acceptable limits for all predictors. The collinearity diagnostics for both regressions, in combination with the tolerance and VIF values, suggested that there was no concern of multicollinearity.

The next two assumptions, Independence of Errors and Linearity, were both determined to be tenable. Independence of Errors was considered through the use of scatterplots comparing studentized residuals against unstandardized predicted values. As all points were within two standard deviations of the mean, this assumption was met for

both regressions. Similarly, the partial scatterplots were examined for the presence of a linear relationship between each independent variable and the dependent variable. The scatterplots displayed a linear relationship for all cases in both regressions, and thus the assumption of Linearity was met.

Normality was examined using the skewness, kurtosis, and histogram of the unstandardized residuals along with the normal probability plots for each regression. While there was evidence of a negatively skewed distribution for both the total score and factor score regressions, the values of kurtosis and information from the probability plots suggested that normality was reasonable in both cases. Therefore, the assumption of Normality was tenable.

Finally, the assumption of homoscedasticity was considered based on the scatterplots of studentized residuals versus the predicted values. The spread of residuals appeared fairly consistent over the range of values of the independent variables, providing evidence of homoscedasticity for both regressions.

Descriptive statistics. After removing cases due to missing data and potential outliers, each regression analysis had a final sample of 227 students. Of the total sample, 90 (39.6%) were male and 137 (60.4%) were female. Additionally, the sample contained 70 (30.8%) students who identified as Asian, 93 (41.0%) students who identified as White, and 64 (28.2%) students who identified as either Black or African American, Hispanic or Latino, or who reported two or more races. Student's median family income was estimated using the zip code of student's home address and ranged from $20,227 to $137,059 with an average of $71,058.54 (*SD* = 22810.21). Moreover, this sample of

students had an average SAT Math (SAT_M) score of 643.92 (*SD* = 67.18) and an average SAT Evidence-Based Reading and Writing (SAT_ERW) score of 625.51 (*SD* = 63.17). Lastly, the average total score achieved on the mathematics placement test was 46.51 (*SD* = 13.61) out of a possible score of 94. The strongest factor score was Algebra 1 with an average of 31.22 (*SD* = 9.36) out of a possible score of 45. The average Geometry and PreCalculus factor scores were much lower with means of 9.57 (*SD* = 2.66) out of 14 and 5.71 (*SD* = 4.30) out of 35, respectively.

Correlations for lower level regression. Correlations were run to examine the relationship between the independent and dependent variables (see Table 13 below). The strongest positive correlation was between SAT Math Score and the Mathematics Placement Test Total Score ($r = .685$, $p < .001$). This strong correlation indicates that students who perform well on the SAT Math exam also perform well on the mathematics placement test. Conversely, the strongest negative correlation appeared between Race 1 and Race 2 ($r_{\Phi} = -.556$, $p < .001$).

Among the independent variables, the Mathematics Placement Test Total Score had the strongest correlation with the dependent variable Percentage Grade in Initial Mathematics Course ($r = .579$, $p < .001$). That is to say that high achieving students on the mathematics placement test are also high achieving students in their initial mathematics course. On the other hand, Race 2 had the only negative correlation with the dependent variable ($r_{pb} = -.029$, $p > .05$), which was not significant.

Table 13

Summary of Correlations for Lower Level Mathematics Courses

#	Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Gender	-													
2	Race 1	.151*	-												
3	Race 2	-.149*	-.556***	-											
4	Median Family Income	.106	.115	-.076	-										
5	SAT Math Subscore	-.110	.299***	.025	.259***	-									
6	SAT ERW Subscore	.020	.120	.064	.242***	.448***	-								
7	Algebra 1 GPA	-.006	.124	-.103	-.069	-.044	-.095	-							
8	Geometry GPA	.097	.182**	.018	-.017	.190**	.118	-.098	-						
9	Algebra 2 Taken	-.013	.089	.102	.008	.100	-.034	-.113	.057	-					
10	MPT Total Score	-.042	.429***	-.102	.280***	.685***	.342***	.040	.260***	.341***	-				
11	Algebra 1 Factor Score	-.018	.434***	-.112	.287***	.675***	.342***	.052	.272***	.286***	-	-			
12	Geometry Factor Score	-.017	.207**	-.028	.159*	.510***	.415***	.025	.179**	-.068	-	.499***	-		
13	PreCalculus Factor Score	-.062	.306***	-.042	.164*	.383***	.081	-.008	.161*	.475***	-	.541***	.156*	-	
14	% Grade in IMC	.029	.266***	-.029	.215**	.492***	.304***	.044	.239***	.112	.579***	.561***	.401***	.362***	-

Note. *p < .05, **p < .01, ***p < .001. ERW = Evidence-Based Reading and Writing, GPA = Grade Point Average, MPT = Mathematics Placement Test, IMC = Initial Mathematics Course

Total score regression. A Hierarchical Multiple Linear Regression was conducted to explore the relationship between the main predictor variable of Mathematics Placement Test Total Score and the criterion variable of the percentage grade received in the student's initial lower level mathematics course. Regression results suggest that a significant proportion of the total variance in students' grades is explained by the collection of independent variables ($R^2 = .366$, $F_{[10, 216]} = 12.479$, $p < .001$). More specifically, the predictors accounted for 36.6% of the variance in the percentage grade students' received in their initial mathematics course.

Overall, there were nine predictors in this model and all three regression blocks were significant. Examining the final block of the regression model, displayed in Table 14 below, it is evident that the student's total score from the mathematics placement test was the only significant predictor of the student's percentage grade received in their initial mathematics course ($t = 5.057$, $p < .001$). Specifically, for each one-point increase in students' Mathematics Placement Test Total Score, the students' grade received in their first semester mathematics course increased by .229 percentage points. Therefore, the Mathematics Placement Test Total Score is predictive of student success in their initial lower level mathematics course (i.e., Mathematical Investigations I or II), providing evidence of Criterion-Related (i.e., Predictive) Validity.

Table 14

Hierarchical Multiple Linear Regression for Lower Level Mathematics Courses - Total Score (n = 227)

Variables	*B*	*SE*	*t*	*β*	95% CI for B	
					Lower	Upper
Gender	.579	.799	.724	.041	-.996	2.153
Race 1	-.193	1.129	-.171	-.013	-2.418	2.033
Race 2	-.084	.976	-.086	-.006	-2.008	1.839
Median Family Income	.000	.000	.592	.035	.000	.000
SAT Math Subscore	.015	.008	1.788	.147	-.002	.032
SAT_ERW Subscore	.008	.007	1.211	.075	-.005	.022
Algebra 1 GPA	1.492	1.673	.892	.050	-1.806	4.789
Geometry GPA	.857	1.173	.730	.043	-1.456	3.170
Algebra 2 Taken	-.533	.874	-.610	-.038	-2.255	1.189
Placement Test Total Score	.229	.045	5.057***	.449	.140	.318

Note. ***$p < .001$. ERW = Evidence-Based Reading and Writing, GPA = Grade Point Average, *B* = Unstandardized Regression Coefficient, *SE* = Standard Error, β = Standardized Regression Coefficient, CI = Confidence Interval

Subscale score regression. To better understand the relationship between the mathematics placement test and students' percentage grade received in their initial mathematics course, a Hierarchical Multiple Linear Regression was conducted with the three factor subscores of Algebra 1, Geometry, and PreCalculus. Regression results indicated that a significant proportion of total variance in students' grades is explained by the collection of independent variables ($R^2 = .367$, $F_{[12, 214]} = 10.335$, $p < .001$). Thus, the predictor variables accounted for 36.7% of the variance in the percentage grade students' received in their initial mathematics course.

Similar to the total score regression, all three regression blocks were significant for the eleven predictor variables. Exploring the final block of the regression model,

Table 15 below shows that the only significant predictor of the student's percentage grade received in their first mathematics course was the Algebra 1 Factor Score (t = 3.321, p = .001). Additionally, for each one-point increase in students' Mathematics Placement Test Algebra 1 Score, the students' grade received in their first semester mathematics course increased by .227 percentage points.

Table 15

Hierarchical Multiple Linear Regression for Lower Level Mathematics Courses - Subscale Scores (n = 227)

Variables	*B*	*SE*	*t*	*β*	95% CI for B	
					Lower	Upper
Gender	.568	.803	.708	.040	-1.014	2.151
Race 1	-.147	1.138	-.129	-.010	-2.389	2.096
Race 2	-.085	.981	-.086	-.006	-2.019	1.849
Median Family Income	.000	.000	.622	.037	.000	.000
SAT Math Subscore	.015	.009	1.732	.143	-.002	.032
SAT_ERW Subscore	.007	.007	1.047	.067	-.007	.021
Algebra 1 GPA	1.486	1.680	.885	.050	-1.825	4.798
Geometry GPA	.842	1.180	.714	.042	-1.483	3.168
Algebra 2 Taken	-.373	.937	-.398	-.026	-2.219	1.473
Algebra 1 Factor Score	.227	.068	3.321***	.306	.092	.361
Geometry Factor Score	.300	.179	1.676	.115	-.053	.654
PreCalculus Factor Score	.198	.117	1.701	.123	-.031	.428

Note. ***$p \leq .001$. ERW = Evidence-Based Reading and Writing, GPA = Grade Point Average, *B* = Unstandardized Regression Coefficient, *SE* = Standard Error, β = Standardized Regression Coefficient, CI = Confidence Interval

Recall that Mathematical Investigations is a four-semester sequence of courses preparing students for Calculus. According to the course syllabus, one objective of Mathematical Investigations I (i.e., the first course the sequence) is to further develop students' understanding of the underlying concepts of algebra and refine their abilities to

apply their algebraic skills. The second course (i.e., Mathematical Investigations II) builds upon this foundation and facilitates student learning in the areas such as linear relationships and equations, exponential functions, and transformations of functions. Therefore, not only does the Mathematics Placement Test Total Score predict student success in their initial lower level mathematics course (i.e., Mathematical Investigations I or II), but more specifically, the subscore obtained from the Algebra 1 section of the mathematics placement test predicts student success in an Algebra-centric course sequence, providing strong evidence of Predictive Validity.

Multiple Regression for Upper Level Courses

The final two Hierarchical Multiple Linear Regressions were conducted for students completing an upper level mathematics course (i.e., Mathematical Investigations III or IV). After all outliers and assumptions were tested for both the total score and factor score regressions, it was determined that there were minor differences between the two samples. To reduce redundancy, the outlier detection and assumption sections will only be presented once. Following that discussion, the descriptive statistics, correlation matrices, and regression results for the total score regression will be presented first, followed by the regression involving the factor subscores as predictors.

Outlier detection. The two upper level mathematics courses had an initial enrollment of 150 students. Through the data cleaning and outlier detection processes, an additional twelve cases were removed for a final sample size of 138 students. Tables 16 and 17 below provide details regarding the outlier testing that was conducted, the acceptable values for each test, the range of values that were obtained, and the action

taken as a result of each outlier check. Two of the twelve cases were immediately removed due to missing data on one or more independent variable. The other ten cases were removed as potential outliers based on the Mahalanobis Distances obtained.

Table 16

Multiple Regression Outlier Checking for Upper Level Mathematics Courses - Total Score

Measure	Recommended Value(s)	Case(s)	Obtained Value(s)	Action
Missing Data	No missing data on any IV	2	≥1 on IV(s)	Removed
Cook's Distance	Cook's D < 1	0	[0, .157]	Retain
DFBETAS	DFBETA ≤ |2|	0	[-.871, 1.158]	Retain
Leverage Values	Leverage < $3p/n$ = .203	7	[.013, .282]	Retain
Mahalanobis Distance	Mahalanobis Distance < χ^2 (i.e., *cv* for 9 IVs and α = .001)	10	Distances > 27.877	Removed
Standardized Residuals	No more than 5% of r_i > |2|	6	Cases ≈4.35% (r_i > |2|)	Retain

Table 17

Multiple Regression Outlier Checking for Upper Level Mathematics Courses - Subscale Scores

Measure	Recommended Value(s)	Case(s)	Obtained Value(s)	Action
Missing Data	No missing data on any IV	2	≥1 on IV(s)	Removed
Cook's Distance	Cook's D < 1	0	[0, .136]	Retain
DFBETAS	DFBETA ≤ |2|	0	[-.830, 1.248]	Retain
Leverage Values	Leverage < $3p/n$ = .243	3	[.021, .305]	Retain
Mahalanobis Distance	Mahalanobis Distance < χ^2 (i.e., *cv* for 11 IVs and α = .001)	10	Distances > 31.264	Removed
Standardized Residuals	No more than 5% of r_i > |2|	6	Cases ≈4.35% (r_i > |2|)	Retain

Assumptions. Following the data cleaning and outlier detection processes, the assumptions of multiple regression were examined. Multicollinearity was considered using values of Tolerance, Variance Inflation Factors (VIFs), and collinearity diagnostics. Tolerance values for the total score regression ranged from .249 to .936 and had VIFs between 1.068 and 4.021, suggesting that all values were within acceptable limits for all predictors. Similarly, the tolerance values for the factor score regression fell between .211 and .961 with VIFs ranging from 1.040 to 4.746, again suggesting that all values were within acceptable limits for all predictors. The collinearity diagnostics for both regressions, in combination with the tolerance and VIF values, indicated that there was no concern of multicollinearity.

Next, the two assumptions of Independence of Errors and Linearity were explored and determined to be tenable. Independence of Errors was examined using scatterplots comparing studentized residuals against the unstandardized predicted values. Since all points fell within two standard deviations of the mean, this assumption was met for both regressions. In a similar manner, the partial scatterplots were used to identify the presence of a linear relationship between each independent variable and the dependent variable. All scatterplots suggested that a linear relationship was evident for all variables in both regressions, demonstrating that the Linearity assumption had been met.

The fourth assumption, Normality, was investigated using the skewness, kurtosis, and histogram of the unstandardized residuals along with the normal probability plots for each regression. Even though there was evidence of a negatively skewed distribution for both the total score and factor score regressions, the values of kurtosis and information

from the probability plots indicated that normality was reasonable in both cases. Therefore, the assumption of Normality was tenable.

Lastly, homoscedasticity was examined using the scatterplots of studentized residuals versus the predicted values. The distribution of residuals appeared relatively consistent across the range of values of the independent variables, providing evidence of homoscedasticity in both regressions.

Descriptive statistics for total score regression. The final sample for the total score regression was 138 students. Of the total sample, there were 82 (59.4%) males and 56 (40.6%) females. Moreover, there were 81 (58.7%) students who identified as Asian, 47 (34.1%) students who identified as White, and 10 (7.35%) students who identified as either Black or African American, Hispanic or Latino, or who reported two or more races. Student's median family income was estimated using the zip code of the student's home address and ranged from $37,846 to $138,178 with an average of $87,772.30 (*SD* = 22662.97). Additionally, this group of students had an average SAT Math score of 735.43 (*SD* = 44.43) and an average SAT ERW score of 659.49 (*SD* = 54.76). Finally, the average total score achieved on the mathematics placement test was 72.00 (*SD* = 8.23) out of a total possible score of 94.

Descriptive statistics for subscale score regression. The final sample for the factor score regression was also 138 students, but had a minor difference among the demographic variables. In the factor score regression total sample, there were again, 82 (59.4%) males and 56 (40.6%) females. Moreover, there were 81 (58.7%) students who identified as Asian, 48 (34.8%) students who identified as White, and 9 (6.52%) students

who identified as either Black or African American, Hispanic or Latino, or who reported two or more races. Similar to before, student's median family income was estimated using the zip code of the student's home address and ranged from $38,313 to $138,178 with an average of $87,620.35 (*SD* = 22800.29). Additionally, this group of students had an average SAT Math score of 736.23 (*SD* = 43.43) and an average SAT ERW score of 659.93 (*SD* = 54.36). Finally, the largest factor score among these students was Algebra 1 with an average of 43.23 (*SD* = 2.08) out of a possible score of 45. The average Geometry and PreCalculus factor scores were similar with means of 11.00 (*SD* = 2.03) out of 14 and 18.88 (*SD* = 6.41) out of 35, respectively.

Correlations for total score regression. Correlations were run to examine the relationship between the independent and dependent variables (see Table 18 below). The strongest positive correlation was between SAT Math Score and the Mathematics Placement Test Total Score ($r = .536$, $p < .001$). This strong correlation indicates that students who perform well on the SAT Math exam also perform well on the mathematics placement test. Conversely, the strongest negative correlation appeared between Race 1 and Race 2 ($r_{\Phi} = -.857$, $p < .001$).

Among the independent variables, the Mathematics Placement Test Total Score had the strongest correlation with the dependent variable Percentage Grade in Initial Mathematics Course ($r = .478$, $p < .001$). That is to say that high achieving students on the mathematics placement test are also high achieving students in their initial mathematics course. On the other hand, Race 2 had the strongest negative correlation with the dependent variable ($r_{pb} = -.173$, $p < .05$).

Table 18

Summary of Correlations for Upper Level Mathematics Courses - Total Score Regression

#	Variable	1	2	3	4	5	6	7	8	9	10	11
1	Gender	-										
2	Race 1	.064	-									
3	Race 2	-.002	-.857***	-								
4	Median Family Income	.031	.121	-.076	-							
5	SAT Math Subscore	-.251**	.128	-.088	-.102	-						
6	SAT ERW Subscore	.074	.072	-.038	-.058	.369***	-					
7	Algebra 1 GPA	.055	-.145	.124	-.065	-.155	-.144	-				
8	Geometry GPA	-	-	-	-	-	-	-	-			
9	Algebra 2 Taken	.081	.074	-.106	-.118	-1.07	-.108	-.033	-	-		
10	MPT Total Score	-.108	.169*	-.149	.292***	.536***	.403***	-.018	-	-.217*	-	
11	% Grade in IMC	-.033	.224**	-.173*	.072	.398***	.236**	.018	-	-.129	.478***	-

Note. *p < .05, **p < .01, ***p < .001. ERW = Evidence-Based Reading and Writing, GPA = Grade Point Average, MPT = Mathematics Placement Test, IMC = Initial Mathematics Course

Correlations for subscale score regression. Similar to the total score regression, the relationship between the independent and dependent variables was explored using the correlation matrix in Table 19 below. The strongest positive correlation present was between SAT Math Score and the PreCalculus Factor Score (r = .427, $p < .001$). This strong correlation suggests that students who score high on the SAT Math exam also score high on the cumulative PreCalculus items from the mathematics placement test. Conversely, the strongest negative correlation was present between Race 1 and Race 2 (r_{Φ} = -.871, $p < .001$).

Examining the independent variables, the SAT Math Score had the strongest correlation with the criterion variable Percentage Grade in Initial Mathematics Course (r = .415, $p < .001$). In other words, students who perform well on the SAT Math exam also perform well in their first mathematics course. On the other hand, Race 2 had the strongest negative correlation with the dependent variable (r_{pb} = -.182, $p < .05$).

Table 19

Summary of Correlations for Upper Level Mathematics Courses - Subscale Score Regression

#	Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
1	Gender	-												
2	Race 1	.064	-											
3	Race 2	-.015	-.871***	-										
4	Median Family Income	.034	.114	-.096	-									
5	SAT Math Subscore	-.266**	.106	-.090	.079	-								
6	SAT ERW Subscore	.073	.032	-.021	.071	.334***	-							
7	Algebra 1 GPA	.022	-.125	.109	-.107	-.124	-.117	-						
8	Geometry GPA	-	-	-	-	-	-	-	-					
9	Algebra 2 Taken	.124	.074	-.109	.002	-.128	-.082	-.034	-	-				
10	Algebra 1 Factor Score	-.062	.209*	-.111	.151	.405***	.167	-.046	-	-.065	-			
11	Geometry Factor Score	-.093	.043	-.122	-.038	.293***	.208*	.005	-	-.206*	.119	-		
12	PreCalculus Factor Score	-.118	.156	-.168*	.313***	.427***	.353***	-.079	-	-.082	.392***	.163	-	
13	% Grade in IMC	-.044	.227**	-.182*	.058	.415***	.241**	-.042	-	-.107	.389***	.247**	.390***	-

Note. *p < .05, **p < .01, ***p < .001. ERW = Evidence-Based Reading and Writing, GPA = Grade Point Average, IMC = Initial Mathematics Course

Total score regression. A Hierarchical Multiple Linear Regression was conducted to explore the relationship between the main predictor variable of Mathematics Placement Test Total Score and the criterion variable of the percentage grade received in the student's initial upper level mathematics course. Regression results suggest that a significant proportion of the total variance in students' grades is explained by the collection of independent variables ($R^2 = .290$, $F_{[9, 128]} = 5.814$, $p < .001$). More specifically, the predictors accounted for 29.0% of the variance in the percentage grade students' received in their initial mathematics course.

Overall, there were nine predictors in this model and the latter two regression blocks were significant. Additionally, the predictor of students' high school Geometry GPA was removed because it was a constant of 4.0 among the sample. Examining the final block of the regression model, displayed in Table 20 below, it is evident that the student's total score from the mathematics placement test was a significant predictor of the student's percentage grade received in their initial mathematics course ($t = 3.712$, $p < .001$). Specifically, for each one-point increase in students' Mathematics Placement Test Total Score, the students' grade received in their first semester mathematics course increased by .288 percentage points. Therefore, the Mathematics Placement Test Total Score is predictive of student success in their initial upper level mathematics course (i.e., Mathematical Investigations III or IV), providing evidence of Criterion-Related (i.e., Predictive) Validity.

Table 20

Hierarchical Multiple Linear Regression for Upper Level Mathematics Courses - Total Score (n = 138)

Variables	*B*	*SE*	*t*	*β*	95% CI for B	
					Lower	Upper
Gender	.909	1.041	.874	.069	-1.150	2.969
Race 1	2.102	1.951	1.078	.161	-1.757	5.962
Race 2	.109	1.995	.055	.008	-3.839	4.057
Median Family Income	.000	.000	-1.137	-.090	.000	.000
SAT Math Subscore	.029	.014	2.166*	.203	.003	.056
SAT_ERW Subscore	.001	.010	.059	.005	-.019	.021
Algebra 1 GPA	1.979	2.950	.671	.052	-3.858	7.817
Algebra 2 Taken	-1.031	2.677	-.385	-.030	-6.328	4.265
Placement Test Total Score	.288	.077	3.712***	.367	.134	.441

Note. $^{*}p < .05$, $^{***}p < .001$. ERW = Evidence-Based Reading and Writing, GPA = Grade Point Average, *B* = Unstandardized Regression Coefficient, *SE* = Standard Error, β = Standardized Regression Coefficient, CI = Confidence Interval

Subscale score regression. To further understand the relationship between the mathematics placement test and students' percentage grade received in their initial upper level mathematics course, a Hierarchical Multiple Linear Regression was conducted with the three factor subscores of Algebra 1, Geometry, and PreCalculus. Regression results indicated that a significant proportion of total variance in students' grades is explained by the collection of independent variables ($R^2 = .308$, $F_{[11, 126]} = 5.096$, $p < .001$). Thus, the predictor variables accounted for 30.8% of the variance in the percentage grade students' received in their initial upper level mathematics course.

Similar to the total score regression, the second and third regression blocks were significant for the eleven predictor variables. Again, the predictor variable of High

School Geometry GPA was removed from the analysis due to it being a constant variable. Exploring the final block of the regression model, Table 21 below shows that both the Algebra 1 Factor Score ($t = 2.075$, $p < .05$) and the PreCalculus Factor Score ($t = 2.188$, $p < .05$) are significant predictors of the student's percentage grade received in their first mathematics course. More specifically, for each one-point increase in students' Mathematics Placement Algebra 1 Factor Score, the students' grade received in their first semester mathematics course increased by .562 percentage points. Likewise, for each one-point increase in students' Mathematics Placement Test PreCalculus Factor Score, the students' grade received in their first semester upper level mathematics course increased by .207 percentage points.

Table 21

Hierarchical Multiple Linear Regression for Upper Level Mathematics Courses - Subscale Scores (n = 138)

Variables	*B*	*SE*	*t*	*β*	95% CI for B	
					Lower	Upper
Gender	.775	1.038	.747	.059	-1.278	2.828
Race 1	2.430	2.098	1.158	.187	-1.722	6.581
Race 2	.550	2.149	.256	.041	-3.702	4.802
Median Family Income	.000	.000	-.894	-.071	.000	.000
SAT Math Subscore	.029	.014	2.132*	.198	.002	.056
SAT_ERW Subscore	.004	.010	.446	.038	-.015	.024
Algebra 1 GPA	.438	3.316	.132	.010	-6.125	7.001
Algebra 2 Taken	-1.051	2.252	-.467	-.036	-5.508	3.406
Algebra 1 Factor Score	.562	.271	2.075*	.182	.026	1.097
Geometry Factor Score	.388	.256	1.520	.123	-.117	.894
PreCalculus Factor Score	.207	.095	2.188*	.206	.020	.394

Note. $^{*}p < .05$. ERW = Evidence-Based Reading and Writing, GPA = Grade Point Average, *B* = Unstandardized Regression Coefficient, *SE* = Standard Error, β = Standardized Regression Coefficient, CI = Confidence Interval

As previously mentioned, Mathematical Investigations is a four-semester sequence of courses preparing students for Calculus. According to the course syllabus, students entering Mathematical Investigations III (i.e., the third course the sequence) should demonstrate a strong background in Algebra and Geometry to be able to expand upon their mathematical thinking throughout this course. The final course of the sequence (i.e., Mathematical Investigations IV) focuses on developing students' learning in the areas of trigonometry, vectors, polar coordinates, and mathematical induction. The strength of the Predictive Validity evidence lies in the Mathematics Placement Test Total Score predicting student success in their initial upper level mathematics course (i.e.,

Mathematical Investigations III or IV). More specifically, the subscores obtained from the Algebra 1 and PreCalculus sections of the mathematics placement test predict student success in courses containing those content areas, providing strong evidence of Predictive Validity.

Discussion

Research has demonstrated the significant impact that course placement decisions can have on a student's future academic preparation (McDaniel et al., 2007; Morgan & Michaelides, 2005). Specifically, students who begin in a mathematics course that is appropriate given their background have an increased chance of succeeding in their initial mathematics course and their subsequent mathematics courses (Akst & Hirsch, 1991; Latterell & Regal, 2003; Marshall & Allen, 2000; Mattern & Packman, 2009; Norman et al., 2011; Shaw, 1997). Therefore, it is critically important to provide evidence of the effectiveness of placement measures and procedures to ensure a reduction in these unintended consequences of misplacement.

Findings from the Hierarchical Multiple Linear Regressions for both lower and upper level mathematics courses demonstrate that the total score students' receive on the mathematics placement test predicts their achievement in their initial mathematics course, above and beyond the contributions of their demographic information and previous academic background. Additionally, the combination of predictor variables in the lower level regression accounted for a greater proportion of variance explained (36.6%) in students' first semester mathematics grade compared to the upper level regression (29.0% variance explained), echoing the use of multiple measures to enhance course placement

decisions (e.g., Armstrong, 1995; Erwin & Worrell, 2012; Marwick, 2004; Ngo & Kwon, 2015; Noble et al., 2003). This finding extends the existing literature by demonstrating the influence of multiple measures on course placement decisions, especially for courses at the lower levels among gifted high school students.

Results from the Hierarchical Multiple Linear Regressions using the factor subscale scores as predictors revealed similar patterns (see Table 22 below). More specifically, the subscale score from the Algebra 1 section of the mathematics placement test was the strongest predictor of student success among the lower level mathematics courses (i.e., Mathematical Investigations I or II). Similarly, both the Algebra 1 and PreCalculus Factor Scores from the mathematics placement test were significant predictors of students' first-semester grades in an upper level mathematics course (i.e., Mathematical Investigations III or IV). These findings may contradict post-secondary education literature which found students' high school coursework and grades received in those courses to be stronger predictors of college course success compared to an incoming placement test score (Latterell & Regal, 2003).

Table 22

Summary of Hierarchical Multiple Linear Regression Results

Course	Type	DV	Block	IV	Direction
Lower Level	Total	Course Grade	(3) Mathematics Placement Test	Total Test Score	Positive
	Subscale	Course Grade	(3) Mathematics Placement Test	Algebra 1 Subscore	Positive
Upper Level	Total	Course Grade	(2) Incoming Performance	SAT Math Subscore	Positive
		Course Grade	(3) Mathematics Placement Test	Total Test Score	Positive
	Subscale	Course Grade	(2) Incoming Performance	SAT Math Subscore	Positive
		Course Grade	(3) Mathematics Placement Test	Algebra 1 Subscore	Positive
		Course Grade	(3) Mathematics Placement Test	PreCalculus Subscore	Positive

Additionally, the regression models used in this study included demographic control variables such as gender, race/ethnicity, and SES. Specific to gender, the literature includes that males take more advanced mathematics courses in high school and obtain higher scores on standardized assessments (Bridgeman & Wendler, 1989, 1991; Catsambis, 1994; Davis & Shih, 2007; Ellison & Swanson, 2018; Gallagher & De Lisi, 1994; Pedro et al., 1981). More recent research has reported that gender differences in math scores on standardized assessments are minimal and non-significant (Else-Quest et al., 2010; Hyde et al., 2008; Lindberg et al., 2010). Still other studies have noted that girls outperform boys with respect to the grades received in their mathematics courses (Arslan et al., 2012; Ding et al., 2006; Gherasim et al., 2013; Wang & Degol, 2017). The collection of results suggests that there is some relationship between gender and mathematics achievement. Therefore, it was surprising that students' gender was not a significant predictor of the outcomes in the regression models.

Likewise, research has continued to examine the effects of race/ethnicity and SES on students' mathematics achievement. In a meta-analysis by Sirin (2005), SES had a medium effect on academic achievement at the student level. This finding strengthened earlier research findings that concluded SES (i.e., income) was an important predictor of mathematics achievement and career decisions, especially for females (Gonzalez & Kuenzi, 2012). Moreover, studies have shown that both race/ethnicity and SES account for a significant and meaningful percentage of variance in students' test scores (White et al., 2016). Similar to the gender variable discussion above, despite the body of research demonstrating relationships between the demographic variables and math achievement, neither race/ethnicity nor SES was a significant predictor of the outcomes in the current study.

Although previous research has demonstrated the effects of demographic information on students' academic performance, that was not the case in the current study. Instead, it is possible that the total and factor subscale scores from the mathematics placement test were overwhelmingly influential and dominated the overall models in this study. This finding is supported in previous literature, which has demonstrated moderate-to-strong relationships between scores received on achievement tests and students' subsequent course performance (Bridgeman & Wendler, 1989; Davis & Shih, 2007; Erwin & Worrell, 2012; Mattern & Packman, 2009; Rueda & Sokolowski, 2004). Another possible explanation is the "recency effect" such that the variables appearing closer in time to the outcome variable become more influential within the model. Thus, since the mathematics placement test was completed approximately six to

eight months prior to students' receiving their grades in their first semester mathematics course, it is possible that the test scores obtained were stronger predictors than the demographic variables included within the models.

Implications

This study examined the relationship between student's mathematics placement test scores and their subsequent performance in their initial mathematics course. Additionally, the models used incorporated students' demographic information and previous mathematical coursework to reflect the reality of placement practices at the high school under study. As such, results of this study provide valuable insight for students and faculty members, as well as administrators and the larger community.

One of educational measurement's core activities is to aid the educational process of each student as they learn (Wilson, 2018). Findings from this study can help students and faculty members identify the academic needs of students so that the proper resources and supports can be implemented. Traditionally, mathematics faculty members have used the total score obtained on the Mathematics Placement Test to guide students' initial course placement. However, by providing evidence regarding the underlying factor structure of the mathematics placement test (Manuscript 2) and developing factor subscale scores (Manuscript 3), students and faculty members can use the newly developed Algebra 1, Geometry, and PreCalculus subscores to gauge student readiness for a particular course. This targeted approach can illuminate both students and faculty about the content students have or have not mastered, allowing the institution to address gaps in student understanding prior to course enrollment. Additionally, as this is the only

study to examine predictive validity of a placement test in the context of a gifted, residential STEM high school, students, parents, and faculty can now have the full gamut of reliability and validity evidence needed to make appropriate course placement decisions.

Similarly, educational administrators and other interested stakeholders can be assured that there is an increased likelihood that the consequences of course misplacement will be minimized. Numerous studies have shown that success in a student's initial mathematics course increases their likelihood of greater achievement in subsequent mathematics courses (e.g., Akst & Hirsch, 1991; Latterell & Regal, 2003; Marshall & Allen, 2000; Norman et al., 2011; Shaw, 1997). Thus, by providing evidence of Criterion-Related Validity, the main purpose of placement testing has been achieved in that the mathematics placement test scores can be used to appropriately match the students' existing level of mathematics knowledge to instruction commensurate with their previous academic preparations. Moreover, in the case where a student completed additional coursework in the summer prior to attending the high school, the development of the three subscale scores (i.e., Algebra 1, Geometry, and PreCalculus) can provide faculty members and administrators with a more targeted placement test without sacrificing reliability and/or validity.

Lastly, the implications of this study go beyond the local context. In the current era of accountability, placement exams and methods that are rigorous and defensible are critical for educational institutions at varying levels to justify their use and to address questions of their impact on students' educational outcomes. A number of studies have

evaluated placement tests at the post-secondary level, with more research needed at lower educational levels. The current study can provide a foundation for other similar high schools to examine the placement tests, procedures, and decisions used at their own institutions.

Limitations and Future Research

Although this study provides evidence of Predictive Validity, there were some limitations. The original sample included student data from across four cohorts, which were determined to be statistically equivalent. However, due to inaccessible and incomplete data, the final regression analyses only included two of the four cohorts. Future research may consider extending this study to more recent cohorts for which complete data may be available in the future.

Another possible limitation was the use of student's median family income based on their home address zip code as an indicator for socioeconomic status (SES). While there is some promising literature that supports the use of neighborhood-level SES indicators (Labovitz, 1975; Sirin, 2005), there is no universally accepted proxy of SES among the educational research literature. Moreover, though the census bureau has median family income data available at the block level, the current data set contained only participants five digit zip code, making coding based on the nine digit zip code not possible. Future research could examine other proxies for SES to determine whether or not they influence the regression models in a different way.

A third limitation to consider is the use of students' SAT scores within the regression models. All applicants are required to submit their current SAT score reports

directly from the College Board for each exam the prospective student completes. Regardless of which test administration the score was from, the high school reports only the highest SAT Mathematics and Evidence-Based Reading and Writing scores as part of the student's admissions application. According to the College Board, robust measures are taken to ensure the accuracy of students' scores across versions of the SAT (College Board, 2018a). This suggests that regardless of the test the student completed, their scores have a consistent interpretation and representation of their underlying knowledge. Future research could examine the impact of using the highest SAT scores within the regression model compared to students' most recent testing administration.

Moreover, on March 1, 2016, the College Board changed the scoring scale for the SAT from a maximum score of 2400 (prior to 2016) to a maximum score of 1600 (after 2016). Therefore, the SAT scores gathered from the admissions applications in this study included both SAT scoring scales, which were all converted to the post-2016 scale using the concordance tables provided by the College Board (2016). Future research may consider extending this study to more recent cohorts for which data will become available so that there is a consistency in the SAT scoring scales reported.

A final limitation to consider is the extent to which grading scales across the state of Illinois are equivalent. The near-universal use of placement tests at the post-secondary level emerged due to the incomparability of unknown factors such as the content and rigor of courses and the grading scales used at different schools (Kossack, 1942; Linn, 1994; Ngo & Kwon, 2015; Noble et al., 2003). In an environment where students with varying experiences and backgrounds from across the state are accepted into the high

school, it is important to consider how comparisons are made among student grades. Future research could explore other ways to measure students' previous academic coursework so that more accurate course predictions can be made.

Similarly, future research could examine the variance of grades received within the high school under study. The grading scale used in mathematics at the current high school is as follows: A [92.5 – 100%], A- [89.5 – 92.5%), B+ [87.5 – 89.5%), B [82.5 – 87.5%), B- [79.5 – 82.5%), C+ [77.5 – 79.5%), C [72.5 – 77.5%), C- [69.5 – 72.5%), and D [0 – 69.5%). However, when critically analyzing the data, it was determined that this scale was not implemented consistently across all students, most likely due to "teacher discretion." Again, future research may consider additional ways to measure student performance and success in coursework.

Conclusions

This study investigated the Criterion-Related Validity of the item scores on a mathematics placement test at a gifted residential high school for students interested in STEM. More specifically, this study examined the relationship between students' mathematics placement test total and factor scores with students' subsequent performance in their first semester mathematics course.

Using a series of four Hierarchical Multiple Linear Regressions, it was determined that the total score obtained on the mathematics placement test was predictive of student success in their initial mathematics course. When examining the predictiveness of the factor scores for students in a lower level mathematics course (i.e., Mathematical Investigations I or II), the Algebra 1 Factor Score was found to be the only

significant predictor of the percentage grade students' received in that course. Likewise, both the Algebra 1 and PreCalculus Factor Scores were determined to be significant predictors of student success in their first upper level mathematics course, either Mathematical Investigations III or IV.

Therefore, the mathematics placement test demonstrates evidence of Predictive Validity and can be used in the course placement decision-making process. In an era of accountability, this study can encourage other educational institutions, at all levels, to validate their placement processes and decisions. In doing so, all stakeholders can be confident that students' future educational outcomes are being optimized while the consequences of misplacement are being minimized.

CHAPTER VIII

CONCLUSIONS

The overarching goal of this study was to investigate the psychometric properties of a mathematics placement test at a gifted residential high school for students interested in STEM. More specifically, the four objectives of this study were: (1) To provide evidence of Content Validity, (2) To provide evidence of Construct Validity and Internal Consistency Reliability, (3) To examine the characteristics and potential bias of the items for males and females and (4) To provide evidence of Criterion-Related Validity. The literature, methodology, results, and discussion for each of the four objectives were presented as four manuscripts within the larger document.

Manuscript 1 examined the Content Validity of the mathematics placement test using a card-sorting technique replicated from a study by D'Agostino et al. (2011). Data were collected from internal and external subject matter experts (SMEs) and were analyzed using Multidimensional Scaling and Hierarchical Cluster Analysis. The final cluster solution revealed six unique clusters that were labeled Algebraic Operations, Solving Equations, Graphing Functions, Evaluating Functions, Trigonometry, and Geometry. Additionally, results demonstrated some congruence between the internal and external SME configurations, indicating marginal evidence of Content Validity.

The second manuscript sought to provide evidence of Construct Validity and Internal Consistency Reliability of the mathematics placement test. Developed by faculty members, the mathematics placement test was designed to measure students' incoming mathematical knowledge prior to entering a Calculus sequence. Existing data from four

cohorts of students were obtained and analyzed using Exploratory Factor Analysis (EFA) and the Kuder-Richardson (KR-20) formula. Results from the EFA suggested that the mathematics placement test was comprised of three factors, which included PreCalculus, Geometry, and Algebra 1. All of these factors had moderate to strong Internal Consistency Reliabilities. Therefore, Manuscript 2 demonstrated evidence of Construct Validity and Internal Consistency Reliability for the population of interest.

The main objectives of Manuscript 3 were to examine the item parameters (i.e., item difficulty and discrimination) and Differential Item Functioning (DIF) of the mathematics placement test. Using the Two-Parameter Logistic (2PL) model from Item Response Theory, existing data from four cohorts of students were analyzed. Due to the unidimensionality assumption of the 2PL model and the results from Manuscript 2, the Algebra 1, PreCalculus, and Geometry factors were examined independently.

Results from the analysis of Algebra 1 and Geometry items indicated that these portions of the mathematics placement test were generally easy for the population of interest. These sections of items also were unable to distinguish between varying proficiencies along the Algebra 1 and Geometry continuums. Item analysis results of the PreCalculus factor suggested that these items from the mathematics placement test were more challenging for the population of interest. Not only were the PreCalculus items able to sufficiently discriminate between individuals of varying PreCalculus knowledge, but the information from the test was also slightly above average.

Finally, Manuscript 4 examined the Criterion-Related Validity of the item scores on the mathematics placement test using existing data from two cohorts of students. Two

Hierarchical Multiple Linear Regressions were conducted for students enrolled in either a lower level mathematics course (i.e., Mathematical Investigations I or II) or an upper level mathematics course (i.e., Mathematical Investigations III or IV; four regression total). The first regression for each group used students' mathematics placement test total score as the main predictor variable. In the second regression for each group, the main predictor variable was students' mathematics placement test factor subscores for the three factors of Algebra 1, Geometry, and PreCalculus.

Results from the regressions for both lower and upper level mathematics courses showed that the total score students received on the mathematics placement test predicts achievement in their first semester mathematics course. More specifically, Algebra 1 scores from the mathematics placement test were the strongest predictor of student success among the lower level mathematics courses (i.e., Mathematical Investigations I or II). Similarly, both the Algebra 1 and PreCalculus Factor Scores from the mathematics placement test were significant predictors of students' grades in their first upper level mathematics course (i.e., Mathematical Investigations III or IV). Each of these findings provide evidence of Criterion-Related (i.e., Predictive) Validity of the items scores on a mathematics placement test for gifted, residential high school students interested in STEM.

Synthesis of Manuscripts 1 – 4

Validity has been argued as the most important criteria to ensure the quality of a test. While there are three major types of validity (i.e., Content, Construct, and Criterion-Related), one is no more important than another. Equally important in judging the

validity of test scores is the analysis of item-level data to determine the quality of the test and the information it generates (Adedoyin & Mokobi, 2013). For this reason, it is vital that each psychometric aspect of a test is examined to appraise the overall quality and the inferences that can be made from the scores.

Manuscripts 1 and 2

As previously mentioned, some literature exists that provides evidence regarding the similarity of results obtained through a Hierarchical Cluster Analysis (HCA) and factor analytic procedures (Capra, 2005; Revelle, 1979). While studies comparing the two techniques are sparse, there is an abundance of literature on the underlying validities that are shared by both analytic strategies. In the current study, HCA and Exploratory Factor Analysis (EFA) were conducted in two separate manuscripts to provide evidence of Content and Construct Validity, respectively. However, the psychometric literature conceptualizes all validities under one overarching framework of Construct Validity (Clark & Watson, 1995; Cook & Beckman, 2006; Loevinger, 1957; Messick, 1989).

In Loevinger (1957), a theoretical approach to scale development is discussed stating that there are three components of Construct Validity, namely substantive validity, structural validity, and external validity. The first component, substantive validity, is described as a critical first step to developing a precise and detailed definition of the target construct and its theoretical context (i.e., content domain; Clark & Watson, 1995; Loevinger, 1957). To develop a detailed construct definition, the scope and range of the content domain should be established. Following this, items are written to ensure that each area of the content domain is well represented (i.e., Content Validity). After writing

items covering the entirety of the content domain, factor analytic procedures can be used to reveal how the items are subdivided into subscales (i.e., factors). These analyses (e.g., factor analysis) may reveal that the number of items is too small to assess each area of the content domain reliably (Clark & Watson, 1995; Loevinger, 1957). To increase the amount of items, the process typically returns to the beginning to re-examine the construct definition (i.e., substantive validity). This cyclical process continues until enough evidence (i.e., objective and subjective) is obtained to support the overarching framework of Construct Validity. Thus, there is an iterative relationship between the traditionally defined concepts of Content and Construct Validity, and obtaining comparable results for the two types (separately) is unsurprising.

The final HCA solution contained six clusters, which were labeled as Algebraic Operations, Solving Equations, Graphing Functions, Evaluating Functions, Trigonometry, and Geometry. While this six-factor solution was considered when conducting the EFAs, it ultimately was unsuitable for these data given the presence of Heywood cases and lack of simple structure. Instead, the final EFA structure was comprised of the three factors – PreCalculus, Geometry, and Algebra 1.

Comparing the results from the HCA and EFA, the following observations were noted. The Geometry cluster from the HCA had a direct relationship to the Geometry factor of the EFA (i.e., the same items in both). Likewise, all items (i.e., except one) from the HCA Trigonometry cluster loaded the highest on the PreCalculus factor of the EFA. This relationship between the Trigonometry cluster and the PreCalculus factor was expected based on the sequence and design of the high school mathematics courses.

Next, the items in the first two clusters of the HCA (i.e., Algebraic Operations and Solving Equations) were mainly located in the Algebra 1 factor of the EFA. However, the clusters of Graphing and Evaluating Functions were split between the Algebra 1 and the PreCalculus factor. The distinction between the two factors appeared to be related to the placement of the items on the exam. Since mathematical knowledge is hierarchical in nature, meaning that you need to know Algebra first before completing PreCalculus, the majority of the earlier items on the exam loaded on the Algebra 1 factor. Conversely, the items that loaded highest on the PreCalculus factor from clusters three and four were the items involving graphing and evaluating higher order functions. Therefore, there appears to be reasonable evidence to support the similarity of results between the HCA and the EFA, further confirming the presence and relationship between Content and Construct Validity.

Validating the scores on a test requires a carefully structured argument where evidence has been collected to support or refute the intended interpretation of results (Cook & Beckman, 2006; Cronbach, 1971; Messick, 1995). Moreover, the validity of an instrument's scores depends on the construct definition, which necessitates an extensive literature review to detail the content of the domain (Clark & Watson, 1995; Cook & Beckman, 2006; Loevinger, 1957). Therefore, to evidence substantive validity in the current study, it was critical to obtain information from multiple different perspectives to ensure a common understanding about the underlying content and constructs of the mathematics placement test. That is, content validity, in essence, "lays the foundation"

for the other types of validity, and without this foundation, future substantive validation evidence is weak or non-existent.

Manuscripts 2 and 3

According to Loevinger (1957), the second component of Construct Validity is structural validity. This type of validity examines the extent to which the internal structure of the assessment reflected in the scores is consistent with the structure of the construct of interest (Messick, 1995). It is important to note that this definition consists of two distinct, but related parts. Before examining the consistency of the scores, it is imperative to understand the underlying structure of the construct of interest. As such, Manuscripts 2 and 3 explored the internal structure of the mathematics placement test by examining the patterns of relationships among item scores and between test scores.

In Manuscript 2, the internal structure (i.e., addressing Construct Validity) of the test was investigated using EFA to evidence the factors in the exam. Findings from the EFA revealed three factors PreCalculus, Geometry, and Algebra 1. To gather more detailed internal structure information to evidence Construct Validity, each of the three factors were subjected to item analysis, which included Differential Item Functioning (DIF). DIF was conducted to uncover the presence of systematic variations in responses to items among subgroups who were expected to perform similarly on the mathematics placement test. According to Crocker and Algina (2008), there are multiple ways to evidence a construct including examining group differences. If a construct is theoretically expected to show differences between groups, and that is demonstrated using t-Tests, ANOVAs, or DIF (as some examples), the evidence supporting the

construct's internal structure increases (or vice versa). In the current context, based on the historical literature, group differences (males versus females) on mathematics performance was expected. Indeed, the results showed that some items displayed significant DIF, which provides more evidence of Construct Validity. However, although this supports the construct under investigation in the current study, for practical and applied purposes and use, DIF should be minimized to ensure that the mathematics placement test is equally valid for both male and female students.

Recall that structural validity examines the extent to which the internal structure of the assessment reflected in the scores is consistent with the structure of the construct of interest (Messick, 1995). Thus, in addition to understanding the structure of the construct of interest, Manuscripts 2 and 3 also examined the consistency (i.e., reliability) of the scores. In Manuscript 2, the Internal Consistency Reliability of each factor was examined using the inter-item correlations and the Kuder-Richardson (KR-20) Formula. Broadly stated, reliability measures the consistency or accuracy of the research and provides evidence to the extent to which the research can be repeated (e.g., Cook & Beckman, 2006; Cronbach, 1951; Nunnally & Bernstein, 1978; Rossi et al., 2003; Wiersma & Jurs, 2009). Thus, items that are intended to measure a single construct should to some degree relate to one another. This was evidenced by acceptable inter-item correlations and moderate-to-strong KR-20 reliability estimates. However, in order to have a holistic understanding of the reliability of an instrument's scores, item analytic procedures must be conducted.

Manuscript 3 used the 2PL model from Item Response Theory to analyze the item-level data. The goal of an instrument is to accurately and consistently measure a student's true score by minimizing measurement error. To do so, requires that the items and test instructions are clearly written and understood and that the scoring of the observed tasks is as objective as possible. Thus, by examining the item-level data, test developers and researchers can gain a better understanding of how particular items are performing (or not in the case of negatively discriminating items). Additionally, item-level diagnostics such as local dependence are useful in determining whether two distinct items are too similar in what they are assessing, which can compromise the reliability of the scores. Thus, to maximize the information gained from an instrument, it is critical that the items on the exam be optimized to reduce measurement error and to fully understand the complexities of score reliability estimates.

Therefore, findings from Manuscripts 2 and 3 demonstrate the complexities of the internal structure of educational assessments and the need to review such information from various perspectives to support the argument of structural validity (i.e., Construct Validity). The previous definition of structural validity suggests that one must understand the underlying structure of the construct of interest first prior to determining the consistency of the scores. However, it has been said that reliability is a necessary, but not sufficient, condition for validity (Cook & Beckman, 2006; Cronbach, 1951). These two statements, both of which are correct, demonstrate the cyclical nature of validity and reliability. When there are concerns regarding item parameters (i.e., item difficulty and item discrimination) and DIF, the reliability of test scores are threatened along with the

interpretations and decisions that are made using the scores. Therefore, educational assessments need to be examined for their psychometric properties as a whole, rather than any one particular property of an assessment.

Manuscripts 1 – 4

Validity is a judgment concerning the extent to which inferences and actions based on test scores are appropriate given the empirical evidence and theoretical rationale (Cook & Beckman, 2006; Cronbach, 1971; Kimberlin & Winetrstein, 2008; Schmitz & delMas, 1991). Underlying each validation argument are assumptions that must be accepted as reasonable or plausible to support the overall interpretations and uses of the test scores (Kane, 1992; Sawyer, 1996). The current study has developed its validity argument through the combination of its four manuscripts.

As previously mentioned, validity consists of a carefully constructed argument where evidence has been collected from multiple sources to support or refute the intended interpretation of results (Cook & Beckman, 2006; Cronbach, 1971; Messick, 1995). Moreover, the validity of an instrument's scores depends on the construct definition, which is why some theorists suggest that all validity should be conceptualized as components of one overarching framework of Construct Validity (Clark & Watson, 1995; Cook & Beckman, 2006; Loevinger, 1957; Messick, 1989). As such, Messick (1989) presented five sources of evidence to support Construct Validity: content, response process, internal structure, relationships with other variables, and social consequences. While many articles cite only one or two sources of validity evidence, the current study included each of the five sources of evidence to support the overarching framework of

Construct Validity (Table 23 below). Furthermore, strong evidence from one source does not negate the need to seek validity evidence from other sources (Cook & Beckman, 2006).

Table 23

Summary of Evidence to Support Construct Validity

Manuscript	Content	Response Process	Internal Structure	Relationships	Social Consequences
1	X		X		
2	X		X		
3	X	x	X		x
4	x		x	X	X

Note. X = Validity evidence that was directly addressed; x = Validity evidence that was indirectly addressed

In Manuscript 1, internal and external subject matter experts (SMEs) were used to explore the congruence of the content domain among the two groups. Using MDS and HCA, it was determined that the content of the mathematics placement test items could be clustered into six mathematical areas, with approximately 63% agreement between internal and external SMEs. Thus, it is reasonable to say that the two groups agreed on the content present on the mathematics placement test, providing content evidence.

Secondly, Manuscript 2 demonstrated the presence of three underlying factors which were labeled as PreCalculus, Geometry, and Algebra 1. While each of these factors had strong Internal Consistency Reliability estimates, Manuscript 3 conducted an item analysis to further explore the quality of the items. Through the item analysis process, the mathematics placement test was refined by deleting items, and the Internal

Consistency Reliability of each factor was reassessed. Thus, Manuscript 2 and Manuscript 3 provided evidence to support the internal structure (and reliability) of the instrument. Manuscript 3 also provided some additional theoretical evidence related to the construct (i.e., the internal structure) via item bias.

Finally, Manuscript 4 provided evidence to support the relationship between students' total and factor scores from the mathematics placement test with students' performance in their first semester mathematics course, based on the revised test from Manuscript 3. Moreover, by establishing the relationship of the mathematics placement test to other variables, Manuscript 4 provided additional information indicating that the consequences of misplacement had been minimized, addressing two sources of evidence as defined by Messick (1989) (i.e., relationships to other variables and social consequences).

Overall, the combination of the four manuscripts provides strong evidence regarding the psychometric properties of the mathematics placement test. More specifically, the current mathematics placement test and procedures appear appropriate for gifted residential students interested in STEM given the empirical evidence demonstrated in the current study. Therefore, the continued use of the revised mathematics placement test in the course placement decision-making process is supported via a compelling validity argument.

APPENDICIES

APPENDIX A

INSTRUCTIONS FOR PART I OF THE MATHEMATICS PLACEMENT EXAM

APPENDIX A

INSTRUCTIONS FOR PART I OF THE MATHEMATICS PLACEMENT EXAM

2018

[Name of High School]
Mathematics Placement Test

Part I

Do NOT turn the page until you are told to do so.

Instructions:

NO CALCULATORS. While calculators will be used in all Academy courses, they will not be permitted on this test.

Time limit for this part of the test is 45 minutes.

On the following pages are 50 short answer questions. There is a box with each problem and a line at the bottom of the box on which to record your answer. Do your calculations in the box; however, only the recorded answer will be graded. You may use the back sides of the pages if you need more space to calculate. No partial credit will be given.

Use pencil only! Write answers neatly.

Name_____

The name of the math course you are currently taking: _____

The last topic you covered in math class was: _____

APPENDIX B

INSTRUCTIONS FOR PART II OF THE MATHEMATICS PLACEMENT EXAM

APPENDIX B

INSTRUCTIONS FOR PART II OF THE MATHEMATICS PLACEMENT EXAM

2018

[Name of High School]
Mathematics Placement Test

Part II

Do NOT turn the page until you are told to do so.

Instructions:

While calculators will be used in all Academy courses, they **will not be permitted on this test**.

The time limit for this part of the test is 85 minutes.

On the following pages are 45 multiple choice questions. Use a soft lead pencil to mark your answers on the separate answer sheet that has been provided. Be careful to fill the answer next to the same number as the problem you are solving. You may use any space on the test to do your calculations. Scratch paper will be provided if you prefer to use it. However, only the recorded answer will be graded.

This test will be machine scored. Make NO stray marks on your answer sheet. Be sure erasures are complete.

PLEASE PRINT YOUR NAME BELOW AND ON THE ANSWER SHEET.

Name_____

APPENDIX C

RECRUITMENT EMAIL FOR SUBJECT MATTER EXPERTS

APPENDIX C

RECRUITMENT EMAIL FOR SUBJECT MATTER EXPERTS

Address Line: This email will be sent individually to allow for confidentiality of the research participants' identities and to address each individual by name along with their relevant experience(s). Additionally, this email will be sent from the Principal Investigator's Kent State University email account (hwilso20@kent.edu) to protect the identity of the participating institution.

Subject: Research Participation Invitation Assessing the Content Validity of a Mathematics Placement Test

Body: This email message is an approved request for participation in research that has been approved or declared exempt by the Institutional Review Boards (IRB) at both the participating location and Kent State University.

Good morning/afternoon/evening [NAME],

You have been selected to participate in a research study regarding the Content Validity of a mathematics placement test due to your [insert relevant experience and research here]. This invitation email will provide you with general information regarding the research project and the tasks requested of you as a participant. Additional information about the research study can be found in the attached consent form. Your participation in this study is voluntary.

Purpose: It is the intent of the current study to examine the psychometric properties of a mathematics placement test at a gifted residential high school focused on STEM. More specifically, this portion of the research project seeks to identify evidence of Content Validity (i.e., whether or not items on an instrument suitably measure a construct of particular interest).

Procedures: Participation in this study is completely voluntary and participants may choose to withdraw from the study at any time without consequence for doing so. Moreover, participation in this research will require each participant to be able and willing to complete a card-sorting tasks of 107 items. The card-sorting task will ask subject matter experts (SMEs) to sort the 107 items into groups based on item similarity

and to record the final groupings on a provided piece of paper. Upon completion of the card-sorting task, participants will return all provided materials to the principal investigator. All participant identities, responses, and contact information will remain confidential through the use of random study identification numbers.

Questions: This project was approved by the Kent State University IRB (#17-475) and the study site's IRB (IRB2017-03) on September 29, 2017. Pertinent questions or concerns about the research, research participants' rights, and/or research-related injuries to participants should be directed to the IRB Research Compliance Coordinator, Kevin McCreary by phone at 330.672.8058 or by email at kmccreal@kent.edu.

If you are willing to participate in this research study or have additional questions about this research, please contact me no later than **Friday, February 8, 2019**.

Thank you for considering this research opportunity.

Sincerely,

Hannah R. Anderson
Ph.D. Student of Evaluation and Measurement
Kent State University
Phone: 234.571.8923
Email: hwilso20@kent.edu

APPENDIX D

INFORMED CONSENT TO PARTICIPATE IN A RESEARCH STUDY

APPENDIX D

INFORMED CONSENT TO PARTICIPATE IN A RESEARCH STUDY

Study Title: A Psychometric Investigation of a Mathematics Placement Test at a Science, Technology, Engineering, and Mathematics (STEM) Gifted Residential High School

Principal Investigator: Hannah R. Anderson

You are being invited to participate in a research study. This consent form will provide you with information regarding the research project, the tasks requested of you as a participant, and the associated risks and benefits of the research. Your participation in this study is voluntary. Please read this form carefully and ask questions, if needed, to ensure that you fully understand the research project in order to make an informed decision. You will receive a copy of this document for your records.

Purpose:

Placement testing has become an integral component of the admissions process within American post-secondary institutions. The overarching goal of administering placement tests is to accurately distinguish between those students who do or do not have the knowledge base to succeed in a particular course (Feldhusen & Jarwan, 1995; J. P. Marshall & Allen, 2000; Mattern & Packman, 2009; Sawyer, 1996; Schmitz & delMas, 1991). In an era of federal regulations such as No Child Left Behind, and a need for increased accountability, American post-secondary institutions are being asked to defend the use and interpretations of their placement testing decisions. While the current study takes place at a gifted residential high school (i.e., Grades 10 through 12), the purpose is the same. Thus, it is the intent of the current study to examine the psychometric properties of a mathematics placement test at a gifted residential high school focused on STEM. More specifically, this portion of the research project seeks to identify evidence of Content Validity (i.e., whether or not items on an instrument suitably measure a construct of particular interest).

Procedures:

Participation in this study is completely voluntary and participants may choose to withdraw from the study at any time without consequence for doing so. Moreover, participation in this research will require each participant to be able and willing to complete a card-sorting task of 107 items. The card-sorting task will ask subject matter experts (SMEs) to sort the 107 items into groups based on item similarity and to record the final groupings on a provided piece of paper. Upon completion of the card-sorting task, participants will return all provided materials to the principal investigator. All participant identities, responses, and contact information will remain confidential through the use of random study identification numbers.

Benefits:

This research study does not provide direct benefits to the participant. However, by assisting in the investigation of Content Validity evidence, the uses and interpretations of the mathematics placement test will be better understood so that future recommendations can be made. Additionally, by exploring psychometric properties of a mathematics placement test and presenting the findings in a scholarly journal, other researchers will be able to replicate and expand upon the current research study in order to move the educational field forward.

Risks and Discomforts:

There are no anticipated risks beyond those encountered in everyday life.

Privacy and Confidentiality:

Your study related information will be kept confidential within the limits of the law. Any responses and identifying information will be kept in a secure location with restricted access by only the principal investigator. Research participants will not be identified in any publication or presentation of research results. Only aggregate data will be used in addition to a general acknowledgement of those who participated in this portion of the research study.

It is important to note that the items used in the card-sorting task are the same items being actively used on the mathematics placement test. Therefore, each participant agrees to the access and use of this confidential data for the sole purposes of this research study. Disclosing confidential information directly or allowing non-authorized access to such information may subject that individual to criminal prosecution.

Voluntary Participation:

Taking part in this research project is entirely your decision. You may choose to not participate or to discontinue your participation at any time without penalty or loss of benefits to which you are otherwise entitled. You will be informed of any new, relevant information that may affect your health, welfare, or willingness to continue your study participation.

Contact Information:

If you have any questions or concerns about this research, you may contact Hannah Anderson by phone at 234.571.8923 or by email at hwilso20@kent.edu. This project has been approved by both the Institutional Review Board (IRB) at Kent State University (#17-475) and the site of the research study (IRB2017-03). If you have any questions or concerns about your rights as a participant or concerns about the research, please contact the Kent State University IRB at 330.672.2704.

Consent Statement and Signature:

I have read this consent form and have had the opportunity to have my questions answered to my satisfaction. I voluntarily agree to participate in this study. I understand that a copy of this consent form will be provided to me for future reference.

_____

Participant Name (Printed)

_____ _____

Participant Signature **Date**

_____ _____

Hannah R. Anderson
Principal Investigator Signature **Date**

References

Feldhusen, J. F., & Jarwan, F. (1995). Predictors of academic success at state-supported residential schools for mathematics and science: A validity study. *Educational and Psychological Measurement, 55*(3), 505-512.

Marshall, J. P., & Allen, B. D. (2000). The Development and Implementation of a Mathematics Placement Program.

Mattern, K. D., & Packman, S. (2009). Predictive validity of ACCUPLACER scores for course placement: A meta-analysis.

Sawyer, R. (1996). Decision Theory Models for Validating Course Placement Tests. *Journal of Educational Measurement*, 271-290.

Schmitz, C. C., & Delmas, R. C. (1991). Determining the Validity of Placement Exams for Developmental College Curricula. *Applied measurement in education, 4*(1), 37-52. doi:10.1207/s15324818ame0401_4

APPENDIX E

SUBJECT MATTER EXPERT'S CARD SORTING TASK RESPONSE SHEET

APPENDIX E

SUBJECT MATTER EXPERT'S CARD SORTING TASK RESPONSE SHEET

Institutional Review Board Approvals
Kent State University (#17-475)
Site of Research (IRB2017-03)

Subject Matter Expert's Card Sorting Task Response

SUBJECT MATTER EXPERT DEMOGRAPHICS	
Current Title:	
Current Employer:	
Type of Institution: ☐ High School ☐ Community College ☐ 4-year College/University	
Current Grade Level(s) Taught:	
Courses Most Commonly Taught:	
Highest Degree Awarded:	
Degree Major:	
Years Teaching Mathematics:	
Years Teaching Gifted Students (i.e., Honors, Advanced Placement):	
Total Years Teaching:	
Gender:	Race/Ethnicity:

Institutional Review Board Approvals
Kent State University (#17-475)
Site of Research (IRB2017-03)

INSTRUCTIONS for SUBJECT MATTER EXPERTS:

DIRECTIONS: Please place each item into a meaningful pile or group based on the **similarity of the content** of the items.

Card-Sorting Rules:

1. Place each item or card into only one pile or group.
2. Refrain from creating as many piles or groups as there are items.
3. Create more than one pile or group.

CARD-SORTING RESPONSES: Please assign each group of items a group title or name.

Group 1 Name:

Item Number(s) in Group 1:

Group 2 Name:

Item Number(s) in Group 2:

Group 3 Name:

Item Number(s) in Group 3:

Group 4 Name:

Item Number(s) in Group 4:

Group 5 Name:

Item Number(s) in Group 5:

Group 6 Name:

Item Number(s) in Group 6:

Group 7 Name:

Item Number(s) in Group 7:

Group 8 Name:

Item Number(s) in Group 8:

Institutional Review Board Approvals
Kent State University (#17-475)
Site of Research (IRB2017-03)

Group 9 Name:

Item Number(s) in Group 9:

Group 10 Name:

Item Number(s) in Group 10:

Group 11 Name:

Item Number(s) in Group 11:

Institutional Review Board Approvals
Kent State University (#17-475)
Site of Research (IRB2017-03)

Group 12 Name:

Item Number(s) in Group 12:

Group 13 Name:

Item Number(s) in Group 13:

Group 14 Name:

Item Number(s) in Group 14:

Institutional Review Board Approvals
Kent State University (#17-475)
Site of Research (IRB2017-03)

Group 15 Name:

Item Number(s) in Group 15:

Group 16 Name:

Item Number(s) in Group 16:

Group 17 Name:

Item Number(s) in Group 17:

Institutional Review Board Approvals
Kent State University (#17-475)
Site of Research (IRB2017-03)

Group 18 Name: [redacted]
Item Number(s) in Group 18: [redacted]
Group 19 Name: [redacted]
Item Number(s) in Group 19: [redacted]
Group 20 Name: [redacted]
Item Number(s) in Group 20: [redacted]

REFERENCES

REFERENCES

Abedalaziz, N., Leng, C.H., & Alahmadi, A. (2018). Detecting a gender-related differential item functioning using transformed item difficulty. *Malaysian Online Journal of Educational Sciences, 2*(1), 16-22.

Adedoyin, O.O., & Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Sciences, 3*(4), 992-1011.

Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., & Belongie, S. (2007). *Generalized non-metric multidimensional scaling.* Paper presented at the Artificial Intelligence and Statistics.

Akst, G., & Hirsch, L. (1991). Selected studies on math placement. *Review of Research in Developmental Education, 8*(4), 6.

Allen, M.J., & Yen, W.M. (2001). *Introduction to measurement theory*: Waveland Press.

Altenhof, J.C. (1984). *Influence of item characteristics on male and female performance on SAT-Math.* City University of New York.

Arce, C., & Gärling, T. (1989). Multidimensional scaling. *Anuario de psicología/The UB Journal of psychology*(43), 63-80.

Armstrong, W.B. (1995). Validating placement tests in the community college: The role of test scores, biographical data, and grading variation.

Armstrong, W.B. (2000). The association among student success in courses, placement test scores, student background data, and instructor grading practices. *Community College Journal of Research and Practice, 24*(8), 681-695.

Arslan, H., Canli, M., & Sabo, H.M. (2012). A research of the effect of attitude, achievement, and gender on mathematic education. *Acta Didactica Napocensia, 5*(1), 45-52.

Atkinson, R.D. (2012). Why the current education reform strategy won't work. *Issues in Science and Technology, 28*(3), 29-36.

Atkinson, R.D., Hugo, J., Lundgren, D., Shapiro, M.J., & Thomas, J. (2007). Addressing the STEM challenge by expanding specialty math and science high schools. *Information Technology and Innovation Foundation*.

Atkinson, R.D., & Mayo, M. (2010). Refueling the US innovation economy: Fresh approaches to science, technology, engineering and mathematics (STEM) education.

Bailey, T. (2009). Challenge and opportunity: Rethinking the role and function of developmental education in community college. *New Directions for Community Colleges, 2009*(145), 11-30.

Balcombe, K., & Fraser, I. (2011). A general treatment of 'don't know'responses from choice experiments. *European Review of Agricultural Economics, 38*(2), 171-191.

Bartlett, M.S. (1950). Tests of significance in factor analysis. *British Journal of Mathematical and Statistical Psychology, 3*(2), 77-85.

Basham, J.D., Israel, M., & Maynard, K. (2010). An ecological model of STEM education: Operationalizing STEM for all. *Journal of Special Education Technology, 25*(3), 9-19.

Bauer, D.J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22*(3), 507.

Becker, B.J. (1990). Item characteristics and gender differences on the SAT-M for mathematically able youths. *American Educational Research Journal, 27*(1), 65-87.

Beede, D.N., Julian, T.A., Langdon, D., McKittrick, G., Khan, B., & Doms, M.E. (2011). Women in STEM: A gender gap to innovation.

Belfield, C.R., & Crosta, P.M. (2012). Predicting success in college: The importance of placement tests and high school transcripts. CCRC Working Paper No. 42. *Community College Research Center, Columbia University*.

Bennett, D.A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health, 25*(5), 464-469.

Betts, J.R., Hahn, Y., & Zau, A.C. (2011). *Does diagnostic math testing improve student learning?* : Public Policy Instit. of CA.

Birnbaum, A.L. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The cognitive domain* (Vol. 19). New York: David McKay Co Inc.

Bloom, B.S., & Sosniak, L.A. (1985). *Developing talent in young people*: Ballantine Books.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2 ed.). New York, NY: Routledge.

Bonett, D.G., & Price, R.M. (2005). Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics, 30*(2), 213-225.

Bradley, K.D. (2018). Item characteristic curves for five dichotomous items. In Models for Measuring (Ed.), *PowerPoint*. University of Kentucky: College of Education.

Breiner, J.M., Harkness, S.S., Johnson, C.C., & Koehler, C.M. (2012). What is STEM? A discussion about conceptions of STEM in education and partnerships. *School Science and Mathematics, 112*(1), 3-11.

Bressoud, D.M., Mesa, V., & Rasmussen, C.L. (2015). *Insights and recommendations from the MAA national study of college calculus*: MAA Press.

Bridgeman, B., & Wendler, C. (1989). Prediction of grades in college mathematics courses as a component of the placement validity of SAT-Mathematics scores. *ETS Research Report Series, 1982*(2).

Bridgeman, B., & Wendler, C. (1991). Gender differences in predictors of college mathematics performance and in college mathematics course grades. *Journal of Educational Psychology, 83*(2), 275.

Brown, R.S., & Niemi, D.N. (2007). Investigating the alignment of high school and community college assessments in California. National Center Report #07-3: National Center for Public Policy and Higher Education.

Brown, S.W., Renzulli, J.S., Gubbins, E.J., Siegle, D., Zhang, W., & Chen, C.-H. (2005). Assumptions underlying the identification of gifted and talented students. *Gifted Child Quarterly, 49*(1), 68-79.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of psychology, 3*(3), 296-322.

Browne, M.W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230-258. doi: 10.1177/0049124192021002005

Burton, N.W. (1996). Have changes in the SAT affected women's mathematics performance? *Educational Measurement: Issues and Practice, 15*(4), 5-9.

Burton, R.F. (2005). Multiple-choice and true/false tests: myths and misapprehensions. *Assessment & Evaluation in Higher Education, 30*(1), 65-72.

Bybee, R.W. (2010). Advancing STEM education: A 2020 vision. *Technology and engineering teacher, 70*(1), 30.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61*(2), 309-329.

Cai, L., Thissen, D., & duToit, S.H.C. (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.

Callahan, C.M. (2005). Identifying gifted students from underrepresented populations. *Theory Into Practice, 44*(2), 98-104.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin, 56*(2), 81.

Capra, M.G. (2005). *Factor analysis of card sort data: An alternative to hierarchical cluster analysis.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Carroll, J.D., & Arabie, P. (1980). Multidimensional scaling. *Annual review of psychology, 31*(1), 607-649.

Catsambis, S. (1994). The path to math: Gender and racial-ethnic differences in mathematics participation from middle school to high school. *Sociology of Education*, 199-215.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate behavioral research, 1*(2), 245-276.

Chen, S.-F., Wang, S., & Chen, C.-Y. (2012). A simulation study using EFA and CFA programs based the impact of missing data on test dimensionality. *Expert Systems with Applications, 39*(4), 4026-4031.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289.

Clark, L.A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319. doi: 10.1037/1040-3590.7.3.309

Cliff, N. (1984). An improved internal consistency reliability estimate. *Journal of Educational Statistics*(2), 151. doi: 10.2307/1164718

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd: Hillsdale, NJ: erlbaum.

College Board. (2016). Instructions for concording new SAT scores to old SAT scores. In C. Tables (Ed.).

College Board. (2018a). *"How the SAT Is Scored"*. Retrieved from https://collegereadiness.collegeboard.org/sat/scores/how-sat-is-scored

College Board. (2018b). SAT Suite of Assessments. 2018

Comrey, A.L., & Lee, H.B. (1992). A first course in factor analysis.

Cook, D.A., & Beckman, T.J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *The American journal of medicine, 119*, 166.e167-166.e116.

Cook, R.D. (1977). Detection of influential observation in linear regression. *Technometrics, 19*(1), 15-18.

Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology, 78*(1), 98-104.

Costello, A.B., & Osborne, J.W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research & evaluation, 10*(7), 1-9.

Crist, C., Jacquart, M., & Shupe, D.A. (2002). Improving the performance of high school students: Focusing on connections and transitions taking place in Minnesota.

Crocker, L., & Algina, J. (2008). *Introduction to Classical & Modern Test Theory*: Cengage Learning.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika, 16*(3), 297-334.

Cronbach, L.J. (1971). Test validation. *Educational Measurement*.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological bulletin, 52*(4), 281.

Culbertson, M.J. (2011). Is it wrong? Handling missing responses in IRT.

Custer, M., Sharairi, S., & Swift, D. (2012). *A comparison of scoring options for omitted and not-reached items through the recovery of IRT parameters when utilizing the Rasch Model and Joint Maximum Likelihood Estimation*. Paper presented at the National Council on Measurement in Education, Vancouver, British Columbia.

D'Ambrosio, A., Amodio, S., Iorio, C., Pandolfo, G., & Siciliano, R. (2020). Adjusted concordance index: An extension of the adjusted rand index to fuzzy partitions. *Journal of Classification*, 1-17.

D'Agostino, J., Karpinski, A., & Welsh, M. (2011). A method to examine content domain structures. *International Journal of Testing, 11*(4), 295-307.

Davis, J.D., & Shih, J.C. (2007). Secondary options and post-secondary expectations: Standards-based mathematics programs and student achievement on college mathematics placement exams. *School Science and Mathematics, 107*(8), 336-346.

Davison, M.L., & Skay, C.L. (1991). Multidimensional scaling and factor models of test and item responses. *Psychological Bulletin, 110*(3), 551.

De Ayala, R.J. (2009). *The theory and practice of item response theory*: Guilford Publications.

De Ayala, R.J., Plake, B.S., & Impara, J.C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of educational measurement, 38*(3), 213-234.

DeLacy, M. (2004). The 'No Child'law's biggest victims? An answer that may surprise. *Education Week, 23*(41), 40.

Denny, J.K., Nelson, D.G., & Zhao, M.Q. (2012). Creating and analyzing the effectiveness of a mathematics placement policy for new freshmen. *PRIMUS, 22*(3), 177-185.

Deville, C.W., & Prometric, S. (1996). An empirical link of content and construct validity evidence. *Applied Psychological Measurement, 20*(2), 127-139.

Dimitrov, D.M. (2013). *Quantitative Research in Education : Intermediate & Advanced Methods*: Oceanside, NY : Whittier Publications, [2013]

2013 edition.

Ding, C.S., Song, K., & Richardson, L.I. (2006). Do mathematical gender differences continue? A longitudinal study of gender difference and excellence in mathematics performance in the US. *Educational Studies, 40*(3), 279-295.

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics-Physics Education Research, 5*(2), 020103.

Doolittle, A.E., & Cleary, T.A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of educational measurement, 24*(2), 157-166.

Ebel, R.L. (1956). Obtaining and reporting evidence on content validity. *Educational and Psychological Measurement, 16*(3), 269-282.

Ebel, R.L. (1965). *Measuring educational achievement*: Englewood Cliffs.

Edelen, M.O., & Reeve, B.B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(1), 5.

Educational Testing Service. (1989). The Gender Gap (Vol. 2). Princeton N. J. Policy Information Center: ETS Policy Notes.

Ellison, G., & Swanson, A. (2018). Dynamics of the gender gap in high math achievement: National Bureau of Economic Research.

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*(1), 103.

Erwin, J.O., & Worrell, F.C. (2012). Assessment practices and the underrepresentation of minority students in gifted and talented education. *Journal of Psychoeducational Assessment, 30*(1), 74-87.

Farley, M. (2007). *Assessing the internal structure of math exam scores with confirmatory factor analysis.* (PhD), University of New Mexico.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Feldhusen, J.F., & Jarwan, F. (1995). Predictors of academic success at state-supported residential schools for mathematics and science: A validity study. *Educational and Psychological Measurement, 55*(3), 505-512.

Feldt, L.S. (1969). A test of the hypothesis that cronbach's alpha or kuder-richardson coefficent twenty is the same for two tests. *Psychometrika, 34*(3), 363-373.

Feldt, L.S., Woodruff, D.J., & Salih, F.A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*(1), 93-103.

Ferketich, S. (1991). Focus on psychometrics. Aspects of item analysis. *Research in Nursing & Health, 14*(2), 165-168.

Flores, R.A. (2007). *Psychometric analyses of a mathematics placement exam.*

Foley-Peres, K., & Poirier, D. (2008). College math assessment: SAT scores vs. college math placement scores. *Educational Research Quarterly, 32*(2), 41.

Ford, D.Y. (1998). The underrepresentation of minority students in gifted education: Problems and promises in recruitment and retention. *The Journal of Special Education, 32*(1), 4-14.

Ford, D.Y., & Grantham, T.C. (2003). Providing access for culturally diverse gifted students: From deficit to dynamic thinking. *Theory into Practice, 42*(3), 217-225.

Frisbie, D.A. (1982). Methods of evaluating course placement systems. *Educational Evaluation and Policy Analysis, 4*(2), 133-140.

Frisbie, D.A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice, 7*(1), 25-35.

Gallagher, A.M., & De Lisi, R. (1994). Gender differences in Scholastic Aptitude Test: Mathematics problem solving among high-ability students. *Journal of Educational Psychology, 86*(2), 204.

Gallagher, J.J. (2004). No Child Left Behind and gifted education. *Roeper Review, 26*(3), 121-123.

Garner, M., & Engelhard Jr., G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12*(1), 29-51.

Article 14A. Gifted and Talented Children § 14A-20 (2005).

Gherasim, L.R., Butnaru, S., & Mairean, C. (2013). Classroom environment, achievement goals and maths performance: Gender differences. *Educational Studies, 39*(1), 1-12.

Giguère, G. (2006). Collecting and analyzing data in multidimensional scaling experiments: A guide for psychologists using SPSS. *Tutorials in Quantitative Methods for Psychology, 2*(1), 27-38.

Gonzalez, H.B., & Kuenzi, J.J. (2012). *Science, technology, engineering, and mathematics (STEM) education: A primer*.

Gorsuch, R.L. (1983). Factor Analysis, 2nd. *Hillsdale, NJ: LEA*.

Grant, J.S., & Davis, L.L. (1997). Selection and use of content experts for instrument development. *Research in Nursing & Health, 20*(3), 269-274.

Grey, T. (2004). Merrow report on gifted education. *The NewsHour with Jim Lehrer*.

Grubb, W.N., & Worthen, H. (1999). *Honored but invisible: An inside look at teaching in community colleges*: Psychology Press.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255-282.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika, 19*(2), 149-161.

Haeck, W., Yeld, N., Conradie, J., Robertson, N., & Shall, A. (1997). A developmental approach to mathematics testing for university admissions and course placement. *Educational Studies in Mathematics, 33*(1), 71-91.

Hahs-Vaughn, D.L. (2016). *Applied multivariate statistical concepts*: Taylor & Francis.

Hair Jr, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (1995). *Multivariate data analysis: With readings* (4 ed.). New Jersey: Prentice Hall.

Hambleton, R.K., Swaminathan, H., Algina, J., & Coulson, D.B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48*(1), 1-47.

Hanna, G.S. (1974). *An investigation of the "Don't Know" option in formative evaluation*.

Harman, H.H. (1976). *Modern factor analysis*: University of Chicago press.

Harris, W.G. (2003). Current issues in educational assessment: The test publisher's role.

Haynes, S.N., Richard, D.C.S., & Kubany, E.S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238.

Hays, R.D., Morales, L.S., & Reise, S.P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*(9 Suppl), II28.

Henson, R.K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*(3), 177.

Hoaglin, D.C., & Welsch, R.E. (1978). The hat matrix in regression and ANOVA. *The American Statistician, 32*(1), 17-22.

Hoand, S.M. (2008). Non-metric multidimensional scaling (MDS).

Hoffman, M., Steinley, D., & Brusco, M.J. (2015). A note on using the adjusted Rand index for link prediction in networks. *Social networks, 42*, 72-79.

Hope, D., Adamson, K., McManus, I., Chis, L., & Elder, A. (2018). Using differential item functioning to evaluate potential bias in a high stakes postgraduate knowledge based assessment. *BMC medical education, 18*(1), 64.

Hopfenbeck, T.N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research, 62*(3), 333-353.

Hopkins, K.D. (1998). *Educational and psychological measurement and evaluation*: ERIC.

Hoyt, J.E., & Sorensen, C.T. (2001). High school preparation, placement testing, and college remediation. *Journal of Developmental Education, 25*(2), 26.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193-218.

Hyde, J.S., Lindberg, S.M., Linn, M.C., Ellis, A.B., & Williams, C.C. (2008). Gender similarities characterize math performance. *Science, 321*(5888), 494-495.

Education and Cultural Resources § 227 (2014).

Impara, J.C., Plake, B.S., & Fager, J.J. (1993). Teachers' assessment background and attitudes toward testing. *Theory into Practice, 32*(2), 113-117.

International Mathematical Olympiad Foundation. (2018). International Mathematical Olympiad. Retrieved 8/23/2018, 2018, from https://www.imo-official.org/default.aspx

Jaworska, N., & Chupetlovska-Anastasova, A. (2009). A review of multidimensional scaling (MDS) and its utility in various psychological domains. *Tutorials in Quantitative Methods for Psychology, 5*(1), 1-10.

Johnson, S.C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*(3), 241-254.

Joint Committee on Testing Practices. (2005). Code of Fair Testing Practices in Education (Revised), 23.

Jones, B.M. (2009). Profiles of state-supported residential math and science schools. *Journal of Advanced Academics, 20*(3), 472-501.

Jöreskog, K.G. (2003). Factor analysis by MINRES. *To the memory of Harry Harman and Henry Kaiser*.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*(1), 141-151.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527.

Katzenmeyer, C., & Lawrenz, F. (2006). National Science Foundation perspectives on the nature of STEM program evaluation. *New Directions for Evaluation, 2006*(109), 7-18.

Keith, T.Z. (2014). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling*: Routledge.

Kim, J.-O., & Mueller, C.W. (1978). *Factor analysis: Statistical methods and practical issues* (Vol. 14): Sage.

Kimberlin, C.L., & Winetrstein, A.G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy, 65*(23).

Kirisci, L., Tarter, R.E., & Hsu, T.-C. (1994). Fitting a two-parameter logistic item response model to clarify the psychometric properties of the Drug Use Screening Inventory for adolescent alcohol and drug abusers. *Alcoholism: Clinical And Experimental Research, 18*(6), 1335-1341.

Kline, P. (1986). A handbook of test construction: Introduction to psychometric design, Methuen & Co. *Ltd, London*.

Kossack, C.F. (1942). Mathematics placement at the University of Oregon. *The American Mathematical Monthly, 49*(4), 234-237.

Kowski, L.E. (2013). Does high school performance predict college math placement? *Community College Journal of Research and Practice, 37*(7), 514-527.

Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*(1), 1-27.

Kruskal, J.B., & Wish, M. (1978). *Multidimensional scaling* (Vol. 11): Sage.

Kuder, G.F., & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika*(3), 151. doi: 10.1007/BF02288391

Labovitz, E.M. (1975). Race, SES contexts and fulfillment of college aspirations. *Sociological Quarterly, 16*(2), 241-249.

Labovitz, S. (1968). Criteria for selecting a significance level: A note on the sacredness of. 05. *The American Sociologist*, 220-222.

Latterell, C.M., & Regal, R.R. (2003). Are placement tests for incoming undergraduate mathematics students worth the expense of administration? *Problems, Resources, and Issues in Mathematics Undergraduate Studies, 13*(2), 152-164.

Lee, S., & Kim, S. (2017). Detecting differential item functioning based on gender: Field of mathematics in the TIMSS 2007. *Journal of Fisheries and Marine Sciences Education, 29*(3), 757-766.

Li, X., & Sireci, S.G. (2013). A new method for analyzing content validity data using multidimensional scaling. *Educational and Psychological Measurement, 73*(3), 365-385.

Lindberg, S.M., Hyde, J.S., Petersen, J.L., & Linn, M.C. (2010). New trends in gender and mathematics performance: a meta-analysis. *Psychological bulletin, 136*(6), 1123.

Linn, R.L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher*(9), 4.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological reports, 3*(3), 635-694.

Lord, F.M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*(1), 95-100.

Lorenzo-Seva, U., & Ferrando, P.J. (2012). TETRA-COM: A comprehensive SPSS program for estimating the tetrachoric correlation. *Behavior Research Methods, 44*(4), 1191-1196.

Lorenzo-Seva, U., & Ferrando, P.J. (2020). Not Positive Definite Correlation Matrices in Exploratory Item Factor Analysis: Causes, Consequences and a Proposed Solution. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-10.

Ludlow, L.H., & O'Leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement, 59*(4), 615-630.

Madison, B.L., Linde, C.S., Decker, B.R., Rigsby, E.M., Dingman, S.W., & Stegman, C.E. (2015). A study of placement and grade prediction in first college mathematics courses. *Problems, Resources, and Issues in Mathematics Undergraduate Studies, 25*(2), 131-157.

Margalit, D., & Rabinoff, J. (2018). Interactive Linear Algebra. *Georgia Institute of Technology*.

Marsh, E.J., Roediger, H.L., Bjork, R.A., & Bjork, E.L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14*(2), 194-199.

Marshall, J.P., & Allen, B.D. (2000). The development and implementation of a mathematics placement program.

Marshall, S.P., McGee, G.W.M., McLaren, E., & Veal, C.C. (2011). Discovering and developing diverse STEM talent: Enabling academically talented urban youth to flourish. *Gifted Child Today, 34*(1), 16-23.

Martinkova, P. (2016). Test and item analysis. In L. s. m.-u. DIF (Ed.).

Martone, A., & Sireci, S.G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332-1361.

Marwick, J.D. (2002). Alternative methods of mathematics placement. *The Community College Enterprise, 8*(2), 41.

Marwick, J.D. (2004). Charting a path to success: The association between institutional placement policies and the academic success of Latino students. *Community College Journal of Research and Practice, 28*(3), 263-280.

Mattern, K.D., & Packman, S. (2009). Predictive validity of ACCUPLACER scores for course placement: A meta-analysis. New York: The College Board.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 11*(3), 71-101.

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2 n contingency tables: A unified framework. *Journal of the American Statistical Association, 100*(471), 1009-1020.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*(4), 305-328.

McClain, M.-C., & Pfeiffer, S. (2012). Identification of gifted students in the United States today: A look at state definitions, policies, and practices. *Journal of Applied School Psychology, 28*(1), 59-88.

McDaniel, M.A., Roediger, H.L., & McDermott, K.B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*(2), 200-206.

McFate, C., & Olmsted III, J. (1999). Assessing student preparation through placement tests. *Journal of Chemical Education, 76*(4), 562.

Mead, A. (1992). Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 27-39.

Medhanie, A.G., Dupuis, D.N., LeBeau, B., Harwell, M.R., & Post, T.R. (2012). The role of the ACCUPLACER mathematics placement test on a student's first college mathematics course. *Educational and Psychological Measurement, 72*(2), 332-351.

Melguizo, T., Hagedorn, L.S., & Cypers, S. (2008). Remedial/developmental education and the cost of community college transfer: A Los Angeles County sample. *The Review of Higher Education, 31*(4), 401-431.

Melguizo, T., Kosiewicz, H., Prather, G., & Bos, J. (2014). How are community college students assessed and placed in developmental math? Grounding our understanding in reality. *The Journal of Higher Education, 85*(5), 691-722.

Mendoza, C. (2006). Inside today's classrooms: Teacher voices on No Child Left Behind and the education of gifted children. *Roeper Review, 29*(1), 28-31.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3 ed., pp. 13-104). New York, NY: American Council on Education and Macmillan Publihing Company.

Messick, S. (1995). Standards of validity and the validity of standards in performance asessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Mondak, J.J. (2001). Developing valid knowledge scales. *American Journal of Political Science*, 224-238.

Morgan, D.L., & Michaelides, M.P. (2005). Setting cut scores for college placement *Connecting Students to College Success*. New York: The College Board.

Moroke, N.D. (2014). Profiling some of the dire household debt determinants: a metric multidimensional scaling approach. *Journal of Economics and Behavioral Studies, 6*(11), 858-867.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (2007). *Rising above the gathering storm: Energizing and employing*

America for a brighter economic future. Washington, DC: The National Academies Press.

National Association for Gifted Children. (2018). What is giftedness? Retrieved 07/05/2018

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. *The Elementary School Journal, 84*(2), 113-130.

National Consortium of Secondary STEM Schools. (2018). Institutional members. Retrieved 07/03/2018, 2018, from http://ncsss.org/membership/institutional-members

Ngo, F., & Kwon, W.W. (2015). Using multiple measures to make math placement decisions: Implications for access and success in community colleges. *Research in Higher Education, 56*(5), 442-470.

Ngo, F., & Melguizo, T. (2016). How can placement policy improve math remediation outcomes? Evidence from experimentation in community colleges. *Educational Evaluation and Policy Analysis, 38*(1), 171-196.

Nitko, A.J., & Brookhart, S.M. (2011). *Educational assessment of students* (P. A. Smith Ed. 6 ed.). Boston, MA: Pearson.

Noble, J.P., Schiel, J.L., & Sawyer, R.L. (2003). Assessment and college course placement: Matching students with appropriate instruction. *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators*, 17.

Norman, K.W., Medhanie, A.G., Harwell, M.R., Anderson, E., & Post, T.R. (2011). High school mathematics curricula, university mathematics placement recommendations, and student university mathematics performance. *Problems, Resources, and Issues in Mathematics Undergraduate Studies, 21*(5), 434-455.

Nunnally, J.C., & Bernstein, I.H. (1978). Psychometric theory.

Nunnally, J.C., & Bernstein, I.H. (1994). Validity. *Psychometric Theory*, 99-132.

O'Neill, K.A., & McPeek, W.M. (1993). *Item and test characteristics that are associated with differential item functioning.* Paper presented at the Differential Item Functioning: Theory and Practice, Oct, 1989, Educational Testing Service, Princeton, NJ, US.

Oakes, J. (1990). Chapter 3: Opportunities, achievement, and choice: Women and minority students in science and mathematics. *Review of Research in Education, 16*(1), 153-222.

Olszewski-Kubilius, P. (2009). Special schools and other options for gifted STEM students. *Roeper Review, 32*(1), 61-70.

Onwuegbuzie, A.J., & Daniel, L.G. (2002). A framework for reporting and interpreting internal consistency reliability estimates. *Measurement and Evaluation in Counseling and Development, 35*(2), 89.

Oosterhof, A. (2001). *Classroom applications of educational measurement*.

Organisation for Economic Co-operation and Development. (2018). Programme for International Student Assessment. Retrieved 8/23/2018, 2018, from http://www.oecd.org/pisa/

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50-64. doi: 10.1177/01466216000241003

Osborne, J.W. (2000). Prediction in multiple regression. *Practical Assessment, Research & Evaluation, 7*(2), 1-9.

Parker, M. (2005). Placement, retention, and success: A longitudinal study of mathematics and retention. *The Journal of General Education, 54*(1), 22-40.

Pedro, J.D., Wolleat, P., Fennema, E., & Becker, A.D. (1981). Election of high school mathematics by females and males: Attributions and attitudes. *American Educational Research Journal, 18*(2), 207-218.

Petrocelli, J.V. (2003). Hierarchical multiple regression in counseling research: Common problems and possible remedies. *Measurement and Evaluation in Counseling and Development, 36*(1), 9-22.

Pett, M.A., Lackey, N.R., & Sullivan, J.J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*: Sage.

Pfeiffer, S.I., Overstreet, J.M., & Park, A. (2010). The state of science and mathematics education in state-supported residential academies: A nationwide survey. *Roeper Review, 32*(1), 25-31.

Pike, G.R. (1991). The effects of background, coursework, and involvement on students' grades and satisfaction. *Research in Higher Education, 32*(1), 15-30.

Pike, G.R., & Saupe, J.L. (2002). Does high school matter? An analysis of three methods of predicting first-year grades. *Research in Higher Education, 43*(2), 187-207.

Plake, B.S., & Impara, J.C. (1997). Teacher assessment literacy: What do teachers know about assessment? *Handbook of Classroom Assessment* (pp. 53-68): Elsevier.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement, 74*(3), 423-452.

President's Council of Advisors on Science and Technology. (2010). *Prepare and inspire: K-12 education in science, technology, engineering, and math (STEM) for America's future*: Executive Office of the President,.

Prieto, G., & Delgado, A.R. (1999). The effect of instructions on multiple-choice test scores. *European Journal of Psychological Assessment, 15*(2), 143-150.

Pugh, C.M., & Lowther, S. (2004). *College math performance and last high school math course.* Paper presented at the Annual Conference of the Southern Association for Institutional Research.

Pyryt, M.C. (2000). Talent development in science and technology. *International Handbook of Giftedness andTalent*, 427-437.

Quaigrain, K., & Arhin, A.K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education, 4*(1), 1301013.

Quilter, S.M., & Gallini, J.K. (2000). Teachers' assessment literacy and attitudes. *The Teacher Educator, 36*(2), 115-131.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846-850.

Raymond, M.R. (1989). Applications of multidimensional scaling to research in the health professions. *Evaluation & the Health Professions, 12*(4), 379-408.

Reilly, D., Neumann, D.L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology, 107*(3), 645.

Renzulli, J.S., & Smith, L.H. (1977). Two approaches to identification of gifted students. *Exceptional Children, 43*(8), 512-518.

Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research, 14*(1), 57-74.

Robitzsch, A., & Rupp, A.A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement, 69*(1), 18-34.

Rose, N., Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT). In D. Eignor (Ed.), *Research Report*. Princeton, New Jersey: Educational Testing Services.

Rossi, P.H., Lipsey, M.W., & Freeman, H.E. (2003). *Evaluation: A systematic approach*: Sage publications.

Roth, J., Crans, G.G., Carter, R.L., Ariet, M., & Resnick, M.B. (2000). Effect of high school course-taking and grades on passing a college placement test. *The High School Journal, 84*(2), 72-87.

Rubio, D.M., Berg-Weger, M., Tebb, S.S., Lee, E.S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27*(2), 94-104.

Rueda, N.G., & Sokolowski, C. (2004). Mathematics placement test: Helping students succeed. *The Mathematics Educator, 14*(2).

Ryan, K.A. (2018). *An investigation of pre-service teacher assessment literary and assessment confidence: Measure development and EdTPA performance.* Kent State University.

Sälzer, C., & Roczen, N. (2018). Assessing global competence in PISA 2018: Challenges and approaches to capturing a complex construct. *International Journal of Development Education and Global Learning, 10*(1), 5-20.

Sarstedt, M., & Mooi, E. (2014). Cluster analysis *A concise guide to market research* (pp. 273-324): Springer.

Sawyer, R. (1996). Decision theory models for validating course placement tests. *Journal of Educational Measurement, 33*(3), 271-290.

Schaefer, L., Raymond, M., & Stamps White, A. (1992). A comparison of two methods for structuring performance domains. *Applied Measurement in Education, 5*(4), 321-335. doi: 10.1207/s15324818ame0504_3

Schafer, W.D. (1993). Assessment literacy for teachers. *Theory into Practice, 32*(2), 118-126.

Schmeiser, C.B. (1995). Code of professional responsibilities in educational measurement. *Educational Measurement: Issues and Practice, 14*(3), 17-24.

Schmidt, F.L., & Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*(5), 529.

Schmitz, C.C., & delMas, R.C. (1991). Determining the validity of placement exams for developmental college curricula. *Applied Measurement in Education, 4*(1), 37-52. doi: 10.1207/s15324818ame0401_4

Schuelka, M.J. (2013). Excluding students with disabilities from the culture of achievement: The case of the TIMSS, PIRLS, and PISA. *Journal of Education Policy, 28*(2), 216-230.

Schumacher, P.A., & Smith, R.M. (2008). A comparison of placement in first-year university mathematics courses using paper and online administration of a placement test. *International Electronic Journal of Mathematics Education, 3*(3), 193-202.

Scott-Clayton, J. (2012). Do high-stakes placement exams predict college success? . *Community College Research Center, Columbia University, Working Paper No. 41*.

Scott, C.E. (2012). An investigation of science, technology, engineering and mathematics (STEM) focused high schools in the US. *Journal of STEM Education: Innovations and Research, 13*(5).

Shaw, P.J. (1997). *An analysis and evaluation of policies and practices of student placement into college algebra classes at Paradise Valley Community College.* Nova Southeastern University.

Sheel, S.J., Vrooman, D., Renner, R.S., & Dawsey, S.K. (2001). *A comparison of neural networks and classical discriminant analysis in predicting students' mathematics placement examination scores.* Paper presented at the International Conference on Computational Science.

Siegler, R.S., Duncan, G.J., Davis-Kean, P.E., Duckworth, K., Claessens, A., Engel, M., . . . Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological Science, 23*(7), 691-697.

Sireci, S.G. (1998a). The construct of content validity. *Social Indicators Research, 45*(1), 83-117.

Sireci, S.G. (1998b). Gathering and analyzing content validity data. *Educational Assessment, 5*(4), 299-321.

Sireci, S.G., & Geisinger, K.F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement, 16*(1), 17-31.

Sireci, S.G., & Geisinger, K.F. (1995). Using subject-matter experts to assess content representation: An MDS analysis. *Applied Psychological Measurement, 19*(3), 241-255.

Sirin, S.R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research, 75*(3), 417-453.

Slavin, R.E. (2007). *Educational research in an age of accountability*: Pearson Boston, MA.

Smith, M.L., & Fey, P. (2000). Validity and accountability in high-stakes testing. *Journal of Teacher Education, 51*(5), 334-344.

Sondergeld, T.A. (2014). Closing the gap between STEM teacher classroom assessment expectations and skills. *School Science and Mathematics, 114*(4), 151-153.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271-295.

Stedman, L.C. (1994). The Sandia report and US achievement: An assessment. *The Journal of Educational Research, 87*(3), 133-146.

Steiger, J.H. (2016). Notes on the Steiger–Lind (1980) handout. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(6), 777-781.

Steiger, J.H., & Lind, J.C. (1980). *Statistically based tests for the number of common factors*. Paper presented at the Psychometrika Society Meeting, Iowa City, IA.

Steinberg, J. (2010). Exploring the dimensionality of large-scale standardized educational assessments using PROC FACTOR. *NESUG 2010 Statistics and Analysis*, 1-8.

Stevens, J.P. (2012). *Applied multivariate statistics for the social sciences*: Routledge.

Steyvers, M. (2002). Multidimensional scaling. *Encyclopedia of Cognitive Science*.

Stiggins, R.J. (1991). Assessment literacy. *Phi Delta Kappan, 72*(7), 534-539.

Stiggins, R.J., & Bridgeford, N.J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*(4), 271-286.

Stone, C.A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40*(4), 331-352.

Subotnik, R.F., Duschl, R.A., & Selmon, E.H. (1993). Retention and attrition of science talent: A longitudinal study of Westinghouse Science Talent Search winners. *International Journal of Science Education, 15*(1), 61-72.

Subotnik, R.F., Olszewski-Kubilius, P., & Worrell, F.C. (2011). Rethinking giftedness and gifted education: A proposed direction forward based on psychological science. *Psychological Science in the Public Interest, 12*(1), 3-54.

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics*: Allyn & Bacon/Pearson Education.

Tai, R.H., Liu, C.Q., Maltese, A.V., & Fan, X. (2006). Planning early for careers in science. *Life Science, 1*, 0.2.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*: American Psychological Association.

Trochim, W.M.K. (1989). An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning, 12*(1), 1-16.

Tucker, L.R., & MacCallum, R.C. (1997). Exploratory factor analysis. *Unpublished manuscript, Ohio State University, Columbus*.

U.S. Department of Education, Office of the Secretary, & Office of Public Affairs. (2004). A guide to education and No Child Left Behind. *Washington, DC: US Department of Education. Retrieved April, 14*, 2006.

Uebersax, J.S. (2006a). Factor analysis and SEM with tetrachoric and polychoric correlations. Retrieved October 23, 2018, 2018, from https://www.john-uebersax.com/stat/sem.htm

Uebersax, J.S. (2006b). Introduction to the tetrachoric and polychoric correlation coefficients. Retrieved October 23, 2018, 2018, from http://www.john-uebersax.com/stat/tetra.htm

Wampold, B.E., & Freund, R.D. (1987). Use of multiple regression in counseling psychology research: A flexible data-analytic strategy. *Journal of Counseling Psychology, 34*(4), 372-382. doi: 10.1037/0022-0167.34.4.372

Wang, M.-T., & Degol, J.L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational psychology review, 29*(1), 119-140.

Warrens, M.J. (2008). On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. *Journal of Classification, 25*(2), 177-183.

Wattenbarger, J.L., & McLeod, N. (1989). Placement in the mathematics curriculum: What are the keys? *Community College Review, 16*(4), 17-21.

Whaley, A.L., & Longoria, R.A. (2009). Preparing card sort data for multidimensional scaling analysis in social psychological research: A methodological approach. *The Journal of Social Psychology, 149*(1), 105-115.

White, G.W., Stepney, C.T., Hatchimonji, D.R., Moceri, D.C., Linsky, A.V., Reyes-Portillo, J.A., & Elias, M.J. (2016). The increasing impact of socioeconomics and

race on standardized academic test scores across elementary, middle, and high school. *American Journal of Orthopsychiatry, 86*(1), 10.

Widaman, K.F. (2006). III. Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development, 71*(3), 42-64.

Wiersma, W., & Jurs, S.G. (2009). *Research methods in education: An introduction* (9 ed.). Boston, MA: Pearson.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *The Phi Delta Kappan, 70*(9), 703-713.

Wilkinson, L. (2002). Multidimensional scaling. *Systat, 10*(2), 119-145.

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine, 8*(3).

Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice, 37*(1), 5-20.

Zhang, X. (2013). The I don't know option in the Vocabulary Size Test. *Teachers of English to Speakers of Other Languages Quarterly, 47*(4), 790-811.