CONSTRUCT RELEVANT AND IRRELEVANT VARIABLES IN

MATH PROBLEM SOLVING ASSESSMENT

by

LISA E. BIRK

A DISSERTATION

Presented to the Department of Educational Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Education

June 2013

DISSERTATION APPROVAL PAGE

Student: Lisa E. Birk

Title: Construct Relevant and Irrelevant Variables in Math Problem Solving Assessment

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Education degree in the Department of Educational Methodology, Policy, and Leadership by:

Dr. Gerald Tindal	Chairperson
Dr. Julie Alonzo	Core Member
Dr. Gina Biancarosa	Core Member
Dr. McKay Sohlberg	Institutional Representative

and

Kimberly Andrews Espy	Vice President for Research and Innovation; Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2013

DISSERTATION ABSTRACT

Lisa E. Birk

Doctor of Education

Department of Educational Methodology, Policy, and Leadership

June 2013

Title: Construct Relevant and Irrelevant Variables in Math Problem Solving Assessment

In this study, I examined the relation between various construct relevant and irrelevant variables and a math problem solving assessment. I used independent performance measures representing the variables of mathematics content knowledge, general ability, and reading fluency. Non-performance variables included gender, socioeconomic status, language proficiency and special education qualification. Using a sequential regression and commonality analysis, I determined the amount of variance explained by each performance measure on the Oregon state math assessment in third grade. All variables were independently predictive of math problem solving scores, and used together, they explained 58% score variance. The math content knowledge measure explained the most variance uniquely (12%), and the measures of math content and general ability explained the most variance commonly (16%). In the second analysis, I investigated whether additional variance was explained once student demographic characteristics were controlled and how this affected the unique variance explained by each independent performance measure. By controlling for demographics, the model explained slightly more than 1% additional variance in math scores. The unique variance explained by each independent measure decreased slightly.

This study highlighted the influence of various construct relevant and irrelevant variables on math problem solving scores, including the extent to which a language-free measure of general ability might help to inform likely outcomes. The use of variance partitioning expanded understanding of the unique and common underlying constructs that affect math problem solving assessment. Finally, this study provided more information regarding the influence demographic information has on outcomes related to state math assessments.

CURRICULUM VITAE

NAME OF AUTHOR: Lisa E. Birk

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

      University of Oregon, Eugene
      University of Idaho, Moscow

DEGREES AWARDED:

      Doctor of Education, Educational Leadership, 2013, University of Oregon
      Master of Education, Educational Leadership, 2010, University of Oregon
      Bachelor of Science, Special Education, 2005, University of Idaho
      Bachelor of Science, Elementary Education, 2005, University of Idaho
      Bachelor of Science, Secondary Education: Mathematics-major, French-minor,
          2005, University of Idaho

AREAS OF SPECIAL INTEREST:

      Building Effective Teacher Teams and Positive School Culture
      Educational Assessment Systems
      Diversity in Schools
      Early Intervention and Predictive Variables for Academic Success
      Special Education Identification Systems and Flexible Services
      Mathematics Education

PROFESSIONAL EXPERIENCE:

      Student Services Coordinator at Bear Creek Elementary School, Bend-La Pine
          Schools, 2011-present

      Teacher on Special Assignment (Mathematics and Data Support), Bend-La Pine
          Schools, 2010-2011

      Special Education Teacher at Juniper Elementary School, Bend-La Pine Schools,
          2007-2010

      Special Education Teacher at Fir Grove Elementary School, Roseburg Public
          Schools, 2005-2007

PUBLICATIONS:

Birk, L. (2009). Mathematics Knowledge Development for Special Education Teachers. University of Oregon Scholar's Bank: One-Goal School Improvement Plans. Retrieved from http://hdl.handle.net/1794/10125

ACKNOWLEDGMENTS

To my parents, Bruce and Sandy who are the best teachers I have ever had.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

State assessments in education are intended to measure progress toward proficiency in specific areas of instruction. In mathematics, assessments in each state are developed based on state standards that may include a number of different domains such as (a) measurement, (b) geometry, (c) numbers and operations, and/or (d) algebra (National Council of Teachers of Mathematics [NCTM], 2000). The content standards and embedded domains represent what students must know and be able to do to demonstrate proficiency in mathematics.

Researchers have pointed out that state standards vary widely (Webb, 1999). Thus, a student who demonstrates mathematic proficiency in Oregon will be unlikely to demonstrate the same level of proficiency in Idaho because the assessments are based on different state content standards and proficiency expectations. Additionally, the extent to which current state assessments accurately or adequately measure the standards or domains of interest is a subject of debate (Webb, 1999). Despite the current variability between state content standards in mathematics, problem solving continues to be one of the primary areas of focus in both instruction and assessment (NCTM, 2000, 2006).

In recent years, a group comprised of the National Governors Association Center for Best Practices (NGA Center) and the Council of Chief State School Officers (CCSSO) resolved to eliminate differences and create common standards for use by all states (2010). By May 2012, nearly every state had joined the initiative known as the Common Core State Standards (CCSS). The adoption of the CCSS will mean many changes for states in terms of instruction, focus, and assessment as each adapts to the new

common expectations. However, despite common standards, some variation is likely to continue due to differences in proficiency standards (cut scores) set by each state independently as well as differences in assessment measures. To monitor achievement, states will choose from two major assessment systems created by two different assessment consortia (Center for K-12 Assessment and Performance Management at ETS [ETS], 2010). Although differences may exist in assessments and state designated cut scores, problem solving will remain a constant (NGA Center & CCSSO, 2010). Experts agree that problem solving is and will continue to be a primary focus of instruction in mathematics and is critical for the demonstration of proficiency in the subject area (National Council of Teachers of Mathematics [NCTM], National Council of Supervisors of Mathematics [NCSM], Association of State Supervisors of Mathematics [ASSM], & Association of Mathematics Teacher Educators [AMTE], 2010).

Researchers note that large-scale assessments typically reflect a complex combination of two major constructs: (a) *declarative knowledge* and (b) *developing abilities* in complex tasks (Haladyna & Downing, 2004). State assessments in mathematics are no exception. In order to demonstrate proficiency, students must use information that they know about numbers (*declarative knowledge)* to solve problems in mathematical situations (*a developing ability)*. Mathematical problem solving is a *developing ability* that is difficult to measure using standard assessment systems. In fact, researchers point out that when trying to evaluate proficiency around a complicated construct such as math problem solving, significant limitations exist. They contend that difficult-to-monitor systematic variance will exist within an assessment, despite attempts

to limit its influence (Haladyna & Downing, 2004). One type of this systematic variance is construct irrelevant variance.

Certain skills are clearly related to certain constructs. For example, numerical fluency is a skill that will likely help a student be successful on assessments measuring math proficiency. Such competencies are considered construct relevant because they are clearly related to the measurement topic; however, construct irrelevant variance (CIV) is variance due to the existence of variables that influence an outcome, yet are not otherwise related to the concept measured. In any assessment system, CIV can (and probably does) exist (Haladyna & Downing, 2004). For example, English language proficiency is likely a factor that would influence outcomes on math or science assessment outcomes, yet has little to do with the constructs being measured.

From a research perspective, it would be ideal to eliminate CIV completely; however, this is improbable. In mathematics problem solving, for example, written or spoken language is the medium through which assessment is delivered. Although unrelated directly to ability in math, language proficiency or reading ability may impact math performance outcomes. It is unlikely that large-scale math assessments will change such that language is unnecessary for assessment. This is just one example of how systematic variance continues to exist in the assessment of complex constructs like math problem solving.

As educators work toward student success on state assessments, it is important to identify variables that may impact outcomes on these measures. Further, if these variables can be altered through instruction, teachers will be better able to allocate resources and focus instruction in order to attain better results for student achievement. To do this, one

must quantify, understand, and consider the variance accounted for by various influential variables when interpreting outcomes. Therefore, the purpose of this study is to broaden the identification and understanding of construct relevant and irrelevant variables on math problem solving outcomes as measured by state assessments in mathematics. Educators will more accurately identify proficiencies and make instructional decisions for students in mathematics when they have greater understanding about the degree to which different variables influence state math assessment outcomes. First, we will consider variables relevant to the constructs represented by state math assessments.

Defining the Construct(s) Measured in State Math Assessments

In this age of accountability, state testing programs are of much interest. It is important, however, to remember that they exist not simply to determine whether or not students do well on the test. Rather, state tests are designed as a way to determine if students are on an academic trajectory toward becoming college and career ready (Conley, 2010). In order to demonstrate readiness, students must show proficiency in several different content domains; one of which is mathematics. Mathematics is important not only in daily life, but is also a necessary competency for technological jobs that exist in increasing numbers in today's society (Jitendra, 2005).

Construct is defined as "the concept or characteristic that a test is designed to measure" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, p. 173). Using this definition, presumably, the construct represented by a state math assessment is mathematics. However, mathematics, like all major subject areas, is a multi-dimensional construct and therefore, difficult to teach, learn, and

measure (Haladyna & Downing, 2004). So although state test programs must report different levels of proficiency, proficiency in mathematics is less than clear. What does it mean to be proficient in math?

As previously described, *declarative knowledge* in mathematics refers to skill competency and efficiency (computation), while *developing abilities* refer to the use of efficient skills to solve problems in mathematical situations (math problem solving). State math assessments such as the Oregon Assessment of Knowledge and Skills (OAKS-math) focus primarily on the latter (Oregon Department of Education, Office of Assessment and Information Services [ODE], 2012). Because assessments are to measure progress toward college and career readiness, or practical application, the focus on mathematical problem solving over skill competency makes sense. Math problem solving is a complex idea and logically contains two terms: (a) math (related to content) and (b) problem solving (related to either skill or ability).

Mathematical content knowledge. In 2001, the Mathematics Learning Study Committee of the National Research Council (NRC) identified five strands of mathematics proficiency. They recognized a need for integrated adaptive reasoning, strategic competence, conceptual understanding, productive disposition, and procedural fluency. The National Mathematics Advisory Panel (NMAP, 2008, pp. xvi-xvii) mirrored these conclusions, indicating a balanced need for a *coherent* progression of learning coupled with *proficiency* with key concepts to solve problems (emphasis in original). From a teaching perspective, mathematics would be much easier to teach if only knowledge of key concepts (*declarative knowledge*) was expected; however, because proficiency means that students are able to synthesize the key concepts and use

them to solve problems, skill knowledge is not enough. Further, a hierarchy of skill development in mathematics is not yet clear. The five strands of mathematics proficiency are tightly intertwined at all levels of math learning (NRC, 2001); thus, critical skills and competencies are not easily isolated or measured.

In an effort to support content delivery, the NCTM (2000) outlined what they believed to be the knowledge and skills that students must be able to demonstrate at each grade level. Like the strands identified by the Math Learning Study Committee, these competencies were broken into five standard domains including numbers and operations, algebra, measurement, data analysis and probability, and geometry. In each grade band (K-2, 3-5, 6-8, and high school) the NCTM specified what type of skills a student should master; however, *every* standard area was important (to varying degrees) in *every* grade band (emphasis added). A sixth standard, process, held the same expectation in all grades. This standard stated that students should demonstrate the ability to problem-solve and more specifically, communicate, prove, reason, make connections, and justify in every mathematical task in every grade.

The standards created by the NCTM were influential as states set standards in mathematics, and therefore typically represented (and continue to represent) the content assessed on many state assessments in math. Such is the case in Oregon (ODE, 2012). However, despite the guidance from NCTM, states were not required to use the suggested standards. Therefore, wide variability of state standards existed, which also affected state assessments. According to one study conducted at the Wisconsin Center for Educational Research, the content of statewide mathematics tests appeared quite varied and addressed a number of different domains to different degrees (Webb, 1999). Some states assessed

certain domains more than others and based on Webb's study, the degree to which the number of questions in any domain represented mastery was also in question.

The recent move by the NGA Center and the CCSSO attempts to eliminate, or at least minimize, differences between states in both instruction and assessment. Like standards of the past, the CCSS include a number of different mathematical standards including the process standard of problem solving through reasoning, justification, and communication (Common Core State Standards Initiative, 2010). Additionally, this policy movement includes the development of two common assessments used to measure standard achievement (ETS, 2010).

Because this movement includes nearly all states, it is likely that math content instruction will become more similar among states than when NCTM initially suggested standards. However, the extent to which mathematical content knowledge will be accurately measured on the new common assessments remains to be seen. Current studies in which researchers examine the predictive validity of curriculum based measures of content knowledge to determine likely success on state assessment outcomes provide a foundation for future replication studies using the CCSS assessments. Once the CCSS assessments are in use, researchers can use these previous studies as models to investigate the construct validity of the new assessment systems as well as the construct relevant and irrelevant variance within them. With this knowledge, teachers and researchers will be able to more accurately identify students who are at risk for failure on state assessments and adjust resources and instruction accordingly to support their path toward college and career readiness.

Problem solving ability. As reflected in the math standards creation from the

NCTM and the NGA Center and CCSSO, problem solving in mathematics is a concept that is critical to mathematical success. It includes the ability to reason, model, justify, and communicate mathematical ideas (NCTM, 2000). In this way, problem solving is construct relevant because it is embedded in mathematical content. It is a skill that is developed. In the CCSS, the problem solving concepts are described as the "standards for mathematical practice" (CCSSI, 2010, p. 10). These standards require students to: (a) make sense of problems and persevere in solving them, (b) reason abstractly and quantitatively, and (c) look for and express regularity in reasoning. These standards of problem solving are related to mathematical skill because students use understanding of numbers to solve problems.

Problem solving is also sometimes referred to as ability (Kaufman, 2009). For example, in reflection of the language outlined in the CCSS, words and phrases like *reason*, *communicate*, *make sense*, and *solve problems* are concepts that extend well beyond the subject of mathematics. We reason when we decide what route to take when we go to the grocery store, we communicate with one another in different settings and different ways and we are always trying to make sense of the world around us. In this way, problem solving is not content specific but rather, an important ability that we use in every setting every day. Both concepts of problem solving are important because they relate to mathematical testing outcomes; however, as a skill, problem solving is construct relevant and as an ability, it is construct irrelevant. Because both conceptual frameworks may influence outcomes on large-scale assessments, and they can be uniquely measured as described in the next section, they can be considered separately as different variables of interest.

Content knowledge and problem solving are sub-constructs represented in state math assessments. They are both construct relevant to mathematics proficiency. Other variables that may influence outcomes but are unrelated directly to mathematics are those that are construct irrelevant. These are more difficult to recognize yet still important to identify. Some of these variables are quantified by performance measures and others are inherent student demographic characteristics. In the next section, construct irrelevant performance variables related to mathematical problem solving assessments are described.

Identification of Measurement Construct (Irrelevant) Variables in Math Problem Solving

Another line of research indicates that state testing programs test various dimensions of the skills being targeted but also a number of features that are not relevant to the content area of interest (Abedi & Leon, 1999; Abedi, Leon, & Mirocha, 2003). These are considered to be construct irrelevant and create CIV. Two main types of CIV exist: that which exists within a group and that which exists at the individual level. Categories of group or environmental CIV include test preparation methods, test creation, language load, administration, scoring, and cheating (Haladyna & Downing, 2004). In high stakes assessments, the influence of these variables is mediated by extensive test protocols that cover each of these areas. For example, to assure consistency in administration, test designers often use scripted directions during assessment. Teachers do not create these protocols, but do follow them during test administration. In this way, at the group level, teachers have only an indirect control over potential CIV because they are bound by the protocols designed to support assessment.

9

At the individual level however, teachers have direct control over the interpretations made regarding testing outcomes. CIV for individuals might be from variables like general ability or reading proficiency (Haladyna & Downing, 2004). Other student characteristic variables like language facility, socio-economic status, and disability may influence outcomes also, yet are unrelated to the construct of math problem solving (Abedi, Leon, & Mirocha, 2001). Because teachers have direct control over the influence of CIV at the individual level through assessment interpretation, it is important to understand assessments and their influential variables in depth in order to make accurate decisions about student instruction and intervention. Additionally, according to Haladyna and Downing (2004), more research is needed in this topic area specifically to better understand the influence of verbal abilities and accommodations on assessment outcomes.

The influence of (*g*). Problem solving tends to be a construct that has broad reach and can be conflated with intelligence and ability. It is often viewed as a trait that has permanence and is inherent in people (Kaufman, 2009). The history of this concept (particularly intelligence and ability) in the United States began with the first tests used to operationalize the constructs: Stanford-Binet, Wechsler, and most recently Woodcock-Johnson and Kaufman. Most of these tests purport to measure a general trait that dominates other specific abilities (e.g., motor versus verbal or sequential versus simultaneous processing). This trait, or factor, is described as mental intelligence that underlies performance on any cognitive task (Jensen, 2002).

This content-free concept of problem solving may be important and possibly related to outcomes on state assessments. Researchers agree that this factor, often referred

to as general intelligence, or *g* (Spearman, 1904), exists and has an interesting correlation between cognitive tasks that would otherwise be unrelated. Specifically in education, *g* has received much attention over several years and has been shown to be a reliable predictor for success in various academic areas (Brody, 1992; Spearman, 1904). In mathematics, there is evidence that *g* was highly correlated to math ability outcomes as far back as the early 1900s (Spearman, 1904). Current research indicates a positive correlation between assessments of intelligence and those of math proficiency, and thus problem solving (Fuchs et al., 2006; Hart, Petrill, Plomin, & Thompson, 2009; Mannamaa, Kikas, Peets, & Palu, 2012). The correlations found in studies like these demonstrate a consensus that general intelligence impacts individuals as they complete any cognitive task; and further, the completion of a cognitive task is, at its core, a type of problem solving. However, in mathematics, a well-developed consensus does not exist about the degree to which *g* might uniquely influence high-stakes academic outcomes such as state assessments of math problem solving for the average student.

Some researchers believe that the traditional general intelligence tests using language (verbal or written) are not sensitive to diverse populations (Naglieri & Das, 2002). Nonverbal general ability assessments have emerged as tools, according to their authors, that researchers can use to measure innate problem solving ability (*g)* for all subjects regardless of diverse background or native language (Naglieri, 1997; Naglieri, 2008; Raven & Raven, 2003; Wechsler, 1999; Wechsler & Naglieri, 2006). In correlational studies, researchers demonstrated a correlation between math outcomes on state and other math and reading assessments and outcomes on nonverbal measures of general ability (Fuchs et al., 2005; Fuchs et al., 2006; Naglieri & Ronning, 2000). The

11

correlations found using these types of measures lend support for researchers to further investigate the influence of general intelligence on academic outcomes, particularly in math, while including subjects that represent diverse populations.

With more reliable information regarding the potential link between general ability and math outcomes for all students, researchers and educators can make better decisions as they continue to answer the question of what it means to be proficient in mathematics. By understanding the amount of variance on state math tests that can be attributed to g, teachers will be able to make better instructional decisions to support struggling students in math and researchers will be able to craft more reliable assessment tools to measure proficiency.

The influence of reading. Another potentially influential variable to consider is reading ability. As outlined by Haladyna and Downing (2004), reading is a skill that often is more important than it should be in assessments of math problem solving or other content areas. The impact of reading on mathematics outcomes has been documented from several different angles. For example, Abedi, Lord, Hofstetter, and Baker (2000) found that linguistic modification of math items in assessment decreased the gap between language minority and language majority students. Helwig, Rozek-Tedesco, Tindal, Heath, and Almond (1999) drew similar conclusions. In the study, students were given portions of a math assessment in paper-pencil format and the other portion in video format. Student scores were more positive using video presentations of math problems without the requirement of reading. Tindal, Heath, Hollenbeck, Almond, and Harniss (1998) conducted a study using read-aloud as an accommodation for math assessments and found that students performed better when the reading task was eliminated. Both

studies highlight reading ability as a basic access skill in mathematics for all students, including those representing diverse populations.

One would expect that a disfluent reader would do poorly on a measure of MAZE reading (a short measure of reading comprehension); however, it is less obvious that MAZE measures would positively correlate with measures of math. Various researchers have demonstrated a positive link between outcomes on MAZE reading measures and math measures of problem solving (Jiban & Deno, 2007; Thurber, Shinn, & Smolkowski, 2002; Whitley, 2010). The correlations for MAZE and state testing outcomes were larger than typical in each study and were stronger in the upper grades (fourth and fifth grade) than in the third grade. Additionally, Jiban & Deno (2007) noted that the MAZE task and a task of calculation accounted for much variance in state testing outcomes. These results are evidence that success on math assessments might be controlled to some degree by proficient reading comprehension, particularly in the upper elementary grades.

Whitley (2010) found that the correlation between a measure of oral reading fluency and state outcomes in math was nearly the same as the correlation found between a MAZE measure and state testing outcomes in math. Crawford, Tindal, and Stieber (2001) also found moderate correlations between oral reading fluency measures and math achievement. They demonstrated that students who had very low reading fluency were much more likely to not pass the state exam than those who were proficient readers. These two studies highlight the utility of a one-minute measure of reading for the prediction of math outcomes; however Jiban and Deno (2007), argue that this type of measure should be used as only *one* piece of information to help determine, interpret, and/or predict future outcomes in math [emphasis added]. They demonstrate that single

measures do not account for as much variance as do a combination of outcomes to predict future success.

Rutherford-Becker and Vanderwood (2009) reported that measures of arithmetic fluency and measures of reading comprehension predicted an applied math outcome better than a measure of oral reading fluency alone. In this study, as in several regarding comprehension variables discussed previously, the subjects were in upper elementary school. A clear consensus regarding oral reading fluency is that as student reading ability grows, which is the case in later elementary years, oral reading fluency becomes less of a valuable predictor for outcomes than measures of comprehension (Fuchs & Fuchs, 1993; Silberglitt, Burns, Madyun, & Lail, 2006). Based on this information, fluency measures may provide the most useful predictive information for teachers and researchers regarding proficiency if subjects are in grades three and below. This also tends to be the time in school when early intervention and identification of special supports for students are most often first implemented.

The described studies represent a foundation for the belief that content-free problem solving, often measured by non-verbal ability tests, and reading proficiency, often measured by oral reading fluency probes, may be influential construct irrelevant variables on math problem solving outcomes as measured by state math tests, particularly in early grades. Additional construct irrelevant variables such as student demographic characteristics are also important to recognize and consider when evaluating math outcomes and are further described in the following section.

Identification of Student Demographic Construct (Irrelevant) Variables in Math

Problem Solving

Other variables that may affect outcomes for students include gender, poverty, language facility, and disability. Each variable presents unique, yet related considerations. These are important factors to consider for teachers largely because they cannot be influenced by instruction. The influence of these factors is important for teachers and researchers to understand so assessments are designed to limit the influence of these factors and truly measure aspects of learning over which teachers have direct control.

Gender. Although there is evidence that the lack of women in Science, Technology, Engineering, and Mathematics (STEM) fields is a current reality, it does not appear to be due to a lack of assessment achievement in mathematics by females (Beede et al., 2011; Hyde, Lindberg, Linn, Ellis, & Williams, 2008; Scafidi & Bui, 2010). Using state assessment data from several states, in both 2008 and 2010, researchers demonstrated that girls and boys performed relatively equally on measures of mathematics achievement (Hyde et al., 2008; Scafidi & Bui, 2010). Further, Scafidi and Bui demonstrated that this performance was not moderated by participation in other special population categories (ethnicity, ELL, etc.). These studies were conducted in both middle and high schools.

Despite the lack of evidence for actual difference in state content assessment performance by males and females, gender seems to be a significant factor related to educational experience and likely instruction based on the fact that boys typically are overrepresented in special education categories. According to Wehmeyer (2001), this

could be due to several reasons including biology, behavior, or bias. Wehmeyer conducted a study involving students who initially qualified for special education. Naturally, most of these students were of elementary age at initial referral. Results indicated that IQ was significantly different between males and females (females slightly lower) and males most often had behavioral issues associated with their referral to special education. The study also indicated that the behavioral factors might have created bias for a higher rate of special education referral. From this information, it appears that gender may co-vary with performance indicators such as g as well as non-performance indicators like special education eligibility. It may be an influential, yet construct irrelevant factor to consider when interpreting academic outcomes for students.

Poverty. Another demographic variable of influence in educational success is poverty. Students who experience poverty as a group are at increased risk for failure on educational outcome measures. According to a meta-analysis conducted by Sirin (2005), socio-economic status (SES) was positively correlated to outcomes on academic measures. Specifically in the area of math, the correlation was very high when compared to the correlations between SES and outcomes in other academic areas such as reading or science.

Other researchers have demonstrated that the influence of poverty on math achievement is significant, especially in the early years (Burnett & Farkas, 2009). This may support the belief that once students are exposed to curriculum and good teaching, math deficits can be outgrown. Another study, though, noted that although students may progress once exposed to teaching, they might not ever catch up to their peers who have higher socio-economic status (Jordan, Kaplan, Olah, & Locuniak, 2006). The federal

government identifies students living in poverty (i.e. low SES) as a focus group that receives additional educational resources through Title I due to consistently low performance on academic assessments as compared to students of average wealth. These resources are intended to support additional teachers and materials to deliver instructional interventions to diminish the negative impact of poverty. Again, poverty is unrelated to the construct of mathematics, but appears to be a variable of influence on assessment.

Limited English proficiency. In a similar way, language proficiency correlates highly with state testing outcomes. Testing for students who are learning English has been an area of increased focus during the past 15 years. The Individuals with Disabilities Education Act (IDEA) of 1997 required that all students be included in state testing programs. Jamal Abedi has been noted in the literature for several years on the topic of English Language Learners (ELLs). He argued that by using testing results from current assessment systems, educators are in jeopardy of making decisions that have detrimental consequences for this population (Abedi, 2006). For example, reliability and validity information is greatly affected by the fact that typical state assessments are not normed for this particular population. Therefore, assessments may not fairly reflect abilities of students who are learning English. When this happens, educators might make decisions that are unfair for this group of students based on an inaccurate understanding of proficiency.

Linguistic difficulty of assessments is another feature that impacts ELLs more than native English speakers. This has been documented in several ways. Abedi (2006) describes features such as long phrasing, complex sentences, unfamiliar vocabulary, and conditional clauses, among others, that present unnecessary and unfair negative bias for

17

ELLs. In one study (Abedi, Lord, Kim-Boscardin, & Miyoshi, 2000), researchers presented an assessment in different formats, including one format with a dictionary and another with a translation of the text. They found that ELLs did much better when they had supports for language than without. Additional studies in which researchers modified the language highlighted the alterations as supportive for ELLs (Abedi & Lord, 2001; Abedi, Lord, & Hofstetter, 1998; Abedi, Lord, & Plummer, 1997). Overall, it appears that ELLs are impacted differently than proficient speakers of English in their ability to demonstrate proficiency in mathematics on assessments.

Special education. One reason that the ELL subgroup has gained attention is because students who are learning English are often over identified in special education programs (Sullivan, 2011). In general, both groups (ELLs and students in special education) demonstrate deficits in reading and writing skills when compared to English speaking peers or those without disabilities (Garcia & Tyler, 2010). It is possible that because reading is an access skill to mathematics assessments, both groups also demonstrate lower scores in math. In this way, a danger exists that an uninformed teacher may believe a student has a learning disability in math when, in reality, he or she may be having difficulty with language more than content. In addition to reading disabilities, math disabilities or cognitive impairments are conditions that are likely to negatively affect outcomes on math assessments. It is also important to recognize that not all handicapping conditions pose a threat to state assessment outcomes. These include conditions such as orthopedic impairments and articulation concerns. However, regardless of student characteristic or exceptionality code, students within these

subgroups experience differences in resource allocation, scheduling, and peer interactions than students of the majority in schools.

Like poverty, it may also be related to other performance indicators (like reading ability) or other non-performance indicators like gender. It's possible that like gender, special education eligibility may even co-vary with other variables. In fact, it is likely that all of the construct relevant or irrelevant variables interact in different ways. Within the literature, additional difficulties in working memory, processing speed, attention difficulties, and phonological skills are correlates for math disabilities (Fuchs et al., 2005) too. Other variables may influence student outcomes and be of interest to explore but because of the confines of the methodology for this study described in the next section, special education, like other student demographic variables and performance indicators were the only variables investigated intently.

As with the other student characteristics, special education eligibility appears to be a construct irrelevant variable to further investigate and consider when making decisions for students. It is important for teachers to be thoughtful when using state assessment data for any student within a special population group. With careful consideration, they can more accurately interpret student assessment results and make sound decisions regarding instruction or intervention needs for students.

The Quantification of Construct Relevant and Irrelevant Variables in Math Problem Solving

In order to monitor learning fairly, it is important that the monitoring tools are free from bias, equitable to all groups, and produce equal scores for groups that should be equal (AERA, APA, & NCME, 1999). Several researchers over the past 20 years have

devised different ways to reach this goal (Abedi & Lord, 2001; Abedi, Lord, & Plummer, 1997). Although attempts to accurately measure math problem solving by reducing CIV are supportive for students, the fact remains that elimination of CIV is unrealistic for large-scale tests, specifically state exams measuring complex constructs like math problem solving through the medium of language. For this reason, it is worthwhile for researchers to identify and quantify relevant and irrelevant variables that may influence outcomes. This way, educators can make more informed decisions regarding test scores, instructional implications, and resource allocation to support students toward the end goal of college and career readiness.

Based on the literature previously described and the methodological confines outlined in the next section, three measurement variables may be important to consider in math problem solving. These include (a) content knowledge, (b) general ability, and (c) reading ability/language facility. According to the authors of each assessment, these variables can be measured using three different measures: (a) easyCBM Mathematics Assessment (easyCBM-math), (b) the Naglieri Nonverbal Ability Test-Second Edition (NNAT2), and (c) the Oral Reading Fluency component of the Dynamic Indicators of Basic Early Literacy Skills (DORF). Outcomes on these measures can be compared to outcomes on the Oregon Assessment of Knowledge and Skills (OAKS-math), to determine the variance in math problem solving outcome scores that can be explained by construct relevant and irrelevant performance variables.

Content knowledge. As defined by Alonzo, Tindal, Ulmer, and Glasgow (2006), the easyCBM-math assessment tests math content using universal design for assessment (UDA) principles described by Ketterlin-Geller, Alonzo, and Tindal (2004). UDA is an

ideal that promotes accessible assessment for all by reducing the influence of any external factors (environment, disability, etc.) that may act as barriers to access and outcomes. In essence, UDA diminishes CIV in assessment. UDA options used on easyCBM-math include the increase of white space on a page, use of fewer answer options, and use of read aloud options.

For example, the easyCBM-math in third grade has an average of 3.9 words per question on the first 20 questions. Each question has only three possible answers. This is noticeably different than the practice questions on OAKS-math in third grade. The average number of words per question on the state example test was 15.8 words per question and each question had four possible answer options. Figures 1 and 2 (see Appendix A for Figs. 1-4) visually demonstrate the differences in language load between the two assessments. In technical reports, researchers demonstrate that the easyCBM-math assessment is technically reliable for students in both special and general populations (Nese et al., 2010).

Non-verbal, content-free general problem solving ability. According to Naglieri (2008), the NNAT2 is a measure of general ability that requires no language. Each question is presented in a pictorial format without any words on the page. Figure 3 displays what is presented to students during the testing session. As defined by Naglieri (2011), the NNAT2 is designed for the purpose of measuring general problem solving ability for students who have limited language to the same level of accuracy as their language-proficient counterparts. While easyCBM-math can measure content-embedded problem solving, the NNAT2 can help to quantify the concept of general problem solving that is not content specific (as defined by the authors). This understanding can help to

distinguish between the influence of problem solving as a skill and problem solving as ability. As with easyCBM-math, technical reports conducted by Naglieri note that this measure is reliable for use with special populations.

Reading fluency. According to Good, Gruba and Kaminski (2009), DORF is a measure of oral reading fluency. Each measure consists of a passage with approximately 240 words that a student reads in one minute. Although construct irrelevant, reading is an important access skill in math problem solving assessment and as discussed earlier, researchers have demonstrated a correlation between comprehension measures and outcomes on math assessments (Jiban & Deno, 2007; Thurber et al., 2002; Whitley, 2010). Actually, researchers believe that DORF is a very accurate predictor of future success on both reading assessments *and* measures of comprehension (Center on Teaching and Learning [CTL], 2012). DORF measures can help to determine how much influence reading proficiency (or lack thereof) can have on outcomes of math problem solving. Knowledge of this influence can help teachers make good decisions regarding instructional supports for students who are struggling in math, particularly those representative of special populations who may also be struggling with the access skill of reading.

Demographic variables. Teachers can make instructional decisions to support all students on their path toward college and career readiness when they more fully understand the influence variables have on math assessments. When these influential variables can be altered through teaching, teachers can immediately alter instruction and therefore change the course of success for students. Sometimes, however, these variables cannot be altered through teaching.

Although static, it is also important to quantify the influence of demographic variables on math assessment outcomes, particularly those that impact special populations. Ideally, score variance on subject area assessments should be explained by predictive variables that can be altered by instruction. This would highlight the belief that instruction in a subject area is important, influential, and can change outcomes such as those in mathematics. However, if additional variance in assessment scores is explained by demographic characteristics that are not altered by instruction, this may suggest that the outcome measure actually assesses something other than, or in addition to, the learning that occurs in the classroom. This would be important information for teachers and researchers to consider when making any decisions regarding proficiency for students within special populations. If this occurs, it would also be interesting to know to what degree the unique variance explained by each independent measure changes when demographic information is considered. If little or no change occurs, the influence of the various measurement variables (i.e. content, *g,* and fluency) would continue to be important to consider, regardless of any additional special student factors.

Possible outcomes. Each independent variable related to math problem solving assessment can help researchers better understand the various influences on math problem solving assessment and thus better understand the construct of math problem solving itself. Figure 5 (see Appendix B for Figs. 5-6) shows a graphic representation of possible relations between these measurement variables.

Any two variables may have a certain degree of correlation; however, in order to better recognize the construct or constructs represented in a math problem solving assessment like OAKS-math, one must consider more than correlations. Variance

partitioning is one way to recognize variance explained uniquely and commonly by independent variables within a regression. This type of analysis can also support better understanding as to the similarities that exist between various independent variables. Figure 6 shows a representation of this method of analysis.

For example, if students score high on the NNAT2, low on easyCBM-math, and high on OAKS-math, this may suggest that the quality shared by general problem solving ability as defined by NNAT2 and math problem solving as defined by OAKS-math may be much more notable than the quality shared by problem solving as defined by OAKS-math and content knowledge as defined by easyCBM-math. On the contrary, if students score poorly on the NNAT2, high on easyCBM-math, and high on OAKS-math, content may be more similar to what is measured on OAKS-math than what is measured by NNAT2. If students do poorly on DORF, perform highly on NNAT2, average on easyCBM-math and poorly on OAKS-math, this may indicate that reading ability is very influential to OAKS-math outcomes. Further, it may indicate that reading/language instructional supports may be more critical for students who have this need than instructional supports in math content. If students score high or low on all tests, it may mean that all of the tests are equally similar or that one or two in particular are highly similar to what is measured on OAKS-math assessments and this similarity overshadows the dissimilar third variable. Other combinations of outcomes may reveal other important information related to the construct or constructs represented in a math problem solving assessment and variance partitioning could help to determine specific influences on outcomes.

Research Questions

The purpose of this study is to use an extant data set to better understand the extent to which construct relevant and irrelevant variables explain variance on large-scale, math problem solving assessments. In order to do this, I will use a sequential multiple linear regression with commonality analysis to answer the following two research questions:

1. How much unique and common variance in math problem solving as defined by OAKS-math scores can be accounted for by measures of content-embedded problem solving as defined by easyCBM-math, content-free problem solving as defined by the Naglieri Nonverbal Ability Test (Second Edition), and reading fluency as defined by Dibels Oral Reading Fluency?

2. Is any additional variance explained once student demographic characteristics of gender, FRL status, ELL eligibility, and special education eligibility are controlled, and if so, how does the unique variance accounted for by each performance measure change?

CHAPTER II

METHODOLOGY

As a requirement of the Doctorate of Education program, researchers were required to use extant data sets. Students in the D. Ed. Program intend to be practitioners in the field. As such, extant data are used to answer questions of practice in education. For the purposes of this study, the construct relevant and irrelevant performance and demographic variables related to math problem solving assessment were confined to indicators found within the literature that were also available within the scope of daily data collection at a school district level. This study included extant data from one district in the Pacific Northwest.

As previously indicated, performance indicators could have included several other interesting variables such as attention/memory ability or achievement in multiple content areas that may have included reading, math, writing, executive functioning, etc. Additionally, student demographic characteristics could have included school movement, early school entry, instructional tier, instructional grouping, or attendance rates. All of these variables may be of interest for future similar studies; however, based on the literature reviewed and the extant data that were accessible, the analyses in this study included data from four assessment measures (easyCBM-math, DORF, NNAT2, and OAKS-math) and four demographic factors (gender, FRL status, ELL status, and IEP status) gathered during the spring of 2011 and the spring of 2012. According to the authors of each specific assessment, easyCBM-math is a measure of content-embedded problem solving, DORF is a measure of reading fluency, NNAT2 is a measure of general problem solving ability, and OAKS-math is a measure of math problem solving. Each

measure is explained in more depth in the materials section. Specific information regarding setting, participants, curriculum, measures, procedures, and analyses will be described in the following sections.

Setting

The participants in this study attended school in a mid-sized school district in the Pacific Northwest. The district is located in a community representing a large geographical area of over 16,000 square miles and a population of nearly 158,000 residents. It supports over 16,000 students in 27 different schools. In this community, there are 16 elementary schools, 5 middle schools, one K-8 school, and five high schools. The community is rapidly growing and the unemployment rate was 12.0% in December 2011.

Participants

The participants in this study included all students in the district who took the third grade OAKS-math test during the spring of 2012 and as second grade students, took easyCBM-math, the NNAT2, and DORF during the spring of 2011. This describes 913 students.

During the 2011-2012 school year, there were 629 males (49.0% of the population) and 654 females (51.0% of the population) in the second grade. Students who identified as Caucasian represented 85.9% of the population. The next highest majority group was represented by those who identified as Hispanic or Latino, (10.2%), while other ethnic groups made up the remaining 3.9%. The district Talented and Gifted (TAG) percentage was equal to approximately 7% of the student population at the time of this study.

In the reviewed literature, demographic characteristics of gender, socio-economic status (represented by Free or Reduced lunch qualification, "FRL"), English language learner status (ELL), and special education eligibility (IEP) represent subgroups that are impacted in different ways by academic assessments. Because of this, these groups were specifically investigated. For this study, students were considered part of the FRL and IEP subgroup if they had qualified for the subgroup at anytime during either spring of second grade or spring of third grade since this categorization would entitle students to differences in service allocation and support during their time of qualification. Students who were considered ELL students had at least one score of 4 or less on the English Language Proficiency Assessment (ELPA) during either spring of second grade or spring of third grade. Those scoring at level 5 were grouped as if they did not qualify for services because they did not receive specific supports or instruction for needs in English Language Development (ELD) during the time of the study.

Once all cases that did not have complete data (i.e. all four test scores) were excluded, the study sample contained 913 valid cases. Excluded students (i.e. missing a score for any measure) were investigated in order to determine any common characteristics. Possible reasons for exclusion included a lack of access to the assessments due to very low cognitive ability or absenteeism during a testing window. Specific information regarding students who were not included is displayed in the data collection and subject selection section.

Curriculum

In the spring of 2009, the district adopted the *Bridges In Mathematics (Bridges)* curriculum published by the Math Learning Center. The students in this study were

exposed to the *Bridges* curriculum during second grade. In kindergarten and first grade, teachers used the previous adoption of *Investigations in Number, Data, and Space* curriculum published by Pearson Education, Inc. According to a district official, during the 2010-2011 school year the district had not yet implemented district-wide interventions for mathematics and the district focus was on implementation of the general classroom curriculum with fidelity (L. Nordquist, personal communication, February 7, 2012).

During the time of this study, school district agreements existed regarding time spent in mathematics instruction. In the elementary schools, all students in all-day kindergarten through grade five participated in 60 minutes of math instruction with an additional 15 minutes for Number Corner (another component of the *Bridges* program) each day. These instructional agreements began along with the new math adoption in the fall of 2010. Prior to this, agreements did not exist regarding time spent in direct subject instruction.

Materials

The students in this study took (a) a content-embedded test of mathematics problem solving and skill with limited language (easyCBM-math), (b) a non-verbal measure of content-free general problem solving ability (NNAT2), (c) a measure of the access skill of oral reading fluency (ORF), and (d) the third grade OAKS-math. Students took the first three tests in the spring of second grade and the latter in the spring of third grade.

Each assessment was developed by different groups of researchers within the past six years. All of the assessments have established reliability and validity described in

detail in the following sections. Researchers used them in previous studies regarding achievement outcomes with the exception of the NNAT2. The first edition of the NNAT, a paper-pencil test, has only been used in one study relating to math achievement to date.

easyCBM-math second grade spring benchmark assessment. All students within the sample took a measure called the easyCBM-math second grade spring benchmark, which according to the authors is a measure of content-based problem solving with a light language load (Alonzo, Lai, & Tindal, 2009). Teachers use this measure, like other curriculum-based measures, to identify students who are at risk for educational failure in the area of mathematics. According to Nese et al. (2010), this assessment represents an adequate general outcome measure of math content knowledge and typically correlates well with large-scale math assessments.

In the district, building testing coordinators conduct the assessment three times each year via computer in grades one through eight. Although available in the fall, testing coordinators typically administer the kindergarten assessment beginning in the winter and then again in the spring. For the purposes of this study, participants were included if they were second grade students in the spring of 2011 and completed the second grade spring benchmark assessment.

Although the easyCBM measure has language, researchers designed the assessment so that the language impact is minimal (Alonzo et al., 2009). As previously described, it uses UDA principles that are shown to be supportive for all students. There are fewer words on this test than on a state math assessment (average 3.9 vs. 15.8 per question). Additionally, white space is increased and language is simplified (see Figure 1).

The test consists of 45 multiple-choice questions related to the focal point standards for each grade outlined by the National Council of Teachers of Mathematics (Alonzo et al., 2006). In Oregon, state standard developers used the focal points as a primary document when creating the Oregon state content standards in mathematics (ODE, 2012). Therefore, although the easyCBM math measure is not directly written to match the Oregon state standards, the content is generally the same in terms of focus and emphasis.

To investigate technical adequacy of the easyCBM math measures, researchers used split-half reliability and Cronbach's alpha to represent internal consistency reliability. Using a sample of 283 subject responses, Cronbach's alpha was .82. The split-half coefficient was .79. This demonstrates that the easyCBM math measure has adequate reliability as a measure for which it is described (Anderson et al., 2010; Anderson, Alonzo, & Tindal, 2010a; Anderson, Alonzo, & Tindal, 2010b).

Anderson et al. (2010a, 2010b) also reported on criterion and content validity evidence for easyCBM math. In order to demonstrate criterion validity, researchers determined the relation between the easyCBM math questions and math questions on the TerraNova assessment. In second grade, the three easyCBM math benchmark measures accounted for 66 percent of the variance in the TerraNova score. This was statistically significant. The relation between the spring benchmark score and the TerraNova math score demonstrated concurrent validity. The spring benchmark score accounted for 51 % of the variance in the TerraNova math score (again, statistically significant). To demonstrate content validity, the researchers conducted a Rasch analysis and a confirmatory factor analysis (CFA). All but six spring items had a mean square outfit

between 0.7 and 1.3, with most between 0.8 and 1.2. This is considered adequate content validity for high-stakes test items.

Based on the unique qualities of this assessment described by the authors, for this study, the second grade spring benchmark form was used as a measurement of content-embedded problem solving with limited language influence. Because it represents mathematical content, it is thought to be construct relevant: related to the construct of mathematical problem solving.

NNAT2 second grade spring assessment. The NNAT2 is a measure free of language. This assessment is described as a measure of general problem solving ability. As described on the publisher's website, the test "uses progressive matrices to allow for a culturally neutral evaluation of students' nonverbal reasoning and general problem-solving ability, regardless of the individual student's primary language, education, culture or socioeconomic background" (Pearson, 2012). The author notes that prerequisite skills are not required for the assessment (Naglieri, 2011).

The assessment system has seven different levels; however, each student within this study was given test level C. Levels A-G loosely correlate with the grade in which the student receives instruction. An example of a question on the NNAT2 is shown in Figure 3. The computer generates scores in several different formats. These include stanine score, percentile rank, ability index (standardized score), and a scaled score. For this study, the ability index was used. A sample student report is attached (see Figure 4).

According to the information regarding updated norms (Naglieri, 2011), researchers used split-half reliability and Kuder-Richardson Formula 20 (KR_{20}) to evaluate internal consistency. Using a sample of 99,004 subjects, the mean was 100.0

32

with a standard deviation of 16.0. The split-half coefficient was .90 and the KR_{20} was .88. The Standard Error of Measurement (SEM) was 5.22. This demonstrates that NNAT2 has adequate reliability as a measure for which it is described.

The manual also refers to studies of validity. Researchers correlated mean scores from the previous version of the assessment to those of the NNAT2. The correlation between tests was .998, indicating a very high level of performance consistency across measures. Further, researchers conducted a correlation between the NNAT2 and the Wechsler Nonverbal Scale of Ability (WNV) (Wechsler & Naglieri, 2006) using a subgroup of gifted students who were part of the updated NNAT2 norms study. The 2011 Naglieri Ability Index (NAI) scores were highly correlated to the WNV Full Scale scores and T-scores (.74). There was also a high correlation between the NNAT2 and Matrices indicating the measurement of similar constructs between tests (Balboni, Naglieri, & Cubelli, 2010).

It is important to note that Naglieri himself conducted the technical adequacy studies for this measure so the claims should be interpreted with caution. However, this measure is an unusual performance indicator rarely accessible to researchers on a large scale. Most often, measures of general intelligence or general ability are only administered to students who are part of an evaluation for specialized services. In this particular district, this assessment is used as one measure to identify students with intellectual giftedness and as such, it is given to every second grade student near the end of the year. Because of this situational convenience, the results of this measure can be further used to investigate the influence of content-free general problem solving ability on assessment outcomes.

For this study, the NNAT2 was used as a pure measure of problem solving ability unrelated to content. This is thought to be one of the two major concepts that underlie proficiency on state assessments, the other skill proficiency. Because this measure does not represent mathematical content, it could be considered construct irrelevant. However, because this study examines *mathematical problem solving* as it relates to assessment, this test highlights problem solving as a potentially influential construct relevant variable of interest. Study analysis will help to determine the relevant influence of pure problem solving on mathematical problem solving assessment outcomes.

DORF second grade spring benchmark assessment. A DORF measure was used to determine the potential influence of the construct irrelevant variable of reading ability on math outcomes. Although only a measure of fluency, researchers have demonstrated that this brief measure alone accounts for as much variance on reading performance outcomes as multiple reading measures combined (CTL, 2012). Further, other researchers have demonstrated a relation between reading performance and math performance on state assessments (Jiban & Deno, 2007) that may be due to the importance of this skill for access to large-scale assessments.

The DORF benchmark measure is administered three times each year in the district. The benchmark assessment consists of three one-minute passages (approximately 240 words long) that a student reads. The tester reads scripted directions before and during each administration. For this study, I used the highest score out of the three passages delivered in the spring.

Researchers used alternate form, test-retest, and inter-rater reliability to represent DORF as a reliable tool to measure reading fluency. The coefficients of .96, .91, and .99

represented very high reliability. The concurrent validity coefficient was .73. This was significant at the .001 level, and was a large effect size.

For this study, the spring benchmark measure in second grade was used as a representative measure of reading proficiency that is consistently shown to be an influential variable (although construct irrelevant) on outcomes in math assessments, both predictively and concurrently (Lamb, 2010).

OAKS-math third grade assessment. Teachers administer the OAKS-math during the spring of each school year beginning in third grade. This assessment is designed to measure proficiency in the area of mathematics related to third grade standards. On this assessment, students use skill efficiency and content comprehension to understand and solve problems in mathematical situations. For this study, this is the operational definition used for math problem solving and this test was the measure that represented the multi-faceted construct. Students were able to take the measure up to three times during the test window. For this study, I used the highest score received on the assessment during the testing window.

Researchers continue to reexamine reliability and validity information for this assessment. According to an official from the Oregon Department of Education, they use standard error of measurement for reliability evidence and they explain test development practices for validity evidence. According to this official, the assessment is technically adequate to measure mathematics proficiency in the area of problem solving (S. Slater, personal communication, September 4, 2012). In this study, this assessment was used as the dependent variable representing mathematical problem solving in assessment.

Procedures

The district had many assessment training and administration protocols in place to support the best possible testing opportunities for students. Because the district is quite large, each building used a testing coordinator to support the administration of assessments at the sites. The assessment administration and training procedures used for each test are described in the following sections.

Assessment administration and training procedures. Each assessment had specific instructions that were followed as part of a district training given to site testing coordinators. Many of the testing coordinators had delivered each assessment for more than one year and needed few supports; however, the district testing coordinator was available for any additional questions regarding administration. The district testing coordinator also opened and closed the benchmark window for each assessment.

For easyCBM-math, it was expected that testing coordinators in each school read the easyCBM Teacher Manual prior to administering the assessment. The manual contains answers to many common questions teachers have during testing. Students had unlimited time to complete the measure, although the typical student finished the assessment in approximately 25 minutes. Depending on the building resources, as suggested by the district coordinator, the assessment was delivered either in a lab setting (whole class) or in small groups in the classroom on laptop computers. For this particular test, the assessment window was May 16, 2011 to June 3, 2011.

Testing coordinators administered the NNAT2 assessment on the computer to all second grade students during the spring of 2011. All testing coordinators received training on administration protocols by the district assessment coordinator prior to giving

the assessment. Assessment coordinators have the option of using Spanish as an accommodation instead of English for their verbal directions if they believe that it would benefit the student. However, because pictorial options are also available it is rare that assessment coordinators use this accommodation. The average assessment takes approximately 30 minutes. This assessment, as suggested by the district testing coordinator, was delivered in either a lab setting or on laptops in small groups. The window for this test was May 2, 2011 to June 10, 2011.

Testing coordinators participated in DORF trainings prior to the testing window with staff members who administered the measure. A reading specialist, special education teacher, or building testing coordinator conducted these trainings. The focus of the training was to review protocols outlined by the assessment system, to practice using the measure, and to calibrate scores between those who would deliver the assessment. These assessments were delivered individually using paper copies of passages during the testing window from May 16, 2011 to June 3, 2011. Assessment proctors kept scores in assessment booklets for each individual student. Scores were then entered into the DIBELS database from which the district testing coordinator gathered results for each school.

The OAKS-math assessment had very strict training guidelines to ensure the secrecy of testing items and to support fair and equitable testing practices. Each school testing coordinator participated in a state-standardized training delivered by the district assessment coordinator. School coordinators then trained teachers in the building who delivered the assessment. Each person involved in the testing completed applicable training modules, read the test administration manual, and signed a test security form. By

signing the form, individuals verified that they had completed the module trainings and readings. The only individuals allowed in testing areas were district employees who had completed the training.

Trained testers administered the OAKS-math measure, either on laptops within classrooms or in a computer lab. Some students were given the assessment in a small group or individual setting based on accommodations determined by an accommodation team or listed on their Individual Education Plan (IEP) or Section 504 Plan. The assessment consisted of approximately 45 questions and typically took about 75 minutes to complete. This test was most frequently delivered during more than one testing opportunity and students had unlimited time to complete the assessment during the assessment window. The window for this test was November 8, 2011 to May 17, 2012. For this study, the highest score attained during the testing window was used.

Data collection and subject selection. In this study, extant data were used. The school district initiated the collection of data for each assessment during the window each was conducted and therefore, an exemption from the Independent Research Board was requested. Once approved, the appropriate data set from the district assessment coordinator was requested in early December 2012. The coordinator collected all of the data and converted student names to numbers to protect any sensitive student information beyond the scope of this study. The data set was received in January 2013.

In addition to scores of each measurement variable, the data file included demographic information regarding gender, free and reduced lunch eligibility, language proficiency level and special education eligibility. These data were drawn from district records on June 10, 2011 and June 10, 2012. These dates were chosen because all the

38

testing windows were complete. As previously described, students were considered participants in the free or reduced lunch category or special education category if they had participated in the program at any time in second or third grade. Additionally, students who had a score of 4 or lower on ELPA at any time in second or third grade were considered in the ELL subgroup. These decisions were made based on access to special services or resource allocations during the time of the study. All information remained confidential using guidelines outlined by the American Psychological Association (APA) to maintain records.

As previously mentioned, there were 913 students who took all four assessments. Missing cases were investigated to determine commonalities among non-participants. Information regarding valid cases and those missing for each demographic group and each independent variable is displayed in Tables 1-4.

Table 1

Valid and Missing Test Data by Gender

| | | Cases | | | |
| | | Valid | | Missing | |
	Gender	*n*	Percent	*n*	Percent
OAKS	Female	541	100.0	0	.0
	Male	510	100.0	0	.0
CBM	Female	484	89.5	57	10.5
	Male	446	87.5	64	12.5
NNAT	Female	516	95.4	25	4.6
	Male	490	96.1	20	3.9
DORF	Female	521	96.3	20	3.7
	Male	493	96.7	17	3.3

Table 2

Valid and Missing Test Data by Free or Reduced Lunch Status (FRL)

| | | Cases | | | |
| | | Valid | | Missing | |
	FRL	*n*	Percent	*n*	Percent
OAKS	No	469	100.0	0	.0
	Yes	582	100.0	0	.0
easyCBM	No	428	91.3	41	8.7
	Yes	502	86.3	80	13.7
NNAT2	No	452	96.4	17	3.6
	Yes	554	95.2	28	4.8
DORF	No	452	96.4	17	3.6
	Yes	562	96.6	20	3.4

Table 3

Valid and Missing Test Data by Special Education Eligibility

| | | Cases | | | |
| | | Valid | | Missing | |
	SPED	*n*	Percent	*n*	Percent
OAKS	No	901	100.0	0	.0
	Yes	150	100.0	0	.0
easyCBM	No	799	88.7	102	11.3
	Yes	131	87.3	19	12.7
NNAT2	No	865	96.0	36	4.0
	Yes	141	94.0	9	6.0
DORF	No	870	96.6	31	3.4
	Yes	144	96.0	6	4.0

Table 4

Valid and Missing Test Data by ELL Qualification

	ELL	Cases			
		Valid		Missing	
		n	Percent	*n*	Percent
OAKS	No	992	100.0	0	.0
	Yes	59	100.0	0	.0
easyCBM	No	877	88.4	115	11.6
	Yes	53	89.8	6	10.2
NNAT2	No	949	95.7	43	4.3
	Yes	57	96.6	2	3.4
DORF	No	956	96.4	36	3.6
	Yes	58	98.3	1	1.7

The numbers of missing cases for the easyCBM, NNAT2, and DORF were 121, 45, and 37, respectively. The percentage of missing cases in each demographic group (gender, FRL, IEP, ELL) was nearly the same as the percentage of those missing cases not part of the demographic group. In the original sample of students who took OAKS-math, 50% qualified for free or reduced lunch, 13% qualified for special education, six percent qualified for English Language Learner services, and 48% were male. In the actual sample (including those who took all four assessments), 53% qualified for free or reduced lunch, 14% qualified for special education, six percent qualified for ELL services and 48% were male. These percentages represent 488, 128, 52, and 438 students, respectively. This demonstrates that students who did not take part in an assessment were not markedly different than those who took the assessment in terms of demographic representation.

Analyses. The analyses addressed the unique and common variance in OAKS-

math scores that could be accounted for by three independent performance measures.

Measures included (a) content-embedded problem solving, (b) content-free problem

solving, and (c) reading fluency. Each measure represented a construct relevant or

irrelevant variable of interest in the assessment of mathematics problem solving.

First, descriptive statistics for each measure were outlined including correlation

coefficients for the measures related to one another. Next, step one of a sequential

multiple linear regression was conducted to determine the amount of variance explained

in the dependent variable by the independent variables. In order to investigate the

explained variance fully, a commonality analysis was then used to partition the variance

into that unique to each variable and that common to two or three variables together.

These variances were determined using the following formula:

$$U_{(1)} = R^2_{y.123} - R^2_{y.23}$$

$$U_{(2)} = R^2_{y.123} - R^2_{y.13}$$

$$U_{(3)} = R^2_{y.123} - R^2_{y.12}$$

$$C_{(12)} = R^2_{y.13} + R^2_{y.23} - R^2_{y.3} - R^2_{y.123}$$

$$C_{(23)} = R^2_{y.12} + R^2_{y.23} - R^2_{y.2} - R^2_{y.123}$$

$$C_{(13)} = R^2_{y.12} + R^2_{y.13} - R^2_{y.1} - R^2_{y.123}$$

$$C_{(123)} = R^2_{y.123} + R^2_{y.1} + R^2_{y.2} + R^2_{y.3} - R^2_{y.12} - R^2_{y.13} - R^2_{y.23}$$

where the numbers represent predictor variables (1=easyCBM-math, 2=NNAT2,

3=DORF) and U/C represent unique and common variance respectively (Nimon, Lewis,

Kane, & Haynes, 2008).

A second step within the sequential multiple linear regression was used to determine if any additional variance was explained once student demographic characteristics were controlled. The unique variance accounted for by the performance measures in the second step was compared to that explained in the first step.

These analyses provided information about which variables explained the most variance in performance on a measure of math problem solving (OAKS-math). Additionally, this information provided insight to the represented constructs in math problem solving assessment and the relative importance of each independent variable for success on state outcome measures. Finally, the analyses provided information about the extent to which inherent student demographic characteristics influence outcomes on state assessments in mathematics.

CHAPTER III

RESULTS

Descriptive statistics were calculated for each of the variables in order to determine normal distribution. Correlations were also calculated between all variables. Next, two multiple regression models were run to determine the variance explained by a model including testing variables, followed by a model to determine the additional variance explained by any demographic or non-performance indicator. After the first step in the regression, a commonality analysis was used to determine the amount of variance explained by each variable uniquely, as well as the common variance explained jointly by the variables.

Descriptive Statistics

Descriptive statistics for each variable, as well as intercorrelations, are displayed in Table 5. Each variable had normal distribution (skewness between -1.0 and 1.0). As the correlation values show, all variables were significantly correlated with OAKS-math scores as well as one another (.36 - .71). OAKS-math scores and easyCBM-math scores were most highly correlated (.71) and NNAT2 achievement index scores and DORF scores had the lowest correlation (.36).

Table 5

Means, Standard Deviations, and Intercorrelations for Variables in Math Problem Solving

Variable	M	SD	OAKS	easyCBM	NNAT2	DORF
OAKS	217.30	9.829	---	.71***	.60***	.51***
easyCBM	36.57	7.478		---	.58***	.49***
NNAT2	99.75	13.185			---	.36***
DORF	106.68	39.471				---

***$p < .001$.

44

Analysis One: Performance Measures

A sequential regression was conducted to determine the degree to which each independent construct relevant or irrelevant performance measure predicted OAKS-math scores in third grade (Table 6). After the first step, a commonality analysis was conducted to determine the unique and common variance accounted for by each measure and measures in combination (Table 7).

Table 6

Sequential Regression Analysis Predicting OAKS-math from easyCBM, NNAT2, and DORF

Step and Predictor	B	SE	β	t	R^2	Adj. R^2	r	sr
Step 1	170.30	1.16		105.35***	.58	.58		
easyCBM	.61	.04	.46	16.22***			.71	.35
NNAT2	.20	.02	.27	10.12***			.60	.22
DORF	.05	.01	.18	7.33***			.51	.16

Note. sr = semipartial correlation coefficient. $N = 913$.
***$p < .001$.

Table 7

Variance Partition of $R^2 = 58.1\%$ with easyCBM, NNAT2, and DORF (N=913)

U/C	easyCBM (T1)	NNAT2 (T2)	DORF (T3)	R^2 Partition
U_1	12.11%			12.11%
U_2		4.71%		4.71%
U_3			2.46%	2.46%
$C_{1,2}$	15.74%	15.74%		15.74%
$C_{1,3}$	7.13%		7.13%	7.13%
$C_{2,3}$		0.76%	0.76%	0.76%
$C_{1,2,3}$	15.14%	15.14%	15.14%	15.14%
Sum = r^2	50.12%	36.35%	25.49%	--
Sum = R^2				58.05%

Sequential regression results indicated that each variable (easyCBM-math, NNAT2, DORF) significantly predicted OAKS-math scores and that together they explained 58.1% of the variance in OAKS-math, $F(3, 909) = 419.70$, $p < .001$. Each factor had a positive effect on OAKS-math. For each point increase in easyCBM, an increase of .61 in OAKS-math was predicted, $t = 16.22$, $p < .001$, 95% CI [.53, .68]. For each point increase in NNAT2, an increase of .20 in OAKS-math was predicted, $t = 10.12$, $p < .001$, 95% CI [.16, .24]. For each point increase in DORF, and increase of .05 in OAKS-math was predicted, $t = 7.33$, $p < .001$, 95% CI [.03, .06].

Results from the commonality analysis (variance partitioning) revealed that easyCBM-math and NNAT2 uniquely explained 12.11% and 4.71% of the variance (respectively) in OAKS-math scores. CBM and DORF jointly explained 7.13% of the variance in OAKS-math. The largest variance partitioning percentages came from the variance explained by all three variables commonly (15.14%) and from the variance explained jointly by easyCBM-math and NNAT2 (15.74%). The lowest variance partitioning percentages came from DORF uniquely (2.46%) and the variance explained jointly by NNAT2 and DORF (.76%). Figures 6 and 7 represent a visual display of the partitioning of variance.

Analysis Two: Measures with Student Demographic Characteristics

To determine if any additional variance was explained once student demographic characteristics were controlled, a second step was added to the sequential regression including variables of gender, FRL, ELL, and IEP (Table 8).

Table 8

Sequential Regression Analysis Predicting OAKS-math from Performance and Non-performance Indicators

Step and Predictor	B	SE	β	t	R^2	Adj. R^2	r	sr
Step 1	170.30	1.62		105.35***	.58	.58		
easyCBM	.61	.04	.46	16.22***			.71	.35
NNAT2	.20	.02	.27	10.12***			.60	.22
DORF	.05	.01	.18	7.33***			.51	.16
Step 2	172.29	1.80		95.87***	.60	.59		
easyCBM	.55	.04	.42	14.49***			.71	.31
NNAT2	.20	.02	.27	10.23***			.60	.22
DORF	.04	.01	.18	6.76***			.51	.14
FRL	-1.37	.46	-.07	-3.00**			-.33	-.06
ELL	-.97	.96	-.02	-1.02			-.25	-.02
IEP	-.56	.63	-.02	-.88			-.14	-.02
Gender	2.02	.43	.10	4.72***			.14	.10

Note. sr = semipartial correlation coefficient. $N = 913$.
**$p < .01$. ***$p < .001$.

Results of the second step of the sequential regression indicated that, controlling for non-performance indicators, the model as a whole was a significant predictor of OAKS-math scores, $R^2 = .60$, $F(7, 905) = 189.97$, $p < .001$. A closer investigation revealed that although the control of demographic variables added statistically significant predictive power to the model, R^2 change $= .01$, $F(4, 905) = 7.99$, $p < .001$, only two of the added variables were significantly influential (FRL and gender). Qualification into FRL status had a negative impact on OAKS-math, and males had higher scores (β = -.06 and .10 respectively). Although these variables added predictive power, they explained very little unique variance (sr^2) in OAKS-math. In fact, the control of demographic variables only accounted for an additional 1.4% explained variance in OAKS-math scores.

To determine how the unique variance accounted for by each independent performance measure changed once demographic variables were controlled, the semipartial correlation coefficients were squared. These were then compared with the original unique variances attained from the first step in the regression. Results are displayed in Table 9.

Table 9

Comparison of Unique Variance Attributed to Performance Variables Before and After Control of Demographic Variables

| Predictor | Variance | | Δ Variance | Relative Δ Variance |
	Step 1	Step 2		
easyCBM	12.11%	9.36%	- 2.75%	-22.71%
NNAT2	4.71%	4.67%	- .04%	-0.85%
DORF	2.46%	2.04%	- .42%	-17.07%

Note. Relative Δ Variance = Δ Variance/Step 1 Variance

In all cases, the unique variance accounted for by each independent performance variable decreased when demographic variables were controlled. The variance accounted for by easyCBM-math, which started with the largest amount of explained variance attributed to it decreased the most, both actually and relatively. NNAT2 decreased the least (- .04%). Reduction in uniquely explained variance is to be expected when additional variables are added into a model: the more variables, the less opportunity for uniquely explained variance.

CHAPTER IV

DISCUSSION

This study highlighted specific performance and non-performance variables as influential factors for outcomes on high-stakes assessment measures of math problem solving as defined by OAKS-math. In the first analysis, mathematical content knowledge, content-free problem solving ability, and oral reading fluency were used as independent performance variables. In the second analysis, non-performance variables added to the model were gender, FRL, ELL, and IEP status. In the following sections a summary of outcomes is provided followed by a discussion of the limitations for this particular study. In the interpretations section, this study is compared and contrasted with previous research and important topics regarding the use of predictive measures in assessment and construct definition are discussed. The last section contains a discussion of practical considerations and areas for future research.

Summary

The purpose of this study was to provide additional research on the underrepresented topic of construct validity in large-scale assessments: specifically construct relevant and irrelevant variance as it relates to the assessment of math problem solving. To do this, a sequential multiple linear regression was conducted to determine the relative predictive nature of various performance variables (both construct relevant and irrelevant) to large-scale math assessment outcomes. This was followed by variance partitioning to further understand the unique variance in OAKS-math explained by each variable as well as the variance explained by characteristics held commonly between variables. Next, another regression was used to examine if by controlling for

49

demographic variables more variance in OAKS-math could be explained. Each analysis was conducted in order to better understand the construct of math problem solving as it relates to assessment. As complex constructs, such as math problem solving, are more clearly defined, decisions regarding use of the assessment results can become more valid. This is of particular importance when high-stakes educational decision-making happens based on assessment outcomes. Further, a more complete understanding of systematic error in assessment may allow for better assessment design, thus leading to more accurate assessment results and interpretations (Haladyna & Downing, 2004). Studies such as this also provide a foundation for future investigation of mathematical problem solving and how the construct can best be assessed.

The four assessments (easyCBM-math, NNAT2, DORF, and OAKS-math) were strongly positively correlated to one another. Each correlation was significant at the $p <$ .001 level; however, the correlation was strongest between easyCBM-math and OAKS-math (r = .71) and weakest between NNAT2 and DORF (r = .36). This makes sense in terms of construct representation. The easyCBM-math assessment and OAKS-math clearly represent math content knowledge in assessment, while DORF and NNAT2 represent what appear to be two relatively different constructs: reading fluency and non-verbal problem-solving (Good et. al., 2009; Pearson, 2012). It is inappropriate and beyond the scope of this study to comment on causation among these variables; however, the significant correlations between and among each performance variable indicate that students who do well on one of the assessments will likely do well on another, regardless of the represented construct. Often, high correlations such as these also pose a threat for multicollinearity, which I will discuss in the following section.
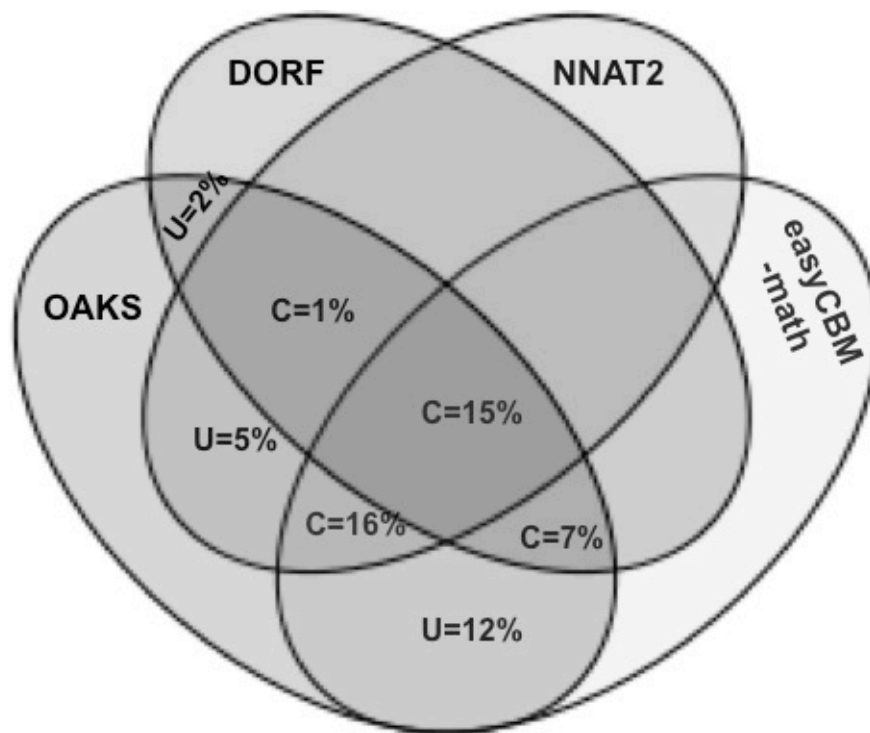
50

Figure 7. Commonality analysis results. This figure illustrates the unique and common variance explained in OAKS-math by three different performance measures. Variance was partitioned using a commonality analysis. U = unique variance, C= common variance.

Note. Figure not drawn to scale.

As displayed in Figure 7, the results of the first analysis indicated that a large amount of variance in OAKS-math (58.1%) was explained by the three independent measures taken together. Additionally, the unique variance contributed by easyCBM-math (12.1%) was more than that contributed by any other performance variable alone. This finding demonstrates that the uniqueness of easyCBM-math, possibly attributed to mathematical content knowledge (Alonzo et al., 2006; ODE, 2012), is more similar to the

construct measured on the OAKS-math than any of the other assessments' unique constructs. Another interesting finding was that the OAKS-math variance explained by combining measures of content knowledge (easyCBM) and problem-solving ability (NNAT2) was more than any other, unique or common (15.74%). This finding suggests that the quality shared by innate problem solving ability as measured by NNAT2 and content knowledge as measured by easyCBM-math is also a quality foundational to math problem solving as measured by OAKS-math. The other variables (NNAT2 and DORF) uniquely explained smaller amounts of variance (4.71% and 2.74%, respectively).

The results of the second analysis indicated that, in general, demographic characteristics did not add much to the variance explained by performance indicators alone. Technically, FRL and gender together accounted for another 1.4% of the variance in OAKS-math scores. While this result is, from a technical standpoint, statistically significant, it is not very interesting. More specifically, the unstandardized beta values for each of the performance indicators did not change from Step 1 of the model to Step 2. This indicates that the variance explained by each of the performance indicators was virtually unaffected by the addition of demographic characteristics to the model. Based on this stability, one could conclude with confidence that gender, FRL, ELL, and IEP status have very little impact on outcomes of math problem solving once math content and problem-solving abilities are controlled, which is what one would hope. After all, state assessments should not be measures of demographics.

Once demographics were controlled, the unique variance explained by each performance indicator was compared to the variance explained prior to demographic control. The variance explained by each independent measure went down slightly. From a

technical standpoint, this result would be expected because whenever more variables are entered into a model, the unique variance attributed to each factor is likely to decrease. The variance attributed to NNAT2 changed the least, while the variance attributed to easyCBM-math changed the most (-0.04% and - 2.75%, respectively). The results of this analysis are more thoroughly interpreted in a following section.

Limitations

As with any study, limitations to the internal consistency and generalizability exist. These include instructional considerations, mortality, extant data use, demographic representation, grade representation, and statistical conclusion validity. Threats to internal and external validity are outlined in the following sections.

Threats to internal validity. A threat to internal validity was instructional controls. For this study, there was no control over the instruction that students received during the year. This threat is important to consider because nearly a year of instruction took place between administration of the independent measures and the OAKS-math test. As explained in Chapter II, the district established instructional agreements for the amount of minimum time that students received instruction in the core mathematics curriculum. Any additional time spent in instruction, including instruction delivered in small groups or individually due to IEP needs, was not investigated. Not only was time in additional instruction not investigated, the quality of instruction due to difference in teachers was not considered. Both of these factors (time and instruction quality) may have affected the results in OAKS-math scores in different ways that without additional study will remain unknown.

Threats to external validity. In this study, 1116 students were part of the original data sample. Only 913 subjects had complete data, meaning they completed all four assessments. This means that the mortality for this study was 203 subjects. The demographic characteristics of the missing cases were outlined in previously displayed Tables 1-4. From these analyses, it appears that subjects not included in the study sample were not unlike those included, meaning that there was little evidence to suggest that students were excluded for specific demographic reasons.

As minimum criteria, students were only considered to be part of the original data sample if they had taken the OAKS-math assessment in the spring of third grade, as this was the dependent variable. For this reason, there are no missing cases listed under the OAKS-math category. Students who did not take part in OAKS-math would undoubtedly be markedly different than those taking the assessment because exclusion from this test most frequently is due to the inability to access the assessment due to extreme educational needs. These students most often qualify for special education and have alternate assessment plans.

For each assessment, normal distribution was investigated. Figures 8-11 in Appendix C show normal distributions for each of the variables. Each was relatively normal without skewness or kurtosis issues of concern. However, the loss of scores in each assessment does impact statistical conclusion validity. If fewer students had incomplete assessment scores, the statistics found in the analyses would be more complete.

Another limitation is due to the use of an extant data set. Because of the confines of these previously gathered data, I was only able to investigate the influence of the

performance and non-performance indicators described in the study. Although this study is relatively small in scope, it does provide a basis for replication using other influential variables.

The lack of subject diversity is another threat. The community from which these results were drawn was relatively homogeneous. Based on district information, few students represent ethnic or racial categories different than the Caucasian majority; however, in this study, gender, FRL, ELL, and IEP were the only demographic categories investigated and therefore are the only categories that can be discussed. Of the 913 subjects, 488 students qualified for FRL, 52 qualified for ELL services, and 128 qualified for special education services. In the case of ELL and IEP qualification, these numbers represent only a fraction of the entire population (5.7% and 14%, respectively). Because the numbers were so small, the ELL levels and special education handicapping conditions were not broken into separate categories. With a larger, more diverse sample, the impact of these levels and conditions could have been more thoroughly represented.

Another threat to the generalizability of this study is the single grade level focus. For this study, the OAKS-math assessment in third grade was used as the dependent variable. As with all state academic assessments, OAKS-math has questions regarding the knowledge and skills that students should have mastered by the end of third grade (ODE, 2012). The state standards for third grade differ from those in other grade levels. They also differ from standards in other states (Webb, 1999). Future similar studies using common standards should reduce state-to-state differences; however, content mastery will remain different at each grade level. It is also possible that as grade levels increase, the correlations among the independent variables chosen for this study and state outcome

assessments at other grade levels may differ. Long-term studies investigating the link between these variables and large-scale assessment outcomes in each grade will help to more completely examine the stability of influence in all grades.

Finally, because all measures were highly correlated, multicollinearity could be considered an issue of concern. In typical regression models, this creates a problem because it becomes difficult to determine what variables account for the variance in outcomes. In this particular study, Tolerance values were greater than .42 for each predictor, so multicollinearity was not an issue; however, this is a problem that is frequently recognized in social science research. One way to minimize this threat is to analyze data using some form of variance partitioning (Zientek & Thompson, 2006). In this study, a commonality analysis was used to support the analysis of predictive and influential variables on OAKS-math outcomes. In the next section, I discuss the use of variance partitioning to support interpretations as well as other interpretations based on the results of this study.

Interpretations

Educational accountability continues to be a topic of much interest and debate in this country. Each year, district leaders all across the United States work hard to ensure teacher quality and student access to current curriculum and instructional tools in order to support educational learning gains. These learning gains are demonstrated using assessment at the classroom, school, district, state, and national levels. Because high-stakes decisions are made based on assessment results, it is important for teachers and researchers to understand deeply what each assessment truly measures (Haladyna & Downing, 2004). Often, educational assessments purport to measure very broad topics or

developing abilities (Messick, 1984). One example of a *developing ability* is mathematics problem solving.

In the following sections, I discuss large-scale issues one might consider in light of the results of this study and how they relate to previous research. First I discuss variables or underlying constructs of particular importance and non-importance in the measurement of math problem solving. Next, I describe the practical use of this information from a formative perspective. Finally, I describe commonality analysis as a useful way to more thoroughly understand variance in high-stakes assessments of complex constructs.

Influential and non-influential variables in math problem solving. As far back as 2000, the NCTM outlined domains of mathematical proficiency, each containing specific content knowledge to be mastered in order for one to be a successful mathematician. Since then, researchers have continued to demonstrate the importance of mathematical content knowledge for success on various state assessments in mathematics (Anderson et al., 2010a; Anderson et al., 2010b). In previous studies, easyCBM-math reliably predicted success on OAKS-math and Measures of Student Progress (MSP) over the course of a single school year focused on a specific grade-level set of standards (Anderson et al., 2010a). The results of this study lend additional support to the predictive nature of easyCBM-math; however, the results indicate a predictive quality spanning more than one grade level. This means that not only is mathematical content important for instruction and assessment during the current year, it also has enduring significance. These findings suggest that content knowledge gained at any point in the

educational career will likely support more successful outcomes on math problem solving assessments in the future.

Based on outcomes from this study and others, problem solving, (*g*)*,* is indeed influential on outcomes relating to mathematical problem solving (Fuchs et al., 2005; Fuchs et al., 2006; Hart et al., 2009; Mannamaa et al., 2012; Naglieri & Ronning, 2000). In each of the reviewed studies, correlations and beta weights were used to demonstrate the relation between general ability and math problem solving. This study adds to the understanding of this link by using variance partitioning. It is noticeable that although *g* can help to explain much variance in OAKS, the variance that it uniquely explains is quite small. Rather, it is what it shares in common with easyCBM-math that contributes to the most explanation of variance in OAKS-math scores (see Figure 6). This may be the difference between the unique construct of IQ and the commonly held construct of problem solving or problem attack. For example, a student may have a high IQ but choose not to spend any time on using their understanding or knowledge to actually solve a problem. The application of this problem solving ability or base of understanding is more commonly reflected in OAKS-math and easyCBM-math than the level of student ability alone (Alonzo et al., 2006; ODE, 2012; Pearson, 2012;). Without application, ability seems to be of little importance in explaining variance in OAKS-math scores.

Similarly, the variance explained by the combination of NNAT2 and easyCBM-math was quite large (~16%). This was expected based on various correlational studies previously described (Fuchs et al., 2005; Fuchs et al., 2006; Naglieri & Ronning, 2000). This result may be a reflection of a quality that is common to all three assessments such as logic. Both content-embedded and content-free problem solving rely heavily on a

logical processes by which to problem solve as well as a common-sense understanding of the reasonableness of a potential answer. This is speculation and more research is necessary to determine the differences in these constructs definitively.

Several of the variables in this study. All of the construct irrelevant variables in this study, including reading fluency and the non-performance variables of gender, FRL, ELL, and special education eligibility were determined to be only marginally influential on math problem solving outcomes, if at all. For example, using variance partitioning, the influence of DORF was partitioned. As a result, the quality that is unique to reading fluency was compared to the quality that it shares with easyCBM-math. The variance reading fluency uniquely explains in OAKS-math performance was not nearly as large as the variance it jointly explained with easyCBM-math (3.5% and 7.1%, respectively). This was surprising given the research from Crawford et al. (2001) and more recently from CTL (2012) that indicates that DORF may be a predictive measure for success on math outcomes. Other research regarding the link between MAZE tasks and math outcomes leads to the same conclusion (Jiban & Deno, 2007; Thurber et al., 2002; Whitley, 2010). This study lends additional support to the predictive nature of DORF for math problem solving outcomes; however, variance partitioning provides additional important information.

Perhaps this marginal unique influence is a reflection of the importance of comprehension over fluency at third grade. As described by other researchers reading comprehension has shown to correlate highly with math outcomes (Jiban & Deno, 2007; Thurber et al., 2002; Whitley, 2010). Although DORF has been shown to be a highly predictive assessment of comprehension (CTL, 2012) it has also been discussed as a

variable that has less of a predictive quality as students move beyond the early years in school (Jiban & Deno, 2007). This is a reflection of the move from students' ability to decode fluently to their skill in comprehending what they have read which is a change that happens approximately during the second or third grade. Because students become fluent readers at different times, it is likely that DORF may be more or less predictive accordingly. According to Jiban and Deno (2007), the correlations between MAZE and state testing outcomes were stronger in the older grades than in the younger grades. The current study utilized the single measure of DORF as a proxy for both reading ability and reading comprehension and although it explained much variance, the unique variance explained was quite small. Perhaps comprehension measures would explain more unique variance in OAKS-math scores at this grade level.

The quality shared between all of the assessments, particularly the common explained variance by various measures and DORF, may be processing speed. Fuchs et al. (2006) describe cognitive correlates to arithmetic as processing speed and decoding. Both DORF and NNAT2 rely on speed of processing as well. Arithmetic is an obvious construct relevant skill to math problem solving, although not a skill investigated in this study; however, the results of this study may be additional evidence of the importance of the construct *shared* between processing speed and numeracy more so than the unique quality of decoding or orally producing words. Again, further research is necessary to determine the underlying constructs definitively.

The results of this study also lend further credence to Jiban and Deno's (2007) assertion that DORF should only be used as one piece of information when predicting outcomes in mathematics. They note that no matter how predictive, most often single

measures do not account for as much variance as do a combination of variables. Results from this study support their claim. The three performance variables, when taken together, accounted for six times the amount of variance in OAKS-math scores as DORF did alone.

When demographic variables in this study were controlled, the explained variance in OAKS-math scores increased only marginally. This means that despite research outlining the influence of each of these construct irrelevant factors on math outcomes (Abedi et al., 1998; Beede et al., 2011; Burnett & Farkas, 2009; Fuchs et al., 2005; Sirin, 2005), the information gathered using performance variables is more predictive of success than any of the non-performance variables in the current study. However, in this study, by controlling for these variables, the variance in OAKS-math scores accounted for uniquely by any of the performance indicators was slightly lowered in all cases. Interestingly, the relative changes in unique variance for easyCBM-math and DORF were far greater than that of NNAT2 (-22.7% & -17.1% vs. -0.9%, respectively).

These findings appear to indicate that demographic indicators affect outcomes related to easyCBM-math and DORF more than outcomes on the NNAT2. This would be expected because the NNAT2 test is a measure of general problem solving, which is thought to be a relatively stable ability throughout life (Davis, Arden, & Plomin, 2008; Gustafsson & Undheim, 1992; Larsen, Hartmann, & Nyborg, 2008; Reeve & Lam, 2005) and, as a measure free of language, it is less likely to affect special populations differently (Pearson, 2012). Additionally, this finding is interesting because it is an indication that NNAT2 is a measurement of something truly unique and unrelated to demographic factors. In essence, this assessment is highlighting a skill or competency

61

that is not overlapping at all with demographic impact. This may lend more support for using a measure such as this in order to gain specific knowledge about the impact of general problem solving ability on math problem solving outcomes; however, it should be considered with cost in mind. This consideration is described in the next section.

The fact that demographic variables did not account for virtually any additional variance in OAKS-math scores was very surprising given the literature base for performance differences demonstrated by these special populations (Abedi, 2006; Burnett & Farkas, 2009; Jordan et al., 2006; Sirin, 2005). One potential reason for this may be attention to test design by researchers and test creators. It is possible that because of the growing body of research related to discrepant performance by these special populations and the identification of the barriers to assessment success, tests like OAKS-math have been designed to limit CIV related to demographic characteristics. This is quite likely. Haladyna and Downing (2004) note an abundant research base in both differential item functioning and test item formatting. This research base was in existence during the creation of the current OAKS-math assessment (ODE, 2012). Additionally, the authors of two of the performance measures used in this study explicitly speak to this consideration in the literature. Both easyCBM-math and NNAT2, according to the authors, were created to limit the influence of access barriers for special populations (Alonzo et al., 2006; Pearson, 2012). This means that in this study, demographic factors would not influence the outcomes on performance variables and therefore, the performance indicators alone would account for any true variance in achievement on OAKS-math.

It is possible that the additional variance explained by demographic characteristics is a function of sample size or grade level more than a lack of additional variance. In this

study, FRL and gender were the only variables that explained any additional variance (albeit small). Previously described research suggests similar academic assessment performance by girls and boys at the middle and high school levels (Hyde et al., 2008; Scafidi & Bui, 2010). Perhaps the difference found in this study was due to a grade level focus in early grades rather than later years. Gender and FRL also represented the largest sample size. The number of students in special education and ELL were far less than those who were male or who qualified in the FRL category. In a larger sample of students, ELL or special education eligibility (not to mention any other demographic factors not explored in this study) may account for more additional variance in OAKS-math scores than performance measures alone although further research is needed to make this determination.

By using three performance variables, the model accounted for approximately 58% of the variance in OAKS-math scores. Use of variance partitioning provided additional important information regarding unique and common variance that may indicate constructs of different underlying importance. From construct validity and construct definition perspectives, it is also important to recognize that 42% unaccounted-for variance still exists. If OAKS-math is purported to be a measure of math proficiency and problem solving, what other factors or constructs make up this score if not related to content knowledge as defined by easyCBM-math, general problem solving ability as defined by NNAT2, or the access skill of reading as defined by DORF? Again, a definitive answer to this question is beyond the scope of this study; however, one can speculate as to reasons for the variance left unexplained.

Although easyCBM-math explained much of the variance in OAKS scores, it is ultimately a screener (Alonzo et al., 2006). As such, it was not designed to have the depth or breadth to completely reflect all of the skills or standards that students are exposed to during a school year (Deno, 1985). Perhaps a complete battery of math assessments may more completely reflect all of the learning one could gain throughout the year, but from a cost perspective this is unreasonable. Additionally, easyCBM-math in this study represents standards at the second grade level (Anderson et al., 2010c), while OAKS-math represents third grade standards (ODE, 2012). For this reason, it makes sense that a large amount of variance would be left unexplained. For example, if the NNAT2, DORF, easyCBM-math and OAKS-math were all given in the spring of third grade, it is likely that the independent measures taken together may have explained even more variance due to the fact that easyCBM-math and OAKS-math would be measures of the same standards. In fact, one would hope that the more instruction a student had, the less single-point-in-time measures would explain outcomes. That easyCBM-math in second grade explained more than 50% of the variance in OAKS-math scores a year later is, in fact, somewhat depressing from an intervention standpoint; however, as stated previously, this may mean that content in mathematics is largely built by broadening understanding each year rather than learning completely new skills in isolation.

Another possible explanation for the unaccounted for variance is teacher use of data. For example, if a particular school has a systematic method to collect and review data, a teacher may recognize struggling students quickly. If a teacher identified a struggling or low-performing student based on end-of-year data in second grade and began to systematically address areas of understanding deficits, it is likely that that

student would not struggle to the same degree they would have, had the teacher not intervened. In this scenario, one would hope that the independent measures used in this study would explain very little variance in OAKS-math scores by the end of the third grade year. This would indicate that the intervention designed by the teacher due to his or her use of data was extremely effective and substantially changed the academic trajectory for the student.

Measurement tools may also impact the potential for explained variance. Certain performance skills such as arithmetic or reading comprehension may be important variables to consider in future studies as well as attention, memory, or executive functioning. Student demographic characteristics that may be of influence might be parent education level, days of attendance, school movement, or instructional grouping. We know, based on this study, that any of these factors may overlap others in terms of unique qualities and what qualities they would share with OAKS-math outcomes. Perhaps there is a combination of skills that accounts for more variance in OAKS-math scores than the model used in this study. If so, it would be important to recognize which measured skills could be influenced through instruction and help support teachers so their instruction can be designed accordingly.

Though not of concern in this study, it is important to recognize measurement characteristics as potential barriers to variance explanation for future studies. For example, sometimes an independent measure may have a ceiling effect or a small amount of score variance. When a ceiling effect occurs, this means that students are unable to show the range of potential that they could demonstrate on the dependent measure. If the independent measure didn't have an adequate score variance, it is unlikely that the

independent measure score could explain much of the variance in scores attained on the dependent measure. Both situations impact the potential for explained variance. Based on the high stakes associated with state achievement assessments, there is obvious reason to continue to explore this complex construct and the predictive variables with which it might be associated. Additionally, as discussed in the next section, there is instructional utility to understanding formative variables that influence summative assessment outcomes.

Utility of predictive measures in assessment. From a public standpoint, outcome assessments such as OAKS-math hold much importance. They are used at the district, state, and federal level in order to reflect progress toward important outcomes like college and career readiness (Conley, 2010). Although they bear much weight on a large scale, summative assessments such as these hold little utility for teachers. Practically, the information gathered from these types of assessments is rarely used at the classroom level except to demonstrate to families in a broad sense if students became proficient on standards of importance for the grade level throughout the course of the school year.

Predictive measures, by contrast, can be widely influential at the classroom level, and results on these assessments will likely influence instructional practices immediately. Seminal work by Deno (1985) and countless studies since demonstrate that Curriculum Based Measures (CBMs) are reliable, fast, and cost effective assessment tools that can help teachers make everyday instructional decisions to support student outcomes. Many studies over the last 25 years have been conducted to investigate the predictive qualities of these measures. If measures at the classroom level are reliable, fast, cost effective, *and*

predictive, teachers are able to use them formatively to support students throughout the year toward success on outcome measures.

The results from this study indicate several construct irrelevant variables that are either only slightly or not at all influential for success in math problem solving as defined by OAKS-math. These include demographic variables and the access skill of reading fluency. Demographic variables are not factors that a teacher can alter. So it is helpful to know that inherent variables like gender, FRL, ELL, and special education status also do not influence math problem solving outcomes greatly. On the other hand, oral reading fluency, although only minimally influential for math problem solving outcomes, is a factor that can be changed through instruction. In addition to instruction, predictive tools such as formative progress monitoring measures, like CBMs, can help teachers gauge progress toward the goal of increased fluency (Good et al., 2009). Based on this study, it is likely that as fluency increases, math problem solving success will also increase, although not necessarily in a causal way.

In this study, the two major construct relevant variables of math content knowledge and general problem solving ability, (*g*), were found to be influential independently and in combination for math problem solving outcomes. Content knowledge, like oral reading fluency, is not a static skill or ability and can be altered by instruction. Similarly too, easyCBM-math is a predictive and formative measure teachers can use to monitor progress toward content knowledge development as the year progresses (Anderson et al., 2010a). It is likely that as knowledge of content increases, scores on easyCBM-math measures will increase and at the end of the year, scores on OAKS-math would be higher than they would have been had teachers not had this

67

predictive formative tool to use. The combination of instruction and formative tools to monitor success has significant potential to support student success on math problem solving outcomes.

Based on the results in this study, general problem solving ability, *g,* as measured by the NNAT2, shares something in common with OAKS-math uniquely (approximately 5% explained variance). It also shares a quality that is common to OAKS-math and the other measures (approximately 30% explained variance). These data suggest that general problem solving ability may indeed influence outcomes in math problem solving, yet according to the literature, it is thought to be relatively stable throughout life (Davis, Arden, & Plomin, 2008; Gustafsson & Undheim, 1992; Larsen et al., 2008; Reeve & Lam, 2005). None of the studies investigated the change in *g* for students in third grade specifically. Additionally, the time span difference between measures in each study ranged from months to several years and included groups of all ages. This evidence suggests that even though *g* is a highly influential variable, teachers may have little success working on changing outcomes for this particular construct.

Although the results of studies reviewed did not indicate that *g* was a factor that could be altered, one study (Davis, Arden & Plomin, 2008) investigated changes among general intelligence among groups of twins. In this study, genetic influence over *g* was most pronounced; however, there was evidence to suggest that environmental influence accounted for 30% variance in *g*. Although limited in scope, this evidence may indicate that certain environmental factors can impact *g* and thus, *g* may be alterable. Obviously, much more research in this area is needed in order to definitively determine if *g* can be altered by instruction and how this change could influence academic outcomes.

Variance partitioning may lend additional information to better understand influential variables on *g*. For example, it is possible that the unique characteristic attributed to *g* (perhaps intelligence) is not alterable, while the characteristic common to easyCBM-math and NNAT2 (perhaps problem attack or strategy) is alterable with instruction. This finding would be important for teachers as they alter instruction to support lagging skills students may have in particular areas. Additional research around the topic of general intelligence stability would help teachers and researchers make sound decisions regarding the use general ability assessment results to help design instructional programs for students.

Defining a complex construct. High-stakes assessments often measure complex constructs like math problem solving. According to Haladyna and Downing (2004) "Each [*developing*] ability involves contextualized mental models, schemas, or frames, and complex performance that may have multiple correct pathways that depend on knowledge and skills" (p. 17). Because of the complexity of these types of constructs, it is very unlikely to have non-overlapping variance between independent variables or underlying constructs of importance. Often, studies use correlations or beta weights to indicate predictive characteristics of specific variables toward outcome measures; however, this may lead to incomplete interpretations. In order to more deeply define a complex construct, variance partitioning may be a better option for analysis (Zientek & Thompson, 2006).

Using variance partitioning, one can recognize unique constructs of influence as well as characteristics held commonly between variables. For constructs with much overlapping variance, this exploration can highlight important qualities of each

independent variable, especially when each variable may overlap significantly with others. By recognizing unique and common variance, teachers can better target instruction based on constructs over which they have control rather than trying to change influential characteristics that are inherent or static. Prediction and correlations only minimally describe the relation among variables but the use of variance partitioning may help to support good decision-making based on a deeper construct understanding (Zientek & Thompson, 2006).

Implications and Future Research

This study highlights several topics of interest for researchers as well as classroom teachers. In the following paragraphs, I will discuss practical implications including cost, early intervention opportunities, and grade specific considerations. The section ends with possible topics for future research and exploration.

Practical considerations. Teachers should always consider the costs associated with any initiative in the classroom, including the addition of assessment. Costs may be monetary expenses, but more often, costs relate to instructional time. Based on this study, a commonality between easyCBM-math and NNAT2 explained the most variance in OAKS-math outcomes. NNAT2, in this district, is delivered once in second grade for all students and easyCBM-math is also mandated in second grade. A teacher in this district might consider using this information since it is already available to them. However, it would not be wise for a teacher from a district not implementing either measure to insist on administering and using both assessment tools. This would waste valuable resources including material gathering, time in training to administer the assessment, actual student time spent in assessment, and time to use the results.

Instead, someone interested in using a predictive formative measure for mathematics problem solving might consider the use of easyCBM-math in his or her classroom, school, or district. Based on results from this study, easyCBM-math uniquely accounted for 12% of the variance in OAKS-math scores but explained 50% of the variance in scores as a whole. This measure alone would give nearly as much information to a teacher as it would in combination with any other variable, while cutting the needed resources in half. Additionally, it is important for teachers to use assessment as just one of several informative tools to determine student needs in the classroom. Although easyCBM may be the best measure in terms of cost, it does not mean that it should be used alone or as a summative measure. Assessments should be considered one of many tools available to teachers (Jiban & Deno, 2007), and should also be utilized the way in which they are intended in order to guarantee validity (Messick, 1984).

The results of this study provide compelling evidence for teachers to have information about their students as early as possible. Typically in schools, each year begins with substantial time spent on creating community within the classroom, followed by assessment, and then teachers begin to create specific instructional groupings. This wait time appears to be unnecessary. This study demonstrated that scores on second grade indicators explained a significant amount of variance in math outcomes even at the end of third grade. In essence, teachers know who is struggling based on information from the previous year. As discussed, although teachers cannot change demographic variables, math content knowledge, reading fluency, and possibly even general ability can be changed through instruction. With a bit of organization, teachers can have access to student information very early, and begin specific instructional interventions quickly to

help change the academic trajectory for struggling students. This type of information may also be helpful in terms of student class placement to ensure that each student receives the best possible instruction for his or her specific needs.

Because math standards are different at each grade level, we do not yet fully understand the influence of variables like content knowledge, general ability, or oral reading fluency on math problem solving outcomes at each level. It is important then that teachers do not apply results of this study freely to any grade level or group of students that they support. For example, it would be inappropriate to claim that oral reading fluency predicts math outcomes in eighth grade. Without further research, this claim is unwarranted. It would be appropriate; however, to be thoughtful about student reading ability when giving an assessment of math problem solving to students. One might consider ways to accommodate students so the skill of reading is less influential to the outcome of the math assessment. As a third grade teacher, it also would be appropriate to rely more heavily on a combination of scores from math content and general ability when creating classroom intervention groups for math rather than DORF scores or NNAT2 scores alone.

Teachers should always be cautious when applying results of specific studies to the classroom due to differences in grade levels as well as population, subject, setting, etc. With continued research, the generalizability of specific claims may increase greatly. The following section outlines four areas of further research in this topic including the use of performance assessments, research using additional independent variables, studies at differing grade levels, and exploration of the assessment of other complex constructs.

Future studies. Recently, the movement toward the CCSS has also begun to change the traditional system of assessment. Future assessments will likely incorporate more performance-based tasks, as well as explanatory components, allowing students to demonstrate their thinking in ways not traditionally utilized (ETS, 2010). Haladyna and Downing (2004) refer to performance-based assessment as the best form of measurement for constructs of *developing abilities* like mathematics problem solving.

This type of measurement may provide a more authentic demonstration math problem solving skill, and studies such as this one provide a foundation for replication studies using the new assessment systems. As explored in this study, researchers can continue to analyze the unique and common variance attributed to several variables thought to be foundational to the construct of math problem solving as measured by the new performance assessments. However, although there may be benefits to performance assessment, other CIV threats exist. For example, it is unlikely that these types of tasks will be able to be assessed solely through technological means. Training for raters and inter-rater reliability will be critically important, as human error becomes a consideration in scoring.

Future studies using other independent variables such as computation skills, comprehension, vocabulary, and race and ethnicity will help us more fully understand influential factors on the construct of math problem solving in assessment. Research in this area may also be important for development of outcome and formative measures in the future, such as performance or multiple-choice assessments. By identifying unique and common variance attributed by additional factors, researchers might more completely understand what skills or competencies are being assessed on math problem solving

73

assessments. This knowledge may also support understanding about the degree to which we can alter foundational or influential constructs in order to better promote student success.

Studies such as this one help to explain predictive characteristics of various variables but only relating to specific grade levels as described in the previous section. Because of this lack of understanding at each grade level, it is important that replication studies at several grade levels be conducted. Although this study attempts to better define the construct of math problem solving and shine light onto construct relevant and irrelevant variables of influence, it only touches the surface. There continues to exist a need for more complete understanding of various skills that influence the outcomes on state assessments in all subject areas and without future research, this void will remain.

Complex constructs are very difficult to define and difficult to adequately assess (Haladyna & Downing, 2004). One interesting factor common among complex constructs is how they are traditionally measured. Reading comprehension, math problem solving, and other academic content areas are typically measured with tests delivered through the medium of language. This presents specific systematic variance that is sometimes completely unrelated to the construct of interest. Studies such as this one help to define what is actually measured on these assessments and how much influence language or other variables have on outcomes. Variance partitioning also offers a deeper understanding as to the underlying constructs of importance for each complex construct. Future studies involving other complex constructs will help to define predictive and alterable factors of importance for successful outcomes. These studies will also help to recognize the impact of certain construct irrelevant variables and variance (such as

reading facility) on outcomes. To measure student achievement fairly and comment on educational quality responsibly, these factors should be identified and minimized, if not eliminated, from assessment.

ASSESSMENT EXAMPLES

You have 1 quarter, 2 dimes, and 6 pennies. How much money do you have?

A. $0.46

B. $0.51

C. $0.55

Figure 1. Example easyCBM question (grade 2). This figure illustrates the minimal wording used in easyCBM-math questions.

Bob emails Matt every 3 days. Carl emails Matt every 4 days. Dave emails Matt every 6 days. Matt gets an email from each today.

How often will Matt receive email from all three on the same day?

A. Every 6th day
B. Every 12th day
C. Every 13th day
D. Every 72nd day

Figure 2. Example OAKS-math question (grade 3). This figure illustrates the relative greater words used in OAKS-math questions compared to easyCBM-math.
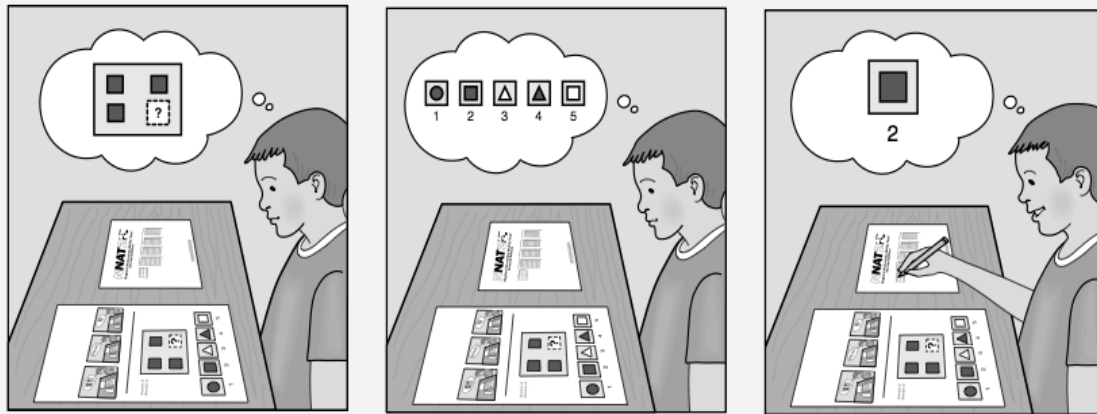
Figure 3. Pictorial representation of NNAT2 items. This figure illustrates the item format

and assessment procedure for the NNAT2.

Qualitative Description

| Extremely Low | Borderline | Low Average | Average | High Average | Superior | Very Superior |

Naglieri Ability Index

70   80   85   90   100   110 115 120   130

Percentile Rank

1   25   50   75   99
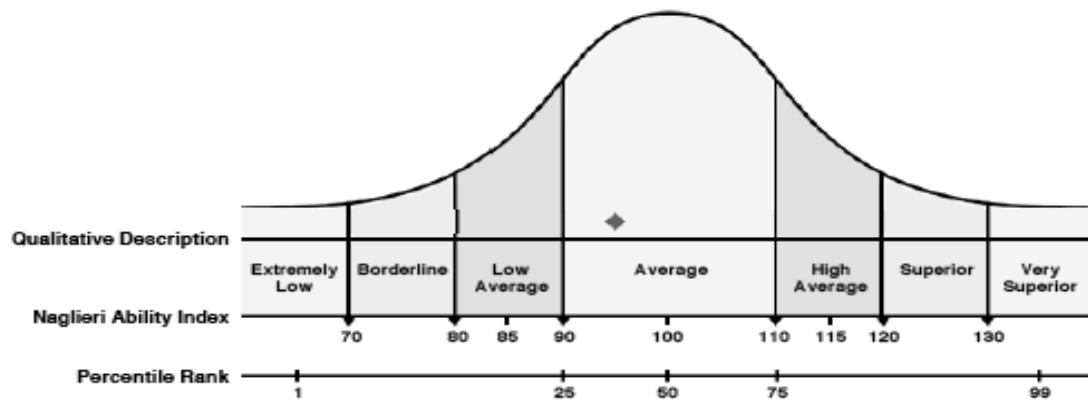
| 95 | **NAGLIERI ABILITY INDEX** |

This student's score has been placed on the graph to show how it relates to the scores of other students of the same age. Most children's scores will fall between 90 and 110. Scores in this range are considered to be average. Fewer children will have scores that fall between 70 to 89 and 111 to 130, and an extremely small number of children will have scores that fall below 70 or above 130. Use the Qualitative Description on the graph to help you understand where the student's score falls.

| 38 | **PERCENTILE RANK** |

Percentile Ranks range from a low of 1 to a high of 99. This score is used to compare the student's performance to that of a national sample of students of the same age. For example, if a child who is six years old has a Percentile Rank of 75, this means the student scored as well as or higher than 75 percent of six year old students who were tested. Please note that Percentile Rank and Percent Correct represent two different performance measurements and have different interpretations.

Figure 4. Student scoring printout (NNAT2). This figure illustrates the information included on the student scoring printout including ability index and percentile rank.
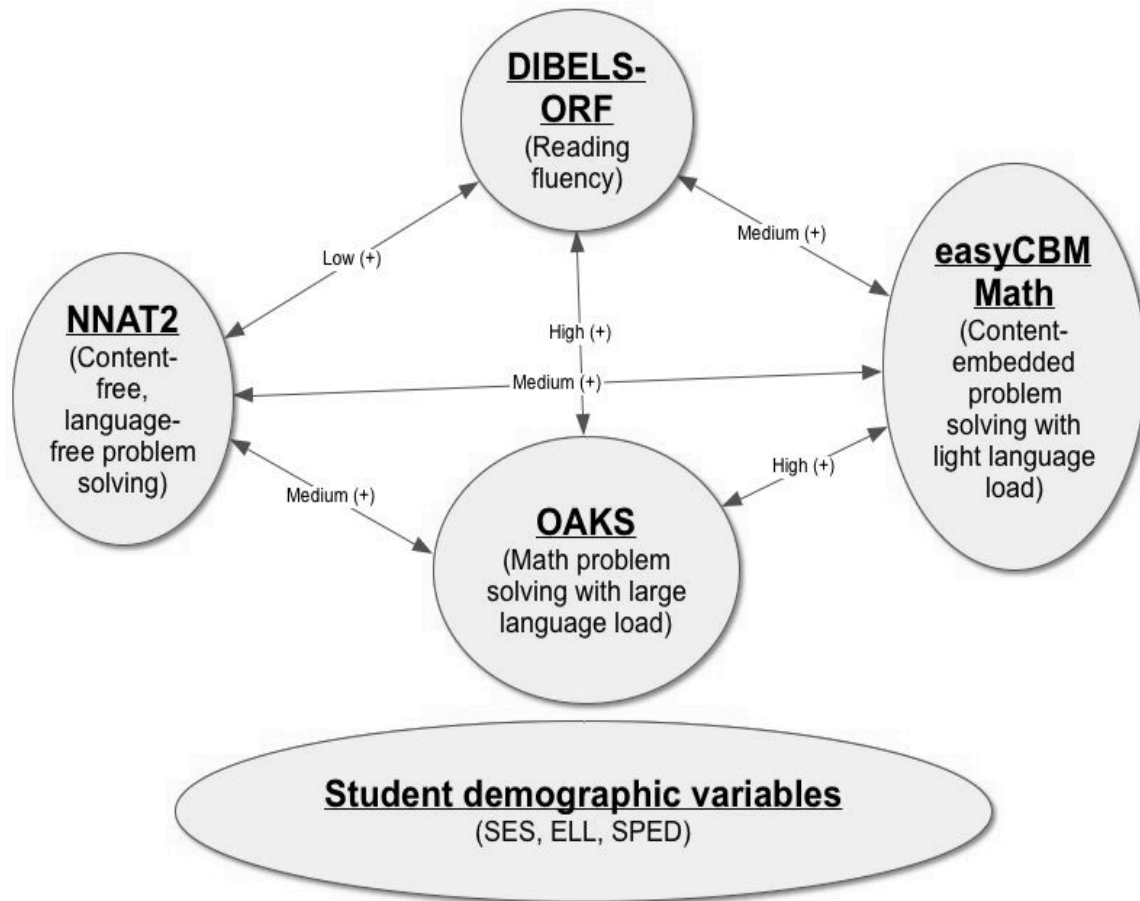
Figure 5. Possible relations among variables in math problem solving. This figure illustrates the possible relations between and among construct relevant and irrelevant variables in math problem solving. High, medium and low refer to possible (not actual) correlation degrees among variables.
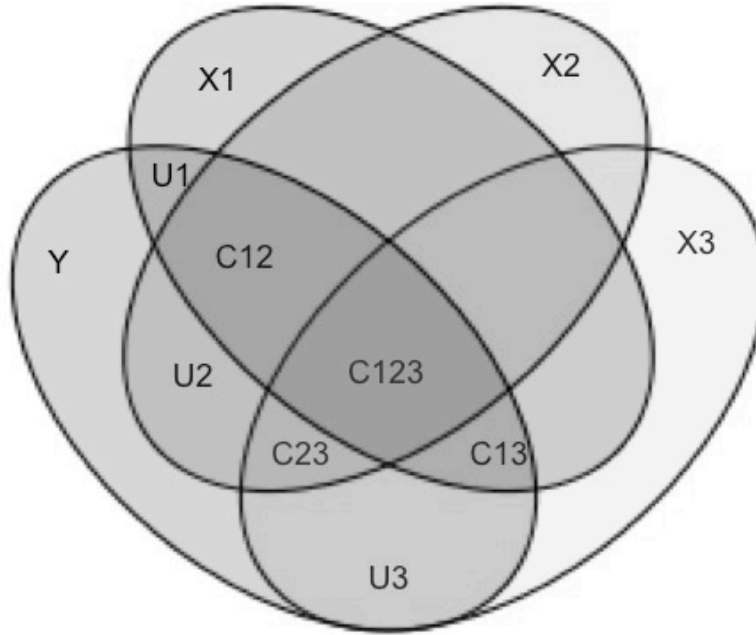
Figure 6. Variance partitioning using a commonality analysis. This figure illustrates the

unique and common variance between variables that were separated using a commonality

analysis. U = unique variance, C= common variance, 1=DORF, 2= NNAT2, 3=

easyCBM-math. Y= dependent variable (OAKS-math).

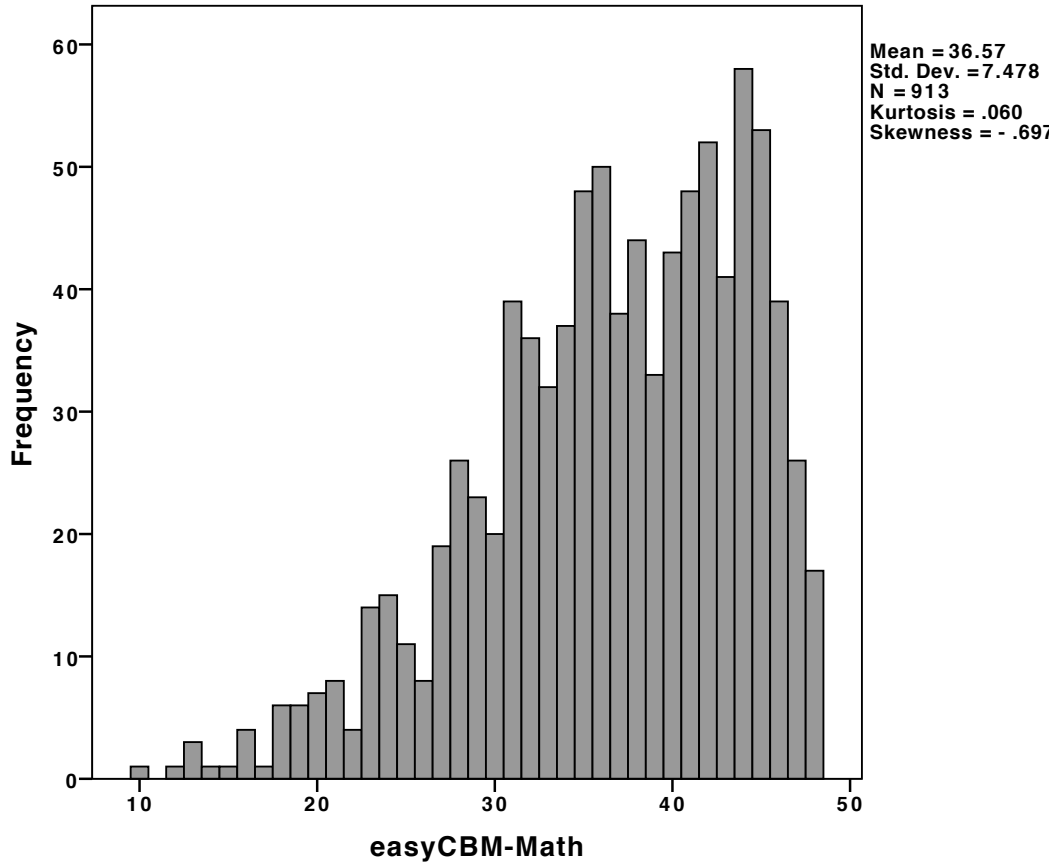**DISTRIBUTION OF SCORES FOR STUDY VARIABLES**

Figure 8. Distribution of easyCBM-math scores. This figure illustrates the mean,

standard deviation, number of cases, skewness, and kurtosis values for easyCBM-math
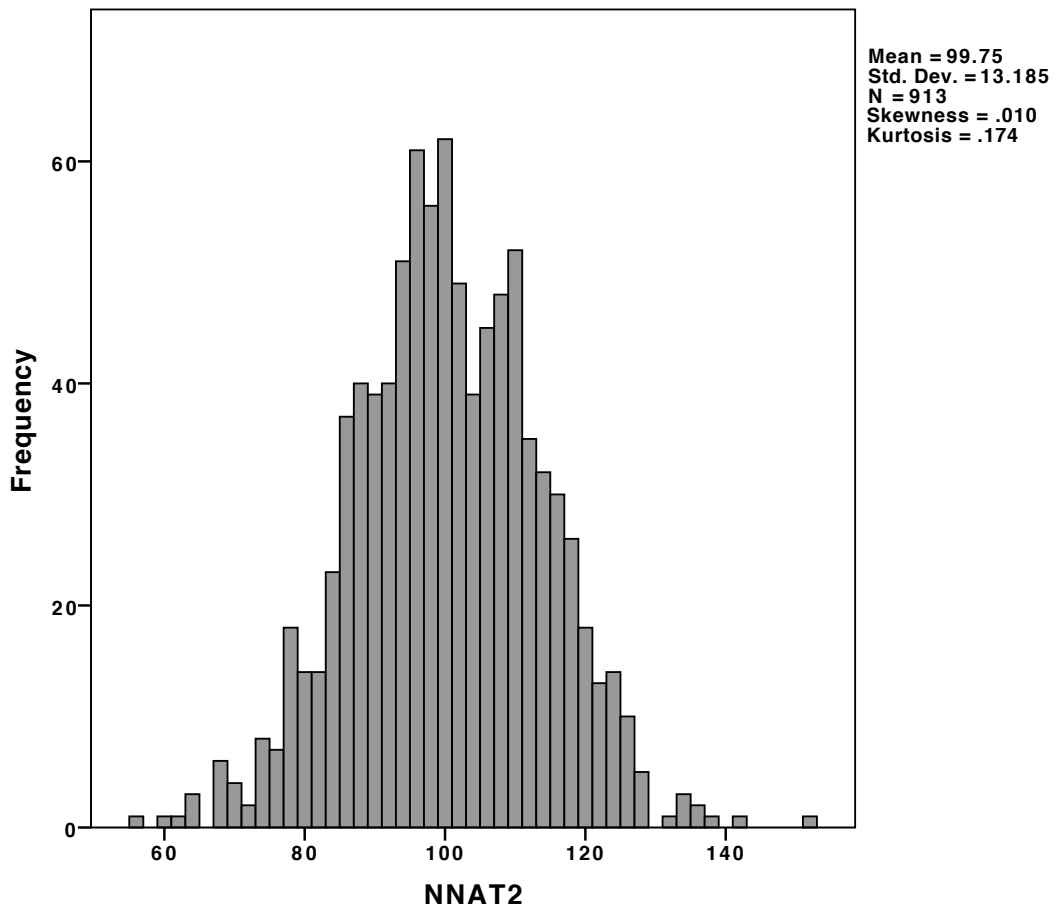
scores (grade 2).

Figure 9. Distribution of NNAT2 scores. This figure illustrates the mean, standard

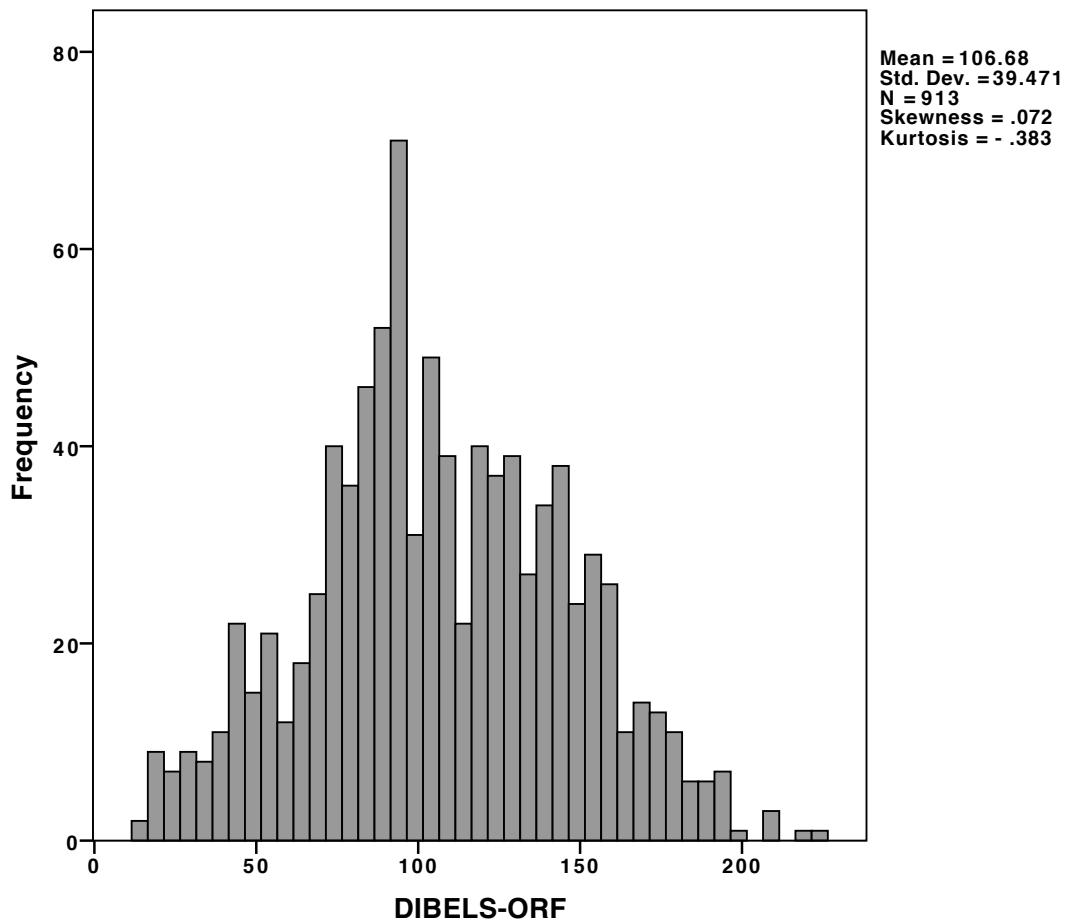deviation, number of cases, skewness, and kurtosis values for NNAT2 scores.

Mean = 106.68
Std. Dev. = 39.471
N = 913
Skewness = .072
Kurtosis = - .383

Figure 10. Distribution of DORF scores. This figure illustrates the mean, standard

deviation, number of cases, skewness, and kurtosis values for DORF scores (grade 2).
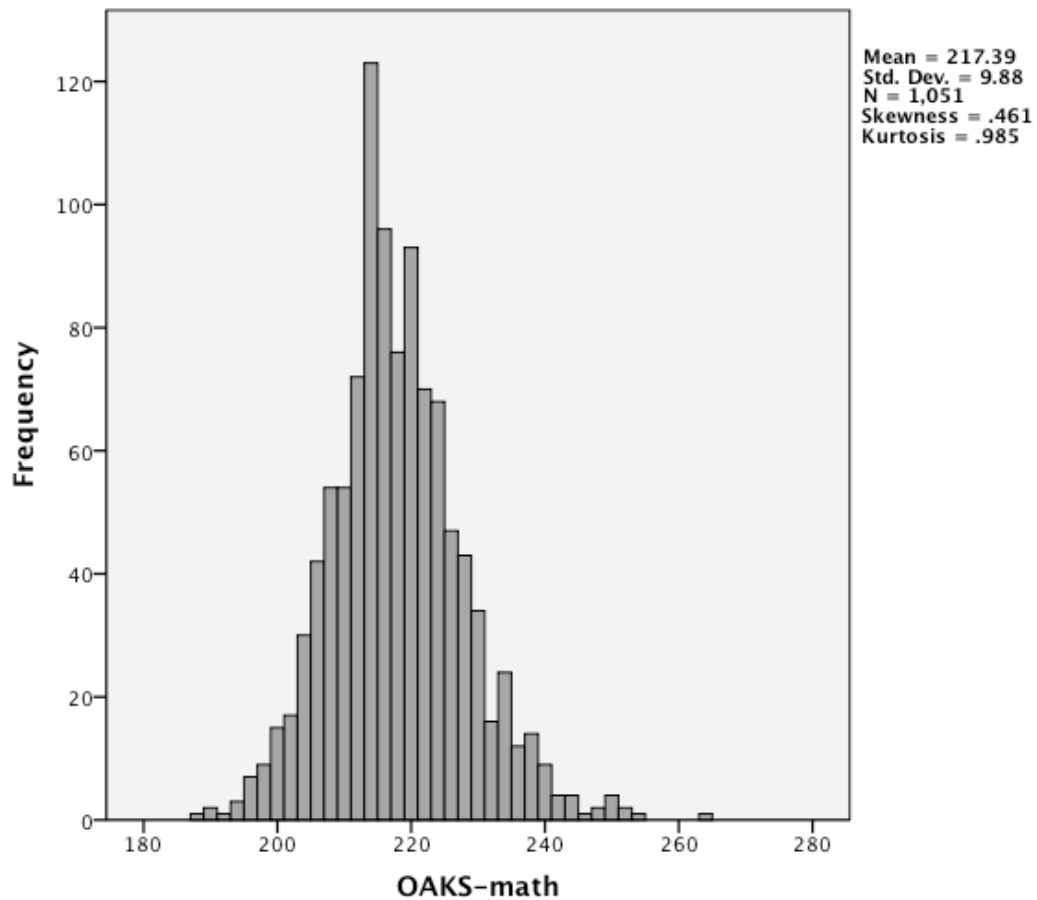
Mean = 217.39
Std. Dev. = 9.88
N = 1,051
Skewness = .461
Kurtosis = .985

Figure 11. Distribution of OAKS-math scores. This figure illustrates the mean, standard deviation, number of cases, skewness, and kurtosis values for OAKS-math scores (grade 3).

APPENDIX D

LITERATURE SEARCH DESCRIPTION

My search for literature on the topic of construct relevant and irrelevant variables in assessments related to math problem solving originated in electronic databases including ERIC, Academic Search Premier, and Google Scholar. I narrowed the search in Google Scholar to retrieve results published since 2006, while the other databases included all date ranges. I searched using various combinations of the following terms: *construct, validity, irrelevant, irrelevance, mathematics, problem solving, cognitive correlates, general ability, assessment, variance, variables, elementary, predictive validity, state assessment,* and *achievement*. The search combinations produced a group of 636 journal articles, theses, book chapters, and reports.

I further narrowed these search results based on my interests in (a) predictive variables related to outcomes on math assessments, (b) construct irrelevant variables studied in math assessments, and (c) the construct of math problem solving. Most often I excluded studies or research that did not address correlations to outcomes in math assessments. I also excluded research focused on the impact of specific interventions or on teacher background or training. With these restrictions, I reviewed 262 articles, chapters, theses, reports and studies in addition to their related reference pages.

I chose to not restrict my search to journal articles because many of the concepts and terms within this study are based on definitions created by national groups and measurement experts and found in books, reports, and presentations. Additionally, this area of research is relatively new, so by including the most recent work of students (reflected in theses) and experts (sometimes reflected in book chapters) I could more

accurately depict the current interest in and impact of construct relevant and irrelevant

variables without being restricted to studies with large effect sizes and large populations

(most frequently published in journals).

REFERENCES CITED

Abedi, J. (2006). Psychometric issues in the ELL assessment and special education eligibility. *Teachers College Record, 108*, 2282-2303.

Abedi, J., & Leon, S. (1999). *Impact of students' language background on content-based performance: Analyses of extant data.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Leon, S., & Mirocha, J. (2001). *Examining ELL and non-ELL student performance differences and their relationship to background factors: Continued analyses of extant data.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data* (CSE Tech. Rep. No. 603). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*, 219–234.

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Tech. Rep. No. 478). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16–26.

Abedi, J., Lord, C., Kim-Boscardin, C., & Miyoshi, J. (2000). *The effects of accommodations on the assessment of LEP students in NAEP* (CSE Tech. Rep. No. 537). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., & Plummer, J. (1997). *Language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Alonzo, J., Lai, C. F., & Tindal, G. (2009). *The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 2* (Technical Report No. 0920). Eugene, OR: Behavioral Research and Teaching: University of Oregon.

Alonzo, J., Tindal, G., Ulmer, K., & Glasgow, A. (2006). *easyCBM online assessment system.* http://easycbm.com. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). *Standards for Educational and Psychological Testing.* Washington, DC: AERA.

Anderson, D., Alonzo, J., & Tindal, G. (2010). *easyCBM Mathematics Criterion Related Validity Evidence: Oregon State Test* (Technical Report No. 1011). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2010). *easyCBM Mathematics Criterion Related Validity Evidence: Washington State Test* (Technical Report No. 1010). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Lai, C. F., Nese, J. F. T., Park, B. J., Sáez. L., Jamgochian, E. M., Alonzo, J., & Tindal, G. (2010). *Technical Adequacy of the easyCBM Primary-Level Mathematics Measures (Grades K-2), 2009-2010 Version* (Technical Report No. 1006). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Balboni, G., Naglieri, J. A., & Cubelli, R. (2010). Concurrent and predictive validity of the Raven Progressive Matrices and the Naglieri Nonverbal Ability Test. *Journal of Psychoeducational Assessment. 28*, 222-235. doi: 10.1177/0734282909343763.

Beede, D. N., Julian, T. A., Langdon, D., McKittrick, G., Khan, B. & Doms, M. E., Women in STEM: A Gender Gap to Innovation (August 1, 2011). Economics and Statistics Administration Issue Brief No. 04-11. Available at SSRN: http://ssrn.com/abstract=1964782 or http://dx.doi.org/10.2139/ssrn.1964782

Brody, N. (1992) *Intelligence*. San Diego: Academic Press

Burnett, K. & Farkas, G. (2009). Poverty and family structure effects on children's mathematics achievement: Estimates from random and fixed effects models. *The Social Science Journal, 46*, 297-318. Retrieved from http://dx.doi.org/10.1016/j.soscij.2008.12.009

Center for K-12 Assessment and Performance Management at ETS (2010, December). *Coming Together to Raise Achievement: New Assessments for the Common Core State Standards.* Retrieved from Education Northwest website: http://educationnorthwest.org/resource/1331

Center on Teaching and Learning (CTL). (2012). *2012-2013 DIBELS Data System Update Part I: DIBELS Next Composite Score* (Technical Brief No. 1202). Eugene, OR: University of Oregon.

Common Core State Standards Initiative (CCSSI). 2010. Common Core State Standards for Mathematics. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers. http://www.corestandards.org/

Conley, D. T. (2010). *College and Career Ready: Helping all Students Succeed Beyond High School*. San Francisco: Jossey-Bass.

Crawford, L., Tindal, G., and Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment, 7*, 303–323.

Davis, O. S. P., Arden, R., & Plomin, R. (2008). g in Middle Childhood: Moderate Genetic and Shared Environmental Influence Using Diverse Measures of General Cognitive Ability at 7, 9 and 10 Years in a Large Population Sample of Twins. *Intelligence. 36*(1), 68-80.

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

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology, 97*, 493-513. doi: 10.1037/0022-0663.97.3.493.

Fuchs, L. S. & Fuchs, D. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22*(1)*, 1-30. doi: 9607083090.

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M.,…Fletcher, J. M. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology, 98*(1), 29-43. doi:10.1037/0022-0663.98.1.29.

Garcia, S. B. & Tyler, B. (2010). Meeting the needs of English language learners with learning disabilities in the general curriculum. *Theory Into Practice. 49*, 113-120. doi: 10.1080/00405841003626585.

Good, R. H., Gruba, J., & Kaminski, R. A. (2009). *DIBELS Next*. Longmont, CO: Cambrium Learning Group.

Gustafsson, J.-E., & Undheim, J. O. (1992). Stability and Change in Broad and Narrow Factors of Intelligence from Ages 12 to 15 Years. *Journal of Educational Psychology. 84*, 141-49.

Haladyna, T. M. & Downing, S. M. (2004). Construct-Irrelevant Variance in High-Stakes Testing. *Educational Measurement: Issues and Practice 23*(1), 17-27. doi: 10.1111/j.1745-3992.2004.tb00149.x

Hart, S. A., Petrill, S. A., Plomin, R., & Thompson, L. A. (2009). The ABCs of math: A genetic analysis of mathematics and its links with reading ability and general cognitive ability. *Journal of Educational Psychology, 101*, 388-402. doi: 10.1037/a0015115.

Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., and Almond, P. J. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *The Journal of Educational Research*, *93*, 113-125.

Hyde, J. S., Lindberg, S. M., Linn, M. C, Ellis, A. B., & Williams, C. E., (2008). Gender similarities characterize math performance. *Science*, *527*(5888), 494-495.

Jensen, A. R. (2002). Psychometric g: Definition and substantiation. In R. J. Sternberg & R. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* Retrieved from http://read.amazon.com.

Jiban, C. L., & Deno, S. L. (2007). Using math and reading curriculum-based measurements to predict state mathematics test performance: Are simple one-minute measures technically adequate? *Assessment for Effective Intervention, 32*(2), 78–89.

Jitendra, A. K. (2005). Mathematics Assessment: Introduction to the Special Issue. *Assessment for Effective Intervention, 30*(2), 1-2. doi: 10.1177/073724770503000201.

Jordan, N. C., Kaplan, D., Olah, L., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development*, 77, 153–175.

Kaufman, A. S. (2009). *IQ Testing 101*. [Google Books version]. Retrieved from http://books.google.com/books?id=Z8i8LeV74m4C&printsec=frontcover&dq=history+of+intelligence+testing&source=bl&ots=cldxCOePny&sig=ZzwfoEOGhrDr_7tnTQ3d8VJx_tA&hl=en&sa=X&ei=IwhhUPDjNejKiwLYn4G4CA&ved=0CEUQ6AEwBA#v=onepage&q=history%20of%20intelligence%20testing&f=false.

Ketterlin-Geller, L. R., Alonzo, J., & Tindal, G. (2004). *Use of focus groups to inform the construction of a universally designed mathematics test* (Technical Report No. 29). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lamb, J. H. (2010). Reading Grade Levels and Mathematics Assessment: An Analysis of Texas Mathematics Assessment Items and Their Reading Difficulty. *The Mathematics Educator, 20*(1), 22-34.

Larsen, L., Hartmann, P., & Nyborg, H. (2008). The Stability of General Intelligence from Early Adulthood to Middle-Age. *Intelligence*. *36*(1), 29-34.

Mannamaa, M., Kikas, E., Peets, K., & Palu, A., (2012). Cognitive correlates of math skills in third-grade students. *Educational Psychology: An International Journal of Experimental Educational Psychology, 32*(1), 21-44.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*, 215-237.

Naglieri, J. A. (1997). *Naglieri Nonverbal Ability Test*. San Antonio, TX: Psychological Corporation.

Naglieri, J. A. (2008). *Naglieri Nonverbal Ability Test- Second Edition (NNAT2)*. San Antonio, TX: Psychological Corporation.

Naglieri, J. A. (2011). *Naglieri Nonverbal Ability Test Second Edition: Manual Supplement- Technical Information and Normative Data.* San Antonio, TX: Pearson.

Naglieri, J. A. & Das, J. P. (2002). Practical implications of general intelligence and PASS cognitive processes. In R. J. Sternberg & R. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* Retrieved from http://read.amazon.com.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author. Available: http://www.nctm.org/standards/content.aspx?id=16909.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through Grade 8 mathematics: A quest for coherence.* Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics (NCTM), National Council of Supervisors of Mathematics (NCSM), Association of State Supervisors of Mathematics (ASSM), Association of Mathematics Teacher Educators (AMTE). (2010). *Mathematics education organizers unite to support implementation of common core state standards.* Retrieved from http://www.nctm.org/standards/content.aspx?id=26088

National Governors Association Center for Best Practices and Council of Chief State School Officers. (2010) *National Governors Association and State Education Chiefs Launch Common State Academic Standards.* Retrieved from http://www.corestandards.org/articles/8-national-governors-association-and-state-education-chiefs-launch-common-state-academic-standards

National Mathematics Advisory Panel. *Foundations for Success: The Final Report of the National Mathematics Advisory Panel*, U.S. Department of Education: Washington, DC, 2008.

National Research Council. (2001). *Adding it up: Helping children learn mathematics.* J. Kilpatrick, J. Swafford, and B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Nese, J. F. T., Lai, C. F., Anderson, D., Jamgochian, E. M., Kamata, A., Sáez. L., Park, B. J., Alonzo, J., & Tindal, G. (2010). *Technical Adequacy of the easyCBM Mathematics Measures: Grades 3-8, 2009-2010 Version* (Technical Report No. 1007). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Nimon, K., Lewis, M., Kane, R., & Haynes, R. M. (2008). An R package to compute commonality coefficients in the multiple regression case: An introduction to the package and a practical example. *Behavior Research Methods, 40*(2), 457-466. doi: 10.3758/BRM.40.2.457.

Nordquist, L. (February 7, 2012). Personal communication.

Oregon Department of Education, Office of Assessment and Information Services. (2012). *Mathematics test specifications and test blueprints (grade 3).* Retrieved from http://www.ode.state.or.us/search/page/?id=496.

Pearson. (2012). *Introduction to the Naglieri Nonverbal Ability Test- Second Edition (NNAT2).* Retrieved from http://www.pearsonassessments.com/haiweb/Cultures/en-US/Site/Community/Education/Products/NNAT2/nnat2.htm.

Raven, J., & Raven, J. (2003). Raven Progressive Matrices. In R. Steve & R. S. McCallum (Eds.), *Handbook of nonverbal assessment* (pp. 223-237). New York: Kluwer.

Reeve, C. L., & Lam, H. (2005). The Psychometric Paradox of Practice Effects Due to Retesting: Measurement Invariance and Stable Ability Estimates in the Face of Observed Score Changes. *Intelligence. 33*, 535-549.

Rutherford-Becker, K. J. & Vanderwood, M. L. (2009). Evaluation of the relationship between literacy and mathematics skills as assessed by curriculum-based measures. *The California School Psychologist, 14*, 23-34.

Scafidi, T. & Bui, K., (2010). Gender Similarities in Math Performance from Middle School through High School. *Journal of Instructional Psychology*, *37*, 252-255.

Silberglitt, B., Burns, M. K., Madyun, N. H., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: a longitudinal comparison of grade levels. *Psychology in the Schools, 43*, 527-535. doi: 10.1002/pits.20175.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A Meta-analytic review of research. *Review of Educational Research, 75*, 417-453.

Slater, S. (September 4, 2012). Personal communication.

Spearman, C.E. (1904). General intelligence objectively determined and measured. *American Journal of Psychology, 15*, 201–293.

Sullivan, A. L. (2011). Disproportionality in special education identification and placement of English language learners. *Exceptional Children*, *77*, 317-334.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., and Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: an experimental study. *Exceptional Children, 64*, 439-450.

Thurber, R. S., Shinn, M. R., Smolkowski, K. (2002). What is measured in mathematics tests: construct validity of curriculum-based mathematics measures. *School Psychology Review, 31*, 498-513.

Webb, N. L. (1999). *Alignment of Science and Mathematics Standards and Assessments in Four States*. (Monograph No. 18). Madison, WI: National Institute for Science Education & Council of Chief State School Officers.

Wechsler, D. & Naglieri, J. A. (2006). *Wechsler Nonverbal Scale of Ability*. San Antonio, TX: Pearson.

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence*. San Antonio, TX: Psychological Corporation.

Wehmeyer, M. L. (2001). Disproportionate Representation of Males in Special Education Services: Biology, Behavior, or Bias? *Education & Treatment Of Children (ETC)*, *24*(1), 28.

Whitley, Samuel, "Oral reading fluency and MAZE selection for predicting 5th and 6th grade students' reading and math achievement on the Illinois Standards Achievement Test " (2010). Masters Theses. Paper 607. http://thekeep.eiu.edu/theses/607

Zientek, L. R. & Thompson, B. (2006). Commonality analysis: Partitioning variance to facilitate better understanding of data. *Journal of Early Intervention, 28*, 299-307. doi: 10.1177/105381510602800405.