The Development and Validation of a Diagnostic Assessment of Algebraic Thinking Skills for Students in the Elementary Grades

Nicole C. Ralston

A dissertation to be submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Dr. Min Li, Chair

Dr. Catherine Taylor

Dr. Elham Kazemi

Program Authorized to Offer Degree:

College of Education: Educational Psychology – Measurement, Statistics, & Research Design

University of Washington

Abstract

The Development and Validation of a Diagnostic Assessment of Algebraic Thinking Skills for Students in the Elementary Grades

Nicole C. Ralston

Chair of the Supervisory Committee:

Dr. Min Li

Educational Psychology: Measurement, Statistics, & Research Design

Elementary school students often exhibit a wide variety of different conceptions associated with algebraic thinking that their teachers fail to recognize or understand (Smith, diSessa, & Roschelle, 1994). It is crucial that elementary school teachers possess knowledge of the variety of different student conceptions and also boast abilities to address alternative conceptions, as students who are not provided with opportunities to think algebraically may continue to struggle with algebra throughout school (National Council of Teachers of Mathematics [NCTM], 2000; Smith et al., 1994). The purpose of this research, therefore, is to create and validate a diagnostic assessment tool for the elementary grades that can be used to help teachers better recognize and understand algebraic thinking conceptions. Although several researchers have created non-validated assessment items for use in intervention research (i.e., Jacobs et al., 2007) and others have begun work creating an algebraic thinking diagnostic tool for the middle and high school grades (i.e., Russell et al., 2009), no formally validated algebraic thinking diagnostic tools were discovered for elementary school students. Further, it appears that very few, if any, of the research studies conducted around algebraic thinking featured random samples or large sample sizes. Because of these gaps in the research, this inquiry seeks to address the following research

purposes: 1) Validate a diagnostic assessment tool for measuring the algebraic thinking skills of elementary school students; 2) Investigate the algebraic thinking skills elementary students currently possess; and 3) Cross-culturally validate this diagnostic assessment tool with students in Singapore. A random stratified cluster sample was therefore used in this study, with 1,745 elementary school students in grades 1-5 from six different schools in six different school districts in urban Washington State completing a diagnostic assessment of algebraic thinking skills. Teachers of these students completed two teacher surveys regarding their views towards teaching algebraic thinking skills and their use of the assessment results. An additional convenience sample of 1,619 students in grades 1-6 from four different schools in Singapore completed the diagnostic assessment. An additional convenience sample of 73 students in grades 1-5 from two different schools in urban Washington State were interviewed via a 'think-aloud protocol' to better understand student thinking while completing the assessment. Initial results demonstrated that the results of this assessment tool may be reliable and valid for the purposes described. Results indicated that although students experienced alternative conceptions discovered by other researchers, these were not occurring at nearly as high of rates as previously reported. Further results and the implications of these findings will be discussed.

Acknowledgements

This dissertation would not have been possible without the help of many people along the way. First and foremost, I would be nowhere without the love and support of my family. My wonderful parents, Grant and Joeleen, started me down this path from the day I was born, instilling in me a love of books and a constant quest for knowledge. Without their unwavering love and support I never would have made it through my undergraduate degree, much less through my Master's as well as this dissertation. In addition, my amazing husband Justin deserves a medal for his patience, support, and confidence in me through these many long years.

Second, I would certainly be lost without my amazing committee. I would like to express the deepest appreciation to my committee chair, Professor Min Li, who started me on this journey with her measurement course during spring quarter, 2008. Without her constant support, expertise, and knowledge this dissertation certainly would not have been possible. Further, it was her encouragement that led me to a summer of once-in-a-lifetime experiences in Singapore.

I would also like to thank my committee member, Professor Catherine Taylor, who constantly went above and beyond the call of duty to support student learning. Similarly, this work also would not have been possible without Professor Elham Kazemi's knowledge and interest in algebraic thinking, which started my interest in the concept in the first place. In addition, Professor Judith Arms deserves a special thank you for joining my committee during my hour of need despite her many other commitments and obligations.

Third, there are many others who could and should be thanked, from professors like Dr. Scott Chambers at Linfield College who first believed that I could accomplish this, to my many fellow graduate students who helped and provided me feedback along the way, to the friends who put up with hearing about my dissertation for so long. Thank you.

Finally, this work certainly could not have been conducted without the financial support from the National Science Foundation (NSF), a Washington Educational Research Association (WERA) grant, and a University of Washington Doi Doctoral Dissertation Award. I am very grateful to the generosity of these institutions.

Table of Contents

List of Tables

List of Figures

Chapter 1: Statement of Problem

Elementary school students often exhibit a wide variety of different conceptions associated with algebraic thinking that their teachers fail to recognize or understand (Smith, diSessa, & Roschelle, 1994). It is crucial that elementary school teachers possess knowledge of the variety of different student conceptions and also boast abilities to address alternative conceptions, as students who are not provided with opportunities to think algebraically may continue to struggle with algebra throughout school (National Council of Teachers of Mathematics [NCTM], 2000; Smith et al., 1994). In order to provide elementary school students with effective algebraic thinking opportunities, teachers must have a substantial knowledge base of algebraic thinking content, algebraic thinking knowledge held by students, and teaching methods (i.e., pedagogical content knowledge). Unfortunately, investigations regarding the pedagogical content knowledge elementary teachers possess has revealed that elementary teachers often hold very narrow views of algebra (i.e., that it involves only the manipulation of symbols) (Stephens, 2006; 2008). Because of these narrow views, it appears beneficial to create and validate a diagnostic assessment of algebraic thinking skills (DAATS) to inform teachers of student knowledge, which would ideally lead to a change in instructional decisions (if needed), and improved student learning.

History of Algebra in the U.S.

Algebra maintains a long-standing history in the U.S. Algebra first appeared as a topic in Benjamin Franklin's academy in 1751, was first acknowledged in Harvard's curriculum in 1786, and was first a required course for entry into Harvard in 1820 (Kilpatrick & Izsak, 2008). This boom in algebra popularity brought the percentage of U.S. high school students taking algebra to

57% in 1910; however, a high failure rate caused this number to plummet to less than 25% by the 1950's (Kilpatrick & Izsak, 2008). In the U.S. today, more than 90% of 17 year-old students and 29% of 13 year-old students have taken at least one algebra course, but algebra remains a considerable source of failure (Campbell, Hombo, & Mazzeo, 2000; Moses & Cobb, 2001). Despite appearing to have a long history of being a constant in America's school system, algebra has primarily existed as a specific subject to be studied only in the upper levels of middle and high schools (Smith & Thompson, 2008; Wagner & Kieran, 1989). Such courses often employ rule-based procedural teaching methods and focus heavily on solving equations (Kaput, Blanton, & Moreno, 2008).

Traditionally, elementary school mathematics focuses on arithmetic while middle and high school mathematics focuses on algebra (Kieran, 1992). Many researchers would disagree with this mentality, instead describing algebra as a set of skills that should be used across all mathematical topics, across all grades – kindergarten through twelfth grade – and not as a stand-alone topic (e.g., Algebra I & Algebra II in high school). Researchers have hypothesized that, by incorporating algebraic thinking skills in the curriculum earlier, the success rates in algebra in middle and high schools may increase. Kaput and Blanton (2005), for example, clarified this: "algebraic reasoning needs to develop over a long period of time in students' mathematical experience, beginning in the early grades and engaging most mathematical topics" (p. 100). This desire to include algebra in the math curriculum earlier - specifically providing opportunities to engage in algebraic thinking in the elementary grades - has only recently come to the attention of researchers due to poor algebra results in secondary schools (Carpenter, Levi, Berman, & Pligge, 2005). These disappointing results are perhaps best illustrated by the 2007 Trends in International Mathematics and Science Study (TIMSS, 2007), in which 8th grade U.S. students

scored 508 points on mathematics ability, only slightly above the international average of 500, but well behind the top three international leaders (i.e., Chinese Taipei at 598, South Korea at 597, and Singapore at 593). 8th grade students scored even lower (i.e., 501 points) on the algebra strand, which again was nearly exactly the international average (500) and statistically significantly lower than the top three international leaders (Chinese Taipei at 617, South Korea at 596, and Singapore at 579).

These results are not only worrisome because the U.S. is performing well below other developed countries, but also because in the U.S., algebra is a significant gateway course and an important determinant in students' mathematical lives; in that, often, it is after their experience with algebra that they determine if they would like to continue to pursue math (Mason, 2008). Students who successfully complete Algebra II, for example, are more likely to be successful in college and have higher employment earnings post-college: "students who complete Algebra II are more than twice as likely to graduate from college compared to students with less mathematical preparation" (National Math Panel, 2008, p. xiii). Further, the decline in mathematics achievement that begins in late middle school (i.e., often upon students' encounter with algebra) is so worrisome that the National Math Panel (2008) has placed algebra as one of its top concerns. It has even been referred to as a Civil Rights issue, in that "algebra, once solely in place as the gatekeeper for higher math and the priesthood who gained access to it, now is the gatekeeper for citizenship; and people who don't have it are like the people who couldn't read and write in the industrial age" (Moses, 2001, p. 14).

Although researchers recommend including algebraic thinking in the early grades, it appears most elementary teachers are unequipped to address algebraic thinking skills in schools. Researchers widely agree that algebraic thinking is not simply early algebra, yet elementary

teachers are not always on board with this. Instead, elementary teachers possess a fairly limited scope of algebraic thinking, believing it to mean solely symbol manipulation (NCTM, 2000; Stephens, 2006; 2008). Algebraic thinking at the elementary level involves not only symbolization, but also centers strongly around generalization (Blanton, 2008; Kaput, 2008). In terms of mathematics, "generalization is a statement that describes a general truth about a set of mathematical data" (Blanton, 2008, p. 3). Such generalization skills are required across the majority of tasks that might be considered 'algebraic' in nature.

Many elementary classrooms, however, are typically focused on the process of calculating correct answers instead of developing generalization and symbolization skills (Kaput & Blanton, 2005). Researchers have discovered that schools and teachers often believe that the need to teach arithmetic in the elementary grades greatly outweighs the need to teach algebraic thinking skills (Usiskin, 1995). Many of the researchers leading the charge on integrating algebraic thinking skills in the elementary grades have emphasized the connection between arithmetic and algebra, stating that instead of separating the two components they should be integrated (Carpenter, Franke, & Levi, 2003; Carraher & Schliemann, 2007). Carpenter and Levi (2000) fully support this integration: the "artificial separation of arithmetic and algebra deprives children of powerful schemes for thinking about mathematics in the early grades and makes it more difficult for them to learn algebra in the later grades" (p. 1).

Defining Student Conceptions

As discussed previously, students often experience a variety of different mathematical conceptions that their teachers may fail to recognize or understand (Smith, diSessa, & Roschelle, 1994). Alternative conceptions (i.e., preconceptions, misconceptions, etc.) are often defined as occurring when students use incorrectly learned strategies to solve new problems (Russell,

O'Dwyer, & Miranda, 2009). Research surrounding the word 'misconception' has evolved from indicating that students had a complete lack of knowledge in the area to beginning toward thinking of misconceptions as a piece of a conception, sometimes incomplete or lacking (Ely, 2010). Confrey (1990) was one of the first to suggest that the term misconception should be eliminated and replaced with simply conception, claiming that even Piaget "studied conceptions, not misconceptions" (p. 15). This term will therefore be used in this dissertation in accordance with Confrey.

Using this framework, it is certain that not all errors can be considered equal: some may be because of simple mistakes, algorithm issues, fatigue, working memory issues, etc. (Li & Li, 2008). Other errors may be grounded conceptually and can therefore be termed alternative conceptions (Li & Li, 2008). These wrong answers are not simple errors but, instead, are systematic in nature and arise from student past experiences with and misunderstandings of such problems (Russell et al., 2009). Further, not all correct answers can be considered equal either: sophistication and efficiency play a role in advanced student thinking as well. Although past researchers believed in simply replacing alternative conceptions with the correct conception, current researchers believe in helping student thinking evolve (i.e., building on previous understandings to add new knowledge) (Ely, 2010). The purpose of conducting this research would therefore be to better understand the range of different algebraic thinking conceptions, using the understanding of the student conceptions to provide tasks that will evolve the old thinking into new, more algebraically advanced student thinking.

Purpose of This Study

The specific purpose of this study is to create and validate a diagnostic assessment tool for the elementary grades that can be used to help teachers better recognize and understand

algebraic thinking conceptions. If teachers are unaware of such algebraic thinking conceptions, it appears that providing teachers with information regarding these conceptions (i.e., such as through a diagnostic measure of elementary school algebraic thinking skills) would benefit students. In this way, teachers armed with student conception information hold the potential to improve student learning and evolve current conceptions early (i.e., prior to the students' entry into middle and high schools and 'official algebra courses'). Choosing a diagnostic assessment (versus a different assessment method) often furnishes the capacity to provide informative and important information about student conceptions that summative or standardized assessments often do not (Nichols, 1994). Further, a diagnostic assessment was chosen because of NCTM's (2000) herald that, "assessment should enhance students' learning" and be "a valuable tool for making instructional decisions" (p. 22-23). Finally, Black and Wiliam's (1998) seminal work describes the research-base for the formative use of assessment to improve student outcomes. This diagnostic assessment tool will therefore offer such potential to support teachers' formative assessment practice to increase student outcomes in the area of algebraic thinking skills. The need for such a tool is illustrated through the following reasoning.

First, Russell, O'Dwyer, and Miranda (2009) have conducted some research in this area in that they developed an online algebra diagnostic assessment (DAAS) that middle school teachers can use to diagnose algebra conceptions. They discovered that using the full set of DAAS features (i.e., the diagnostic tests plus ability and misconception reports plus an intervention) had statistically significant positive effects on student algebra knowledge. This assessment tool was developed for and measured with students in 6^{th} through 12^{th} grade, not with students in the elementary grades. Therefore creation of a measure of the algebraic thinking

construct for elementary school students could have beneficial outcomes for both students and teachers.

Second, several researchers have created non-validated assessment items for use in intervention research (i.e., Jacobs et al., 2007) and others have begun work creating an algebraic thinking diagnostic tool for the middle and high school grades (Russell et al., 2009). Despite this work, no formally validated algebraic thinking diagnostic tools were discovered for elementary school students.

Third, it does not appear that extensive studies have been conducted on the current status of all dimensions of student algebraic thinking skills in elementary school students, despite the many studies existing investigating specific dimensions of algebraic thinking skills. Fourth and finally, no work was discovered investigating the relationship between student algebraic skills in the U.S. and student algebraic skills in Singapore, despite what we know about Singapore's academic successes. Because of these gaps in the research, this inquiry seeks to address the following research purposes:

1. Validate a diagnostic assessment tool for measuring the algebraic thinking skills of elementary school students;
2. Investigate the algebraic thinking skills elementary students currently possess;
3. Cross-culturally validate this diagnostic assessment tool with students in Singapore.

Significance of This Study

Kaput and Blanton (2005) summarized the issues described above: "disappointing results on student achievement tests in the United States can be traced to our nation's approach to algebra, which is to introduce it abruptly and late, isolate it in courses separated from other

mathematical subject matter, and teach it primarily as a series of procedural symbol-manipulation skills" (p. 99). Because of these issues, many researchers have recommended a call for action in incorporating algebraic thinking skills in the curriculum earlier (Carraher & Schliemann, 2007; Kaput, 2008; NCTM, 2000; RAND, 2003). The new Common Core Standards (2011), for example, which have been adopted by 45 of the 50 states, emphasizes "Operations and Algebraic Thinking" as one of the five core domains in the elementary grades (K-5). Further, the Common Core Standards for Mathematical Practice (SMP) demonstrate the need to incorporate algebraic thinking skills earlier, by highlighting ingraining in students the ability to reason abstractly and quantitatively (SMP 2), construct viable arguments and critique the reasoning of others (SMP 3), model with mathematics (SMP 4), look for and make use of structure (SMP 7), and look for and express regularity in repeated reasoning (SMP 8).

Others have demonstrated that teachers often misconstrue the meaning of algebraic thinking and fail to comprehend the algebraic thinking conceptions their students experience (Asquith et al., 2007). Discovering student conceptions and methods for teachers to recognize and address such conceptions is crucial, as students who possess alternative conceptions may continue to struggle with algebraic thinking throughout school. Improving algebraic thinking skills will ideally impact subsequent algebraic thinking skills of these students as they pursue mathematics throughout their educational careers. The National Math Panel (2008) described the importance of this study best: "longitudinal research is needed to identify early predictors of success or failure in algebra. The identification of these predictors will help to guide the design of interventions that will build the foundational skills needed for success in algebra." (p. 33).

Chapter 2: Algebraic Thinking in the Elementary Grades

Elementary school students often experience alternative conceptions associated with algebraic thinking that their teachers fail to recognize or understand (Smith, diSessa, & Roschelle, 1994). It is crucial that elementary school teachers possess knowledge of their students' conceptions and also boast abilities to address these conceptions, as students not provided with opportunities to think algebraically may continue to struggle with algebra throughout school (Smith et al., 1994; NCTM, 2000). In the next sections, I will first define algebraic thinking. I will then fully discuss the literature base of each of the three dimensions of algebraic thinking. Next, I will review the research on measures of algebraic thinking to demonstrate the research gap in assessing algebraic thinking skills in the elementary grades. If further information is desired, the Appendix A includes a description of both the factors that influence algebraic thinking and the factors influenced by algebraic thinking, including an illustrative nomological net.

Defining Algebra

Algebra and algebraic thinking may be abstract concepts to define, but it remains important researchers agree on common definitions of the terms. Some researchers, for example, have described school algebra as relationships-between-quantities (Chazan, 2000), while others describe it as an emphasis on "relationships among quantities, including functions, ways of representing mathematical relationships, and the analysis of change" (NCTM, 2000, p. 37). Usiskin (1988) provides several conceptions of algebra: 1) "algebra as generalized arithmetic" (i.e., a+b=b+a), 2) "algebra as a study of procedures for solving certain kinds of problems" (i.e., 5x+3=40), 3) "algebra as the study of relationship among quantities" (i.e., y=11x+b), and 4) "algebra as the study of structures" (i.e., factor $3x^2+4ax-132a^2$) (p. 11-15).

The disagreements regarding algebra are not limited to only the definition, but extend to theories concerning cognitive development. Past researchers, for example, have theorized that elementary school students may not be developmentally ready for algebra (i.e., that a cognitive gap or didactic cut exists) (Filloy & Rojano, 1984; Herscovics & Linchevski, 1994; Sfard & Linchevski, 1994). At this point it appears enough research (Brizuela & Schliemann, 2004; Carpenter & Levi, 2000; Carraher, Schliemann, Brizuela, & Earnest, 2006; Jacobs et al., 2007) has described the opposite to be true: under the right instructional circumstances, elementary school students can reason algebraically.

Defining Algebraic Thinking Skills

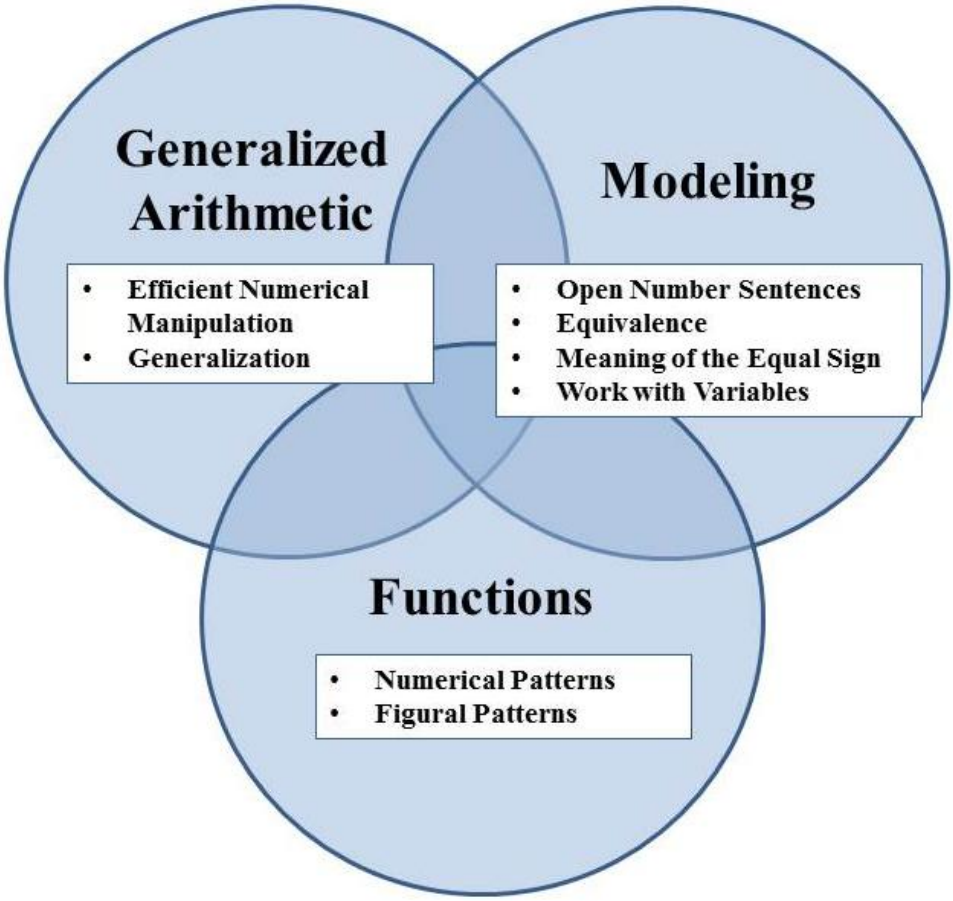
Given that it appears that elementary school students can reason algebraically, what exactly are algebraic thinking skills? Despite the many overlapping yet conflicting views of the definition of algebra and algebraic thinking, many agree that possessing strong algebraic thinking skills requires strong symbolization and generalization skills (Blanton, 2008; Kaput, 2008). Carpenter and Levi (2000) agree, defining algebraic thinking as "a) making generalizations and b) using symbols to represent mathematical ideas and to represent and solve problems" (p. 5).

Kaput (2008) elaborates on Carpenter and Levi's (2000) definition of algebraic thinking into a conceptual framework that will be utilized in this work. He theorizes algebraic thinking as containing two distinct aspects: generalization and symbolizing, which overarch three strands: generalized arithmetic, functions, and modeling (Kaput, 2008). The symbolizing aspect has been defined as "systematically symbolizing generalizations of regularities and constraints" (p. 11), while the generalization aspect has been defined as "syntactically guided reasoning and actions on generalizations expressed in conventional symbol systems" (p. 11). This model, adapted for

this specific work with elementary school students, is displayed below in Figure 1. The work of symbolizing and generalizing is required across all three strands, however the work of generalizing may be more prevalent within generalized arithmetic, while the work of symbolizing might be more prevalent in the work of modeling. It could be argued that many of the algebraic thinking skills described below may fit under more than one of Kaput's categories, and this is why it has been displayed as three over-lapping circles. In the next sections, I will describe the three dimensions in further detail, first defining the dimension, then describing the different types of tasks associated with that dimension, and finally discussing the common conceptions associated with those tasks.

Figure 1.

Kaput's (2008, p. 11) Algebraic Thinking Framework Adapted for This Work



Modeling

Modeling has been defined by Kaput (2008) as "the application of a cluster of modeling languages both inside and outside of mathematics" (p. 11), and typical tasks associated with modeling include unknowns in the sense of both open number sentences and equivalence, work with variables, and further work with equivalence including understanding the meaning of the equal sign.

Solving open number sentences. Though some researchers may disagree with categorizing solving open number sentences as algebraic thinking, solving open arithmetic sentences are often students' first, and most common, experience with algebra in the early grades (Carraher & Schliemann, 2007; Wagner & Kieran, 1989). The Progression of Common Core Standards (2011) has stated that by performing these types of problems, "students thus begin developing an algebraic perspective many years before they will use formal algebraic symbols and methods" (p. 13). Such tasks have long been used as methods of introducing students to thinking inversely about operations (i.e., using subtraction techniques to solve addition sentences) (Groen & Poll, 1973). Many researchers conceptualize algebraic thinking as being implicitly embedded in solving such missing-addend problems when they are forced to invent their own strategies for conducting the calculations.

Traditionally open number sentence problems have existed only in the format with the results unknown (a+b=__). The Progression of Common Core Standards (2011) now recommends students venture further into the algebraic realm by progressing with open number sentences from results unknown (a+b=__) to change unknown (a+__=c) to start unknown (__+b=c). Beyond this, these types of tasks should extend to subtraction and emphasize the conceptual understanding of connecting subtraction to addition. Further, the Common Core

Standards have stressed the importance of mastering such tasks: “the meaning of subtraction as an unknown-addend addition problem is one of the essential understandings students will need in middle school in order to extend arithmetic to negative rational numbers” (p. 13). In addition, students should be able to comprehend nonstandard open number sentence formats, which include sentences such as c=a+__ or c=__-b (Powell & Fuchs, 2010). In the past even simple versions of these non-traditional formats were postponed until official algebra classes in middle and high schools (Wagner & Kieran, 1989). It has now been found that even young children (i.e., in 1st and 2nd grades) can reason with open number sentences (Carpenter & Levi, 2000).

Students may experience alternative conceptions with these simple problems if they fail to understand the inverse relationship that exists between addition and subtraction, as well as commutative properties that exist (Carpenter et al., 2005). For example, students may read __-4=2 as ‘4 minus 2 equals something,’ (Lindvall & Ibarra, 1980). Alternative conceptions often occur surrounding these nonstandard formats when students read 5=10-__ as ‘something minus 10 equals 5’ (Lindvall & Ibarra, 1980). Overall, students tend to face more alternative conceptions with open number sentences when the answer is on the left side of the equal sign (i.e., nonstandard formats like c=a+__ or c=__+b), when the answer is on the left side of the equal sign and subtraction is involved (i.e., c=a-__), or when the blank is in the first position during subtraction problems (i.e., __-b=c) (Grouws & Good, 1976; Lindvall & Ibarra, 1980; Weaver, 1973).

Possessing the skills to solve open number sentences is important because acquiring the ability to reason with open number sentences can make learning other algebraic thinking skills (i.e., reasoning with zero, making conjectures and justifications, etc.) more accessible to young students (Carpenter & Levi, 2000). Further, some research has shown that working with such

nontraditional formats (i.e., ___=9+4) can increase student equivalence knowledge (see below for description of equivalence) (McNeil et al., 2011). It is also possible these types of problems actually require equivalence knowledge as a prerequisite skill.

Equivalence. Equivalence is perhaps the most heavily researched dimension of algebraic thinking skills, and research in the area of equivalence conceptions dates back to the 1970's (Behr, Erlwanger, & Nichols, 1976). Equivalence requires at least three needed knowledge components: "a) the meaning of two quantities being equal, b) the meaning of the equal sign as a relational symbol, and c) the idea that there are two sides to an equation" (Rittle-Johnson & Alibali, 1999, p. 177).

Commonly used prompts to assess the equivalence understanding are similar in nature to the open number sentence formats, however these items have at least two values on each side of the equal sign. For example, these items might include unknowns in the second position (a+__=c+d), third position (a+b=__+d), or even in the fifth position (a+b=c+d+__).

There is a growing research base displaying documentation that students often possess alternative conceptions surrounding equivalence and the equal sign, in that "children in the elementary grades generally consider that the equal sign means to carry out the calculation that precedes it; this is one of the major stumbling blocks when moving from arithmetic to algebra" (Carpenter et al., 2005, p. 84). Mann (2004), for instance, asked a group of 3rd grade students what they thought the equal sign meant, and the majority of the students said that it meant 'the answer is' (versus 'is the same as'). Carpenter, Franke, and Levi (2003) believe that three different alternative conceptions exist when solving equivalence problems: the answer is, using all the numbers (i.e., changing the number sentence), and extending the problem. It is believed

that such procedural errors result from a conceptual misunderstanding of the equal sign (i.e., viewing the equal sign operationally instead of relationally).

There is a long history of research documenting these equivalence conceptions persisting throughout the grade levels. In their often-cited study documenting this problem, Carpenter and colleagues (2005) found that less than 10% of students in all elementary grades, grades 1-6, showed understanding of the correct meaning of the equal sign when given the problem 8+4=__+5. Overwhelming responses involved interpreting the equal sign as 'put your answer here', in that the most popular student answers were 12 or 17. Surprisingly, grade level was not correlated with understanding of equality: 5% of 1st and 2nd graders answered correctly, 9% of 3rd and 4th graders answered correctly, yet only 2% of 5th and 6th graders answered correctly (Carpenter et al., 2005). Some researchers have argued that students may understand equivalence more than we realize, but they may simply be confused about the equal sign symbol itself. This idea was bolstered by research demonstrating that preschool-aged and 2nd grade students may understand the idea of equivalence when the equal sign is not featured (Mix, 1999; Sherman & Bisanz, 2009). 2nd grade students who were presented with equivalence problems without the conventional symbols, for example, performed at much higher rates than students presented with standard symbolic equivalence problems (Sherman & Bisanz, 2009). Other researchers have suggested that the concept of equivalence does not develop linearly with age. McNeil (2007), for example, sought to better understand the effects that age (i.e., students 7-11 years old) had on equivalence problems (i.e., 7+4+5=7+__) and disputed previous presumptions of a positive and linear association. She found poor performance overall and a contrasting U-shaped association between age and equivalence knowledge, which was replicated in a follow-up

study demonstrating that equivalence knowledge declines between ages 7 and 9 and then begins to improve between ages 9 and 11 (McNeil, 2007).

Student equivalence knowledge does not appear to improve drastically when students move from elementary school to middle school: in one study only 29%, 37%, and 44% of 6th, 7th, and 8th grade students, respectively, held relational views of the equal sign, while the majority held operational views (Asquith et al., 2007). These alternative conceptions held by middle school students have been discovered repeatedly in other studies (McNeil & Alibali, 2000; McNeil & Alibali, 2004; McNeil et al., 2006; Rittle-Johnson, 2006). Alibali and colleagues (2007) followed equivalence performance of middle school students longitudinally and discovered that improvement was gradual (i.e., students holding relational views of the equal sign grew from 20% at the beginning of 6th grade to 60% at the end of 8th grade) yet student equivalence knowledge remained worrisome even at completion of the study (i.e., nearly 30% of 8th grade students still held an operational view of the equal sign).

Little research was found with high school students. One of the few studies discovered included that of Russell, O'Dwyer, and Miranda (2009), who analyzed the results of their Diagnostic Algebra Assessment System (DAAS) and discovered that 11% of 905 students enrolled in an algebra class in 6th-12th grade held alternative conceptions surrounding equivalence. It may appear that this is a wide range of grade levels, but 90% of their student participants were in 8th or 9th grade, as these were Algebra I students. These researchers found significantly lower levels of equivalence alternative conceptions than found by Alibali and colleagues' (2007) findings described above, however these students were slightly older and the methods for discovering equivalence knowledge by Alibali and colleagues were more conceptual in nature than those used by the DAAS.

Several have argued that such alternative conceptions occur because of the types of instruction students receive in school (Kieran, 1981). Teachers, for example, often overestimate their students' abilities to solve equivalence tasks; believing that the majority of their students do indeed hold a relational view of the equal sign (Asquith et al., 2007). Other have argued that our emphasis on and methods of teaching arithmetic may actually hamper equivalence knowledge (McNeil, 2008), or that the way in which textbooks present such problems may be perpetuating such alternative conceptions (Seo & Ginsburg, 2003). Many textbooks reviewed by Seo and Ginsburg (2003), for example, displayed number sentences only in the traditional format: a+b=c, while others discovered that popular middle school textbook series (i.e., *Saxon Math*, *Prentice Hall Mathematics*, *Connected Mathematics*, and *Mathematics in Context*) rarely included equations with operations on both sides of the equal sign (McNeil et al., 2006). Capraro and colleagues (2007) summarize this issue: "findings indicate that misconceptions are still manifest in the U.S., and textbooks do little to mitigate the problem" (p. 87).

Mastering equivalence early appears to be a crucial skill prior to being successful in other areas of algebra. Alibali and colleagues (2007), for example, discovered that possessing a relational understanding of the equal sign was related to better performance on solving equivalence problems, and that the earlier students understood the equal sign relationally versus operationally, the higher they would perform on equivalent equation items. Other researchers also discovered that holding a relational view of the equal sign is actually predictive of higher scores when solving equations, even after mathematics ability is controlled for (Knuth, Stephens, McNeil, & Alibali, 2006).

Work with variables. The term variable is quite elusive and often difficult to define, which is quickly realized through the common exercise of trying to define it with only one word

(Schoenfeld & Arcavi, 1988). Because "the meaning of variable is variable", the wide variety of different acceptable meanings can make the term variable difficult for students to understand (Schoenfeld & Arcavi, 1988, p. 425). Researchers have long agreed that work with variables is an extremely important topic in middle and high schools; but the requirements of the skill itself are confusing and appear to change over time (Usiskin, 1988). The Common Core Progression (2011) recommends that "students in Grade 3 begin the step to formal algebraic language by using a letter for the unknown quantity in expressions or equations" (p. 27). Many researchers believe that the use of symbols in this manner does not denote true variable use because these symbols stand for one number only. This is unlike the true meaning of variable, for which varying quantities can replace it (Fujii & Stephens, 2008). While elementary school students may not be ready for this true meaning of variable, they certainly are ready to understand that the symbol itself stands for a number, which can occur through use of the literal symbols, and may help eliminate common alternative conceptions (i.e., that the letter is an alphabetic number) that may hinder more complex variable use later in middle and high school. Therefore the definition of work with variables in this work will involve using algebraic letters.

As with equivalence, past researchers believed that students were not capable of using algebraic letters until they had reached the appropriate Piagetian stages, and that therefore most 13-15 year old students could not interpret such algebraic letters correctly (Kuchemann, 1981). Again like equivalence, this theory has been disputed by more recent researchers, and it is instead believed that many students simply exhibit a wide variety of different conceptions in this area and may be in need of remedial instruction. "Students frequently base their interpretations of letters and algebraic expressions on intuition and guessing, on analogies with other symbol systems they know, or on a false foundation created by misleading teaching materials… Their

misinterpretations lead to difficulties in making sense of algebra and may persist for several years if not recognized and corrected" (MacGregor & Stacey, 1997, p. 15).

Kuchemann (1978; 1981) developed a framework of what he considered the six different student interpretations of variables: letter evaluated (i.e., the letter is a specific number, for example $a+3=5$), letter ignored (i.e., the letter is not given meaning, for example $a+b=43$ so $a+b+2=?$), letter as object (i.e., the letter stands for an object, for example *s* stands for students), letter as specific unknown (i.e., the letter is a specific yet unknown number, for example add 4 onto $n+5$), letter as generalized number (i.e., the letter can represent several numerical values), and letter as variable (i.e., the letter can represent a large range of unknown numerical values). Students' first introduction to work with variables often involves adding in literal symbols (i.e., *x* or *y*) into simple open number sentences or equivalence problems instead of blanks or boxes (i.e., $x+8=23$ instead of __+8=23) (Fujii & Stephens, 2008).

Despite variable use being stressed in the Common Core Standards starting in Grade 3, there is very little research surrounding the conceptions held by elementary school students in this area. It is clear, however, that alternative conceptions surrounding variables appear to persist with age. In Kuchemann's studies, many of the 13, 14, and 15 year-old students he assessed on variable knowledge only functioned at the lower levels of algebra knowledge, and many held alternative conceptions in the interpretation of letters (Booth, 1984). MacGregor and Stacey (1997) found that these common interpretations of variables for students ages 11-15 included believing the letter to stand for an abbreviated word, alphabetical value, particular numerical value, use of different letter, letter ignored, letter as label for an object, and letter equals 1. Further, when presented with the task '2n+3, what does the symbol stand for?' (Knuth et al., 2005; Asquith et al., 2007), the percentage of 6^{th}, 7^{th}, and 8^{th} grade students who

understood was 46%, 63%, and 76%, respectively. The majority who answered incorrectly either didn't know, or believed it to stand for an object, word, or a specific digit. This indicates that middle school students may still be interpreting letters as letters (or specific corresponding numbers) instead of understanding their purpose in representing numbers or a range of numbers. One of the most popular alternative conceptions held by middle and high school students appears to be misunderstanding literal symbols as labels (i.e., *c* stands for cat, so 4*c* might mean 4 cats) (Booth, 1984; MacGregor & Stacey, 1997). McNeil, Weinberg, and colleagues (2010) investigated how these conceptions manifest when utilizing variables using mnemonic letters (i.e., *c* for price of a cake), non-mnemonic letters (i.e., *x* or *y*), or Greek letters (i.e.,Φ or Ψ). They presented 322 middle school students with a commonly used problem adapted from Kuchemann (1978, 1981) using these three different types of variables as conditions: 'Cakes cost *c* dollars each and brownies cost *b* dollars each. Supposed I buy 4 cakes and 3 brownies. What does 4*c* + 3*b* stand for?' Students in the mnemonic letters condition misinterpreted the expressions the most often, and often considered the labels as standing for objects. Approximately 37% of students in the mnemonic (i.e., c and b) condition interpreted the variables correctly while approximately 56% of students in the non-mnemonic conditions interpreted the variables correctly. There was no difference between the non-mnemonic letters and the Greek letters groups (McNeil, Weinberg et al., 2010).

Mastering an understanding of variables and work with variables is significant because, if not remedied early, these alternative conceptions appear to persist into high school and even into adulthood. Russell, O'Dwyer, and Miranda (2009) analyzed the results of the DAAS and discovered that 14% of 905 students enrolled in an algebra class in 6^{th}-12^{th} grade held alternative conceptions surrounding variables. One of the most common alternative conceptions held by

older students is the 'reversal error', which occurs when the variables are reversed in formulas and has been found to be highly prevalent in both high school and college students (MacGregor & Stacey, 1993; Fisher, Borchert, & Bassok, 2010). A seminal study utilizing the notoriously famous 'Student-Professor Problem' (Kaput & Clement, 1979) demonstrated that even mathematically-proficient adults (i.e., college students pursuing traditionally mathematically based majors) often seriously struggle with representing mathematical relationships with variables and commit reversal errors. This was demonstrated when between 40% and 60% of adults solved the following problem incorrectly: "Write an equation using the variables *S* and *P* to represent the following statement: 'There are six times as many students as professors at this university.' Use *S* for the number of students and *P* for the number of professors" (p. 288). The most common error made involved reversing the solution: '6*S*=*P*'. In a follow-up study Rosnick (1981) revealed that a large proportion (i.e., 40-43%) of college students could not identify the *P* to mean number of professors or the *S* to mean number of students, revealing the common alternative conception that occurs when students misunderstand the *S* to mean students (instead of number of students) and therefore reading *S*=6*P* as "one student for every six professors" (p. 419). Because of these alternative conceptions that even adults appear to possess, it is crucial that students gain experience working with variables at an early age.

Generalized Arithmetic

According to Kaput (2008), generalized arithmetic is "the study of structures and systems abstracted from computations and relations, including those arising in arithmetic (algebra as generalized arithmetic) and in quantitative reasoning" (p. 11), and includes: efficient numerical manipulation (i.e., simplify calculations using number relations and compensation strategies) and generalizing (i.e., utilizing mathematical properties like the commutative property, property of

zero, etc.) (Kaput, 2008). Blanton (2008) elaborates on this, defining generalized arithmetic as "building generalizations about operations on and properties of numbers" (p. 5). It appears, however, that Blanton (2008) may have considered the modeling strand as combined with generalized arithmetic.

Efficient numerical manipulation. It is important students learn to manipulate numbers and use relational thinking strategies to solve problems more efficiently (Carpenter et al., 2003). Possessing the capacity to do so has been renamed 'efficient numerical manipulation' for this work. Take the problems 67+83=__+82 and 54+37–36 =__, for example. Although students can solve such problems by calculation alone, understanding the mathematical relationships between the numbers and possessing the algebraic thinking skills to identify the potential to simplify such problems are necessary for more complex problems in later grades (Stephens, 2008). Such tasks require more than just the arithmetic 'tricks' and algorithms; the tasks require students to identify the relationships between numbers and to reason about numerical transformations (Jacobs et al., 2007).

In the examples displayed above, students are not experiencing alternative conceptions but are instead simply using less sophisticated, inefficient, or less advanced strategies. An inefficient conception exhibited by students in this area, for example, is that they often fully calculate the equation (i.e., to solve 25+59-59 they will add 25+59 to obtain 84, then subtract 59 from 84 to obtain 25) instead of recognizing the simplification procedure that can take place. Although this is the correct conception, it is lacking in efficiency and sophistication. Such simplification exercises can also benefit students in the area of work with variables, because while elementary school students may not be able to fully grasp the concept of varying quantities such as 56-a+a=56, they can begin to demonstrate that they understand that 56-22+22=56 (Fujii

& Stephens, 2008). Further, students capable of generalizing and using the distributive and associative problems (i.e., such as being able to solve 101x46 by splitting it into 100x46+1x46) are usually more prepared to learn conventional algebra in official algebra courses in middle and high school (Baek, 2008).

Little research at any grade level was discovered in the area of efficient numerical manipulation, despite the fact that it is agreed that such tasks are important and meaningful. This lack of research may be partially due to the fact that this is a very difficult skill to measure, as it is nearly impossible to identify student thought processes without the use of an interview (i.e., a think-aloud protocol) (Ericsson & Simon, 1984). Regardless, it is agreed upon that possessing the ability to manipulate numbers efficiently is an important and significant ability needed to progress to higher order algebraic thinking skills.

Generalization. Generalization, or generalizing, requires possessing the ability to identify generalizations concerning the fundamental properties of numbers. Generalization requires the capacity to make a general statement using inferences from specific cases, as this relates to mathematics. Students, for example, should be able to understand and apply common mathematical properties such as odd/even numbers, doubling, commutative property, distributive property, etc.

Kilpatrick and Izsak (2008) provide "examples of deliberate generalization that appear to be accessible to elementary school students include perceiving and expressing patterns and deducing some general properties of numbers, such as a-a=0 for all whole numbers" (p. 12). The Progression of Common Core Standards (2011) recommends that over the course of the elementary grades students will begin to "build their understanding of the properties of arithmetic: commutativity and associativity of addition and multiplication, and distributivity of

multiplication over addition" (p. 3). To fully demonstrate understanding of the associative and distributive properties students must also demonstrate understanding of the order of operations and the uses of parentheses (Progression of Common Core Standards, 2011).

An alternative conception commonly experienced by students in this area is that they do not recognize that a mathematical procedure that worked in one equation may apply to another equation. Students who see a+b=b+a may, for example, not understand that this same property can be illustrated with x+y=y+x, or that the important feature in this statement is the + (not the variables themselves), as this is a statement about the commutativity of addition (Schifter, 1999). Further, many students who may understand the concept of commutativity when numbers are involved (i.e., 2+3=3+2) may fail to understand the principle when variables are included (i.e., a+b=b+a) (Baroody & Gannon, 1984). It is therefore likely that work with variables is a prerequisite skill for generalization tasks.

Stephens (2005), for example, sought to investigate these conceptions by asking 371 middle school students if h+m+n=h+p+n was always, sometimes, or never true. While less than 50% of these students answered correctly ('sometimes'), a large proportion of the students answered 'never'. A smaller group of students were interviewed and Stephens found that the alternative conception that two different variables cannot be the same value persisted across a wider variety of questions. These results were deemed unsurprising because the idea that two different variables (i.e. *m* and *p*, for example), can be equivalent "is a mathematical convention, not a notion that is intuitively obvious" (Stephens, 2005, p. 97). It is further particularly important that students understand the different properties associated with zero and 1, "because students often confuse these three patterns: $n + 0 = n$ but $n \times 0 = 0$, and $n \times 1$ is the pattern that does not change n" (Progression of Common Core Standards, 2011, p. 26).

Understanding such mathematical properties is important because of the inherent requirements of knowledge of decomposing and recombining numbers; students must understand that larger numbers are made up of smaller numbers that can be manipulated (Geary, 2006). This understanding of the properties and meanings of operations will benefit students as they progress to more difficult algebraic problems in the upper grades. Carpenter and colleagues (2005) have further emphasized the importance of such skills, by stating that "the best students have always figured out generalizations, and by doing so they make mathematics easier to learn and apply. Making generalizations explicit so that they are available to all students can address important issues of equity and access to powerful ideas of mathematics" (p. 97).

Functions

Kaput (2008) defines functions as "the study of functions, relations, and joint variation" (p. 11). Although explicit work with functions is often a term saved for middle and high school grades, students can begin this work in the elementary grades in working with patterns, recursive thinking (i.e., describing sequential change), and the overall study of mathematical change (NCTM, 2000; Warren & Cooper, 2005). "The construction and use of functions is considered to be central to most mathematical investigations and has been found to be notoriously difficult for most students at all levels of learning" (Warren, Cooper, & Lamb, 2006, p. 209). Although again past researchers have believed elementary school students to not be capable of functional thinking many have shown the opposite to be true: young children are capable of taking on a functional perspective (Martinez & Brizuela, 2006; Nunes, Schliemann, and Carraher, 1993; Stacey, 1989; Warren, Cooper, & Lamb, 2006).

Work with patterns. Patterns may be the algebraic thinking skill most emphasized by teachers in the elementary grades, and often is one of a student's first experiences with algebraic

thinking (NCTM, 2000). At the elementary level, working with patterns is defined as possessing the ability to recognize, describe, extend, and create patterns. Patterns could be defined as something that is predictable or exhibits some form of regularity: a 'rule' of sorts could be used to define that grouping of numbers, shapes, figures, etc. Understanding that a rule can be found in such patterns, even if the students cannot themselves find it, is crucial in moving from using recursive strategies to more algebraic strategies (Moss, Beatty, Barkin, & Shillolo, 2008). The Progression of Common Core Standards (2011) recommends students begin "reasoning about number or shape patterns, connecting a rule for a given pattern with its sequence of numbers or shapes" by at least the 4th grade, often generalizing such patterns to the 100th term (p. 30). Three different categories of patterns exist: repeating patterns, linear patterns, and nonlinear patterns. Further, these three types of patterns can be numeric or figural in nature.

Students often encounter patterns early in school with repeating patterns, which repeat over time in some predictable way and while they can be numerical (i.e., 1, 3, 6, 1, 3, 6, 1, 3, 6), they are more often figural (i.e., ). Because patterns are so heavily emphasized in early elementary school, it is likely the majority of elementary school students will have mastered repeating patterns, and may not be experiencing any alternative conceptions in this area. Repeating patterns is instead a type of prerequisite for mastering more difficult types of patterns. It is therefore appropriate for elementary school students to begin generalizing about what the 10th or 20th shape or number will be (Warren & Cooper, 2008).

Students are next introduced to linear patterns, which grow or shrink in some predictable linear way (i.e., add 4 to each number or subtract 2 from each number). Again, these can be either numerical or figural, but young students tend to have many more opportunities to experience either repeating figural patterns or linear numerical patterns; it is rarer for them to

experience figural linear patterns. In the elementary grades these numerical linear patterns often typically exist in the form of input-output tables (Tanish, 2011; Warren, Cooper, & Lamb, 2006). The majority of elementary students have had opportunities to experience numerical linear patterns and likely are not experiencing many alternative conceptions in this area. Again, this is a prerequisite for solving more difficult, generalized types of pattern items. It is unfortunate that figural patterns are not often emphasized, however, because figural patterns can serve three important roles: to represent numerical patterns visually, to serve as an initial induction to the idea of variable, and to create equivalent expressions (Warren & Cooper, 2008). One study, for example, found that only about half of 45 elementary school students solved figural linear patterns correctly (Warren & Cooper, 2008). Common alternative conceptions held by students included misunderstanding the number the pattern is growing by leaving the prompt blank or not seeing how the figure grows in more than one direction, if applicable (Warren & Cooper). When students are confronted with figural patterns, a common student approach is to simply translate the figural pattern into a numerical pattern, which unfortunately eliminates the opportunity for the student to better understand the relationship between the change in the physical figure and the change in value (Billings, Tiedt, & Slater, 2008).

Nonlinear patterns change over time in some predictable, yet nonlinear, way. Students in elementary and middle schools often experience alternative conceptions when faced with both numerical and figural nonlinear patterns because they often expect patterns to always be linear in nature. It becomes more difficult for students to understand how the pattern is changing when the change is nonlinear in nature. Nonlinear patterns are considered the most challenging by students, particularly those patterns that use more than one type of operation (i.e., the pattern 4, 10, 22, 46 involves adding 1 then multiplying by 2) (NCTM, 1997).

The more complicated step beyond any of these basic types of numerical or figural patterns involves combining these with generalizations (i.e., what will the 20th shape / figure / number be?) (Warren & Cooper, 2008). Common alternative conceptions when generalizing, include the failure of linking the figure number to the pattern, of generalizing too quickly, of generalizing the wrong aspect of the pattern, or of spotting trivial patterns (Driscoll, 1999). It is also likely that many students simply haven't experienced many generalizing pattern problems, and therefore will not be sure how to approach such a problem.

Early work with functions through patterning has been stressed as essential, especially in providing early experiences of working with variables and relationships between variables in the movement toward eventual work with functions (Wagner & Kieran, 1989). Some have even stated that "recursive thinking is a vital part of algebraic thinking and reasoning at all levels", and that students need opportunities to experience creating, observing, describing, and continuing a variety of different kinds of patterns (Bezuszka & Kenney, 2008, p. 81). NCTM (1997) also recommends student participation in patterning activities, because as students gain experience with patterns, they also gain initial experiences with functions, and these experiences provide a "bridge from numeric work at the elementary level to more general, symbolic algebra at the secondary level" (p. 228). By working recursively, students also start to understand the limitations of such recursive processes (i.e., when asked to provide the 10th or 20th term in a sequence) and can begin to work towards higher levels of abstraction (Bezuszka & Kenney, 2008; Lannin, Barker, & Townsend, 2006).

Standards Featuring Algebraic Thinking Skills

The above sections detail a thorough discussion of the three dimensions of algebraic thinking and the research that has been conducted in each area. It is important to consider where

each of these dimensions fits into the standards required by both the local state government and the national government. Therefore, Table 1 below details the skills under each dimension and how they are defined or found in both the Washington State Standards and the Common Core Standards (2011). Further, as discussed previously, many of the algebraic thinking skills are emphasized through the Common Core Standards for Mathematical Practice.

Table 1.

Common Core and Washington State Standards for Algebraic Thinking Skills

Algebraic Dimension	Dimension Tasks	Washington State Grade Level & Standard	Common Core Grade Level & Standard
Modeling	Solving open number sentences	2nd: "Solve equations in which the unknown number appears in a variety of positions" (2.2.G)	1st: "Determine the unknown whole number in an addition or subtraction equation relating three whole numbers" (1.OA) 3rd: "Determine the unknown whole number in a multiplication or division equation relating three whole numbers" (3.OA)
	Understanding equivalence	2nd: "Solve equations in which the unknown number appears in a variety of positions" (2.2.G) 3rd: "Determine whether two expressions are equal and use '=' to denote equality" (3.5.A)	1st: "Understand the meaning of the equal sign, and determine if equations involving addition and subtraction are true or false" (1.OA)
	Work with variables	4th: "Represent an unknown quantity in simple expressions, equations, and inequalities using letters, boxes, and other symbols" (4.4.A)	3rd: "Solve two-step word problems using the four operations. Represent these problems using equations with a letter standing for the unknown quantity" (3.OA)
Generalized Arithmetic	Efficient Numerical Manipulation	2nd: "Add and subtract two-digit numbers efficiently and accurately" (2.2.C)	1st: "Understand the meaning of the equal sign, and determine if equations involving addition and subtraction are true or false" (1.OA) 2nd: "Fluently add and subtract within 20 using mental strategies" (2.OA)
	Generalization	3rd: "Determine whether two expressions are equal and use '=' to denote equality" (3.5.A)[a]	1st : "Apply properties of operations as strategies to add and subtract" (1.OA) 3rd: "Apply properties of operations as strategies to multiply and divide" (3.OA)
Functions	Repeating Patterns	K: "Copy, extend, describe, and create simple repetitive patterns" (K.2.A).	4th: "Generate a number or shape pattern that follows a given rule. Identify apparent features of the pattern that were not explicit in the rule itself" (4.OA)
	Linear patterns	1st: "recognize, create, and extend number patterns" (1.2.I) 2nd: "Create and state rule for patterns that can be generated by addition and extend the pattern" (2.2.F)	3rd: "Identify arithmetic patterns and explain them using properties of operations" (3.OA) 4th: "Generate a number or shape pattern that follows a given rule. Identify apparent features of the pattern that were not explicit in the rule itself" (4.OA)
	Nonlinear patterns	5th: "describe and create a rule for numerical and geometric patterns and extend the patterns" (5.4.A)	4th: "Generate a number or shape pattern that follows a given rule. Identify apparent features of the pattern that were not explicit in the rule itself" (4.OA)

[a] Explicit generalization does not appear to exist in the Washington State Standards

Review of Measures on Algebraic Thinking Skills

The Need for a Diagnostic Measure of Algebraic Thinking Skills

It has been well documented that, although students often exhibit a wide variety of different conceptions in the area of algebraic thinking skills, even young students are developmentally ready (despite hypotheses of the past) and can learn algebraic thinking skills (Carraher, Schliemann, Brizuela, & Earnest, 2006). It has been further documented that students offer a wide variety of different conceptions surrounding algebraic thinking skills. Discovering these student conceptions as well as methods for teachers to recognize and address such conceptions is crucial, as some have hypothesized that students who possess alternative conceptions may continue to struggle with algebraic thinking throughout school (NCTM, 2000; Smith et al., 1994).

It appears that some teachers are unaware of the continuum of algebraic thinking conceptions. Because of this, it appears that through providing teachers with information regarding these conceptions (i.e., such as through a diagnostic measure of elementary school algebraic thinking skills), teachers may gain the potential to address such conceptions early (i.e., prior to the students' entry into middle and high schools and 'official algebra courses'). Access to a diagnostic assessment tool for the elementary grades may help teachers better recognize and understand algebraic thinking conceptions. Choosing a diagnostic assessment (versus other different assessment methods) often furnishes the capacity to provide informative and important information about student conceptions that summative or high-stakes assessments often do not (Nichols, 1994). The purpose of diagnostic assessment, after all, is to diagnose, or "the act of precisely analyzing a problem and identifying its causes for the purpose of classification-based decision making" (Rupp, Templin, & Henson, 2010, p. 1). Further, diagnostic assessments have been recommended by NCTM (2000) because: "assessment should enhance students' learning"

and be "a valuable tool for making instructional decisions" (p. 22-23). Black and Wiliam's (1998) seminal work describes the extensive research-base for the use of assessment to improve student outcomes. The development of a diagnostic measure of the algebraic thinking construct for elementary school students could, therefore, have the potential to produce beneficial outcomes for both students and teachers.

Current Measures of Elementary School Algebra

Because algebraic thinking skills have recently become a popular topic of research, quite a few assessment tools in this area exist.

Large scale efforts. Currently, elementary school algebraic thinking skills are measured primarily (if not solely) by high-stakes standardized tests, which measure all strands of mathematics ability including algebra. The Washington State Measurements of Student Progress (MSP), for example, measure the mathematics knowledge of students in grades 3-12 in Washington State and include a subscale measuring Number Sense and Algebraic Sense (for more information see OSPI's website at http://www.k12.wa.us/assessment/StateTesting/ScoreReportMSP.aspx). Teachers, parents, and students receive both overall scores and subscale scores (i.e., a score for Number Sense and Algebraic Sense); but further detail regarding specific algebraic thinking skills tested or the items used are not provided. High schools also generally have end-of-course exams that test algebra knowledge. The results of these tests are generally non-diagnostic in nature and the results arrive long after the tests occur (i.e., such test results can rarely be used by teachers to modify instruction). Due to the cost and effort involved in their development, these high-stakes tests generally have been validated and excel psychometrically. Despite this, these high-stakes tests

generally fail in providing teachers with information useful in making instructional decisions (Nichols, 1994).

Several other large-scale algebraic thinking assessments exist or are in development. The majority of these assessments only assess a portion of the algebraic thinking dimensions and also target middle and high school students exclusively. First, an online 21-question Algebra Readiness Check-up created by Sylvan and Let's Go Learn, (see http://letsgolearn.com/fit4algebra/) can be taken to determine if a student is 'ready' for traditional algebra. The assessment was created using the National Math Panel (2008) recommendations as critical foundations for algebra, but no other technical or validation information regarding the assessment is provided.

Second, other researchers have created an online Diagnostic Algebra Assessment System (DAAS), which is "a classroom assessment and instructional tool that comprises a battery of online diagnostic algebra tests that provides estimates of students' abilities and misconceptions" (Russell, O'Dwyer, & Miranda, 2009, p. 415) (see also http://www.bc.edu/research/intasc/researchprojects/DiagnosticAlgebra/daa.shtml). The skills focused on by this assessment include concept of variable, equality, and graphing. The assessment includes 34 multiple-choice items (10-12 items each in the three areas) that include the correct answer, an alternative conception, and two distractors (i.e., common arithmetic or problem-interpretation mistakes). In this way specific student conceptions can be diagnosed and this information relayed to teachers. This type of assessment poses a severe limitation, however, in that trait underrepresentation likely occurs due to utilizing only one measurement method (Kane, 2006). By utilizing multiple-choice questions only, for example, one risks drawing student attention to the answer choices instead of the problem itself (Kazemi, 2002). This large-

scale test progressed through full assessment development including pilot studies, validity studies, and a three-year trial that investigated the full and partial effects of the system on 905 students enrolled in an algebra class in 6^{th}-12^{th} grades. The developers (Russell, O'Dwyer, & Miranda, 2009) found that using the full DAAS system (including the diagnostic tests, ability reports, diagnostic misconception reports, and interventions) resulted in statistically significant improvements in algebraic skills.

Third, McGraw-Hill has created a "comprehensive online assessment solution aligned to NCTM standards in Algebra 1" entitled Acuity Algebra (see http://ctb.com/ctb.com/control/ctbProductViewAction?p=products&productId=709). The set includes a readiness exam, 13 formative assessments, and a proficiency exam, and is designed for students in grades 6-12. No other technical or validation information is provided.

Fourth, Kuchemann (1981) created an algebra test for 13, 14, and 15 year old students. He categorized his items by level (Levels 0-4) and intended to capture the different meanings of variable he described in his previous work (see work with variables section). The results and conceptions experienced by the student participants are described in Booth's (1984) publication. He provided convergent validity evidence by correlating his test with the Concepts in Secondary Mathematics and Science (CSMS) mathematics tests (Hart, 1980) and the assessment was also validated by conducting an intervention study and obtaining an increase in scores on the assessment after the intervention (Kane, 2006).

Fifth, Foegen, Olson, and Impecoven-Lind (2008) have worked towards the development of progress monitoring measures in the area of secondary algebra. Project AAIMS (Algebra Assessment and Instruction: Meeting Standards) seeks to develop and validate algebra assessment tools which test both basic skills and what is termed 'algebra foundations'. Their

validation efforts include reliability analyses, criterion validity correlations (i.e., including teacher ratings, grades, and standardized test scores), parallel-form reliability, growth in the construct over time (i.e., growth studies), and teacher use studies. Unfortunately, these teacher use studies revealed that there was no difference in student achievement over time between teachers who did not have access to the assessment data versus teachers who not only had access to the assessment data but also discussed and shared the data with others (Foegen & Olson, 2007).

Sixth and finally, one study involving the development of an assessment tool for elementary students exists, but for only one of the algebraic thinking dimensions (Rittle-Johnson et al., 2011). Using a construct-modeling approach to create a construct map, these researchers implemented their 37-item assessment measuring equivalence knowledge with 175 students in grades 2-6. They used a variety of different item types including true or false sentences, multiple-choice, answer only, short answer, etc. The researchers also obtained standardized student assessment data from the Iowa Test of Basic Skills (ITBS). They pilot-tested their data, four experts determined face validity (i.e., extrapolation), and factor analysis techniques determined internal structure validity (Kane, 2006). Item statistics were reported via Item Response Theory. Internal consistency, test-retest, inter-rater, and alternate forms reliability coefficients were reported. Like the DAAS system (Russell, O'Dwyer, & Miranda, 2009), much research has been conducted to ensure this assessment is valid and reliable; but its focus on only one trait (i.e., equivalence) limits its usefulness diagnostically.

Classroom assessments. A plethora of research has been conducted across all grade levels in which researchers have created assessments to measure student ability in algebraic thinking skills, discover student conceptions, or determine the success of an intervention

program or curriculum. All of these tools appear to be created for a specific purpose and appear to be generally without validation support. Jacobs and colleagues (2007), for example, created an algebraic thinking assessment tool to be used to measure gains from a professional development program. Like the potential diagnostic tool needed, these researchers created differing assessment tools for the various grade levels and also assessed multiple strands of algebraic thinking (they termed it relational thinking): equivalence, simplifying, work with variables, and generalization. Many of the items proposed for the Diagnostic Assessment for Algebraic Thinking Skills (DAATS) described below are similar to the items on these classroom assessments. Differences lie in that the proposed assessment items on the DAATS vary more by grade (i.e., five versions for five grade levels instead of three versions for five grade levels), incorporate more strands of algebraic thinking (i.e., open number sentences and functions are not included on Jacobs and colleagues' tool), and are diagnostic in nature (versus as a tool to measure intervention growth). Their sample size is also relatively small per grade level. No reliability or validity information is reported on the assessment tool developed by Jacobs et al. (2007).

Asquith, Stephens, Knuth, and Alibali (2007) also created an assessment for middle school students regarding work with variables and equivalence; the results were used in a comparison study with teacher predictions of how students would perform. These items were nearly identical to those used to measure teacher pedagogical knowledge in Stephens' (2006) study investigating pre-service teachers' pedagogical content knowledge surrounding algebraic thinking. Alibali and colleagues (2007) also used a short, three-item assessment to measure equivalence knowledge across the middle school grades. Rittle-Johnson and Alibali (1999) developed an assessment to measure both procedural and conceptual equivalence knowledge.

The equivalence items appear to be used repeatedly by Hattikudur and Alibali (2010) as a pretest and posttest for before and after conducting an equality intervention with 3rd and 4th grade students. The items were also used by McNeil and Alibali (2005a) as a measure of their study for students ages 7-11, and again in a follow-up study with elementary school students, middle school students, and college students (McNeil, & Alibali, 2005b), and with students in grades 1-4 by McNeil (2007). None of these studies appeared to report evidence supporting validation claims.

Finally, Fuchs and colleagues (2008) developed a 3rd grade algebra Dynamic Assessment (DA), which is designed to 'index' student skills through assisted learning (i.e., while two students might receive the same score on a diagnostic test, one student might be able to do more than the other "with help"). Items measured skills that included open number sentences with variables (i.e., x+5=11), multiplication equations with variables (i.e., 3x=9), and more than one missing variable in equations (i.e., x+2=y-1; y=9). Convergent validity evidence through correlations with state assessment scores was provided (i.e., correlations with the DA). While work towards validation has been conducted, this type of assessment differs largely from the proposed diagnostic assessment in the assessment methods and requirements on the teacher.

A table displaying summary information regarding these different types of assessments has been included in Appendix F.

Measurement Issues

Despite the creation of a seemingly large quantity of assessments in algebraic thinking skills, the majority do not appear to have been validated through a structured process. High-stakes tests, although often validated, do not provide the information needed to diagnostically change teacher instruction and improve student learning. The large-scale assessments, for the

most part, have striven to at least incorporate some validation claims and have usually been developed with care through a thorough process. These large-scale assessments are primarily targeted at middle and high school students, as only one larger-scale study was discovered for students in the elementary grades (Rittle-Johnson et al., 2011). This study was also produced in a thoughtful, structured manner; however it was designed for only one type of algebraic thinking skill: equivalence.

Six large-scale efforts with some aims at validation were discovered in the literature review, in addition to many more classroom assessments developed without such validation aims. A small percentage of these classroom assessments may have used some method of validation, but all appear to be lacking in some major way per Kane's (2006) validity framework. Kane's (2006) framework described validation to involve both an interpretive argument (i.e., designate all of the proposed test uses, interpretations, decisions, etc. a priori) and a validity argument (i.e., assess the interpretive argument to ensure it is reasonable). Interpretive arguments often involve developing 'if-then' statement rules that make inferences about the assessment results. An example of an interpretive argument for this specific algebraic thinking assessment would be: 'if the assessment results are stable over time, then the test-retest reliability coefficient should be at least 0.80.' The validity argument, on the other hand, involves evaluating this 'if-then' statement to ensure it is plausible, coherent, and reasonable (Kane, 2006).

Large-scale efforts, including standardized test items, often include at least one interpretive argument. Classroom assessments, on the other hand, rarely included either interpretive arguments or the associated validity arguments. Further, the majority of the classroom assessments do not appear to go through a purposeful development stage, in which the

test interpretations and uses are explicitly stated, a plan to achieving these interpretations and uses is developed, and the developers identify and also strive to control irrelevant variance (Kane, 2006). It therefore appears that the majority of these test developers fail to enter the 'appraisal stage', in which one strives to validate or evaluate the goals and uses of the assessment (Kane, 2006). Although it is possible these processes are done through 'behind the scenes' and not publicized research, it appears that the majority of the classroom assessment tools are not vetted through this process.

Perhaps the most popular validity argument made, often inexplicitly, is that of scoring: that the scoring rules used for the assessment are not only appropriate but are also implemented routinely and correctly. Both large-scale and smaller-scale assessments do this well: scoring criteria are appropriately developed, reviewed, and used. If rubrics are used the assessments commonly report inter-rater reliability, fulfilling the validity requirements of generalization from one rater to a universe of raters (Kane, 2006). This is only one small subset of generalization, and the majority of such assessments use a small numbers of items (i.e., they do not include a large enough sample from the universe of items) and rarely appear to conduct generalizability studies. Internal structure reliability studies in the form of calculating Cronbach's alpha are sometimes conducted.

The extrapolation inference is rarely referenced, in which researchers would strive to move the inferences from the assessment to more generalized settings (Kane, 2006). Often these validity judgments arise from expert opinion alone, but inter-rater agreement is also sometimes used, and this is commonly reported in the studies reviewed. The studies also often make claims regarding 'face validity', as this is an easily employed facet of validity relying heavily on expert judgment but also highly subjective and limited in usefulness (Kane, 2006). Reliability or

generalizability studies are also appropriate, and although some studies report reliability information few use generalizability analysis to generalize to a greater universe. Further, it appears that few of the studies conduct an analysis of consequences related to the test results or uses (Kane, 2006). Finally, the larger-scale efforts may conduct some think-aloud protocols to examine student cognitive processes, but very few of the classroom assessment developments went so far in-depth. In the defense of these researchers, even experts agree that it is difficult to balance extrapolation and generalization, because "we can strengthen extrapolation at the expense of generalization by making the assessment tasks as representative of the target domain as possible, or we can strengthen generalization at the expense of extrapolation by employing larger numbers of highly standardized tasks" (Kane, 2006, p. 37).

Last but not least in Kane's (2006) framework is trait underrepresentation, which is also a major problem, especially for largely underdeveloped assessments that were developed quickly and without a test map. Many of these assessments include a small number of items that feature only one method of measurement (i.e., multiple-choice). To ensure extrapolation and to avoid underrepresentation, it is desirable that a wide sample of the target domain is collected and that multiple measurement methods are used to reduce construct irrelevant variance (Kane, 2006). These issues are particularly relevant for diagnostic assessment because, by not adequately representing the trait, a teacher may miss diagnosing and addressing an important alternative conception. While others have successfully created assessments for a particular skill or algebraic thinking dimension (i.e., equivalence), it is important to more fully represent the entire continuum of skills on a diagnostic assessment.

Research Gap

It is clear that there is a need for an algebraic thinking assessment tool that solves the measurement issues described above. This research need does not end here. First, the majority of developed algebraic thinking assessments appear to measure only one dimension of algebraic thinking (i.e., only equivalence), while it does not appear that a comprehensive assessment designed to measure the various dimensions has been created. This method of assessing only one skill or dimension may be beneficial in fully representing that skill (i.e., reducing trait underrepresentation and increasing generalization) and in increasing reliability (i.e., by including a large quantity of items for one dimension). Despite these benefits, assessing only one skill also severely limits the assessment's diagnostic capacities, usefulness, and social validity from a teacher's point of view (Kane, 2006).

Second, the majority of the tools created do not appear to be diagnostic in nature (i.e., designed with the intent of informing instruction and improving student learning). One truly comprehensive diagnostic tool that has been validated targeted middle and high school students (Russell, O'Dwyer, & Miranda, 2009) and a second (not necessarily diagnostic) validated tool targeted elementary school students but focused only on equivalence (Rittle-Johnson et al., 2011). Most of these assessments were designed with a specific research purpose mostly around intervention and program evaluation in mind (i.e., to measure equivalence knowledge following an equivalence intervention to determine if the intervention is successful or not), so it is important to fully develop and validate an instrument with the diagnostic purpose in mind (Kane, 2006). This development will be discussed further below.

It is also important to consider the need for a validated instrument to be utilized in such intervention and other research. Hill and Shih (2009), for example, discovered that a very small

proportion of mathematical research studies reported validity and reliability information on the measures used: only approximately 17% of the studies reported validity information and only 38% of the studies reported reliability information. These results mirror the findings described above and demonstrate the intense need for validated tools for use in research.

Third, of the studies that focused on elementary school students, the majority of these studies have focused on one subset of age (e.g., 7-9 year olds only or 3rd grade students only, for example) instead of focusing on how algebraic thinking skills develop across the elementary school years. This focus on a small subset of age limits the information we know about how students' algebraic thinking skills develop across the grade levels. Studies that examine algebraic thinking skills across the spectrum of elementary grades (i.e., K-5) are needed to better understand this progression of knowledge.

Fourth, the majority of the research samples were primarily small (i.e., less than 100 students) convenience samples. No randomized samples were discovered. Despite the plethora of research in the area of equivalence, for example, it appears that very few of the research studies conducted featured random samples, large sample sizes, or cross-country international comparison samples (Capraro, Ding, Matteson, & Capraro, 2007). Convenience samples can provide biased results that are not representative of the general population, and small sample sizes are not appropriate for statistical techniques like Item Response Theory (IRT). It is therefore important to utilize a random sampling technique and include sample sizes larger than 100 at each grade level.

Because of these research gaps, and the fact that the development of a measure of the algebraic thinking construct for elementary school students could have beneficial outcomes for

both students and teachers, it appears that there may be a need for a comprehensive diagnostic algebraic thinking skills assessment designed for students across all of the elementary grades.

Chapter 3: Methods

This primarily quantitative measurement-based research study will strive to: a) validate a diagnostic assessment of algebraic thinking skills of elementary school students, b) investigate the algebraic thinking skills elementary students possess, and c) cross-culturally validate the diagnostic assessment tool with students in Singapore.

Participants

A large randomized sample of students and their associated teachers must be obtained to allow for generalizability of results. Student participants completed a diagnostic assessment and teacher participants completed a teacher survey.

Sampling Procedures

A stratified cluster random sample was used in this study, with schools stratified by mathematics achievement (i.e., high achievement, medium, and low) and students clustered by school. A randomized sample was chosen because nonrandom samples may display biased results (Allen & Yen, 1979). Cluster sampling was chosen to significantly reduce costs despite the severe drawbacks in reduction of precision associated with cluster sampling. To increase precision it is desirable to increase the number of clusters rather than the number of participants within each cluster (Kalton, 1983). Because all students within a school are included (i.e., a large number of participants within the cluster), as many schools (i.e., clusters) as affordably possible should be included in this sample. A stratified sample was chosen to ensure representativeness of the local population; but coverage error, sampling error, and nonresponse will remain major concerns (Dillman, Smyth, & Christian, 2009). In addition, test booklets were spiraled within classrooms. Four test forms were developed for each grade level and were

distributed in sequence within a classroom. This helped to ensure that between test bias is not an issue.

Although a sample of all schools in Washington State would be ideal, all schools in King and Pierce Counties were chosen to further reduce travel costs. Only public schools serving students in grades 1-5 were included in this sample (i.e., "primary schools", or those serving grades K-2, for example, were omitted). The final frame of elementary schools in King and Pierce counties included 397 schools. These schools were further sorted into three lists depending on mathematics achievement utilizing the state standardized test scores: high achievement (top third), medium achievement (middle third), and low achievement (low third). It was determined that stratification could occur based on either mathematics achievement or socioeconomic status because of the high correlation between the two variables (r = -0.71). Schools on each of these three lists were sorted again into large schools (top 50% in size) and small schools (low 50% in size). This resulted in six lists of schools: 1) high achievement large size, 2) high achievement small size, 3) medium achievement large size, 4) medium achievement small size, 5) low achievement large size, and 6) low achievement small size. Assuming a participation rate of approximately 33% of schools contacted, a total of 18 schools were randomly selected to participate; three schools were randomly drawn from each of the six lists using Research Randomizer (see http://www.randomizer.org/form.htm for more information). The school principals of these 18 schools were contacted regarding potential participation in the study. Participation was encouraged by offering each participating school a $300 stipend. Principals were asked to facilitate asking teachers if they would like to participate, distributing packets (described in procedures below), collecting data, etc.

Demographics of Schools, Teachers, and Students

The demographics of the six schools, the 81 teachers, and the 1,745 students in the Washington State sample are described below.

School Demographics. Six elementary schools from six different school districts participated in this research study. Three schools (50%) served students in grades K-5 while three schools (50%) served students in grades K-6. Students in kindergarten and 6th grade were excluded from this study. School specific demographics are displayed in Table 2 below.

The six participating schools were chosen randomly and stratified by 3rd grade school mathematics achievement (see methods section). Although mathematics achievement across the three grade levels remained fairly constant, 3rd Grade math achievement scores tended to be the highest of the three. As you can see in Table 3 below, percent passing the 3rd grade standardized test ranged from 37% to 83% with a mean of 65% (*SD* = 17%). Percent free or reduced price meals, which is often used to measure socioeconomic status, also varied widely, ranging from 30% to 77% with a mean of 45% (*SD* = 18%). Although in this case it is clear the correlation is not a perfect relationship, the overall correlation between percent free or reduced price meals and achievement scores was large enough to allow the two variables to be considered interchangeably (r = -0.71). 2010-2011 data was used for sampling techniques as this was the data available at the time the study took place. However, at the time of this writing, 2011-2012 data was also available, therefore it has been included as that is the school year the study actually took place. These data are available in Table 3.

Table 2.

School Specific Demographics

	Strata: Math Achieve-ment	Number of Classroom Teachers	Years of Classroom Teaching Experience	Free or Red. Price Meals (in Percent)	White (in Percent)	Special Education (in Percent)	English Language Learners (in Percent)	Student Enroll-ment	Student Participants
School 3	Low	34	6.4	77	30	20	36	540	289
School 6	Low	21	13.8	37	86	11	0	340	243
School 5	Medium	33	8.9	36	72	16	2	490	275
School 1	Medium	28	13.7	30	79	14	0	482	260
School 4	High	21	7.4	58	44	14	21	317	269
School 2	High	32	13.8	34	43	6	15	557	409

Table 3.

School Specific Mathematics Standardized Test Achievement Test Data

	Strata: Math Achieve-ment	2010-2011 3rd Grade Math Standardized Test Passing (in Percent)	2010-2011 4th Grade Math Standardized Test Passing (in Percent)	2010-2011 5th Grade Math Standardized Test Passing (in Percent)	2011-2012 3rd Grade Math Standardized Test Passing (in Percent)	2011-2012 4th Grade Math Standardized Test Passing (in Percent	2011-2012 5th Grade Math Standardized Test Passing (in Percent)
School 3	Low	37	41	37	77	30	20
School 6	Low	56	63	39	61	77	53
School 5	Medium	68	68	79	63	65	80
School 1	Medium	70	51	68	77	70	58
School 4	High	76	52	46	55	59	89
School 2	High	83	75	75	84	78	77

Teacher Demographics. Eighty-one teachers participated in this research: 12 (15%) from School 1, 17 (21%) from School 2, 16 (20%) from School 3, 12 (15%) from School 4, 12 (15%) from School 5, and 12 (15%) from School 6. In terms of grade level taught, 17 teachers (21%) taught 1st grade, 13 (16%) taught 2nd grade, 15 (19%) taught 3rd grade, 15 (19%) taught 4th grade, 16 (20%) taught 5th grade, and 5 (6%) taught multi-grade classrooms (i.e., 1 taught a 1st/2nd split, 2 taught a 2nd/3rd split, and 2 taught a 3rd/4th split). For the purposes of the remaining analysis of survey items by grade, multi-grade classrooms were categorized in the lowest grade (i.e., 2nd/3rd splits were categorized as a 2nd grade classroom). The average class size was 22.83 students (*SD* = 4.46).

Teacher participants were asked to complete a short teacher survey. Seventy-two teachers out of 81 completed this survey, for a response rate of 89%. The majority of teachers had obtained at least their Master's Degree as only 26% of the teachers had obtained only their Bachelor's Degree or their Bachelor's Degree plus a Professional Certificate. Several teachers went beyond their Master's Degree, with 13 (18%) also obtaining their Professional Certificate, 1 (1%) obtaining their Master's Degree, Professional Certificate, and their National Board's, 1 (1%) obtaining their Master's Degree, Professional Certificate, and their Administrative Certificate, and 2 (3%) obtaining their Master's Degree and an Ed.D. The number of years of teaching experience ranged widely from 1 to 39, with a mean of 15.25 years (*SD* = 9.22). Teachers were then asked to complete a series of Likert-style survey questions regarding their teaching of and student knowledge of math and algebraic thinking skills. Because these are Likert questions, and therefore ordinal level data, it is not appropriate to report the mean and instead only the frequencies of each answer are reported below in Table 4 (Clason & Dormody, 1994). It appears that the majority of teachers sometimes teach algebraic thinking skills, believe

the math skills of their students to be “good”, and believe the algebraic thinking skills of their students to be “fair”.

Additionally, teachers were asked to describe their general preparation regarding the teaching of mathematics. Teachers in general felt well prepared both to teach math and to teach algebraic thinking skills, however they felt more strongly towards general mathematics. While 94% of teachers agreed or strongly agreed that they felt well prepared to teach math, only 79% of teachers agreed or strongly agreed that they felt well prepared to teach algebraic thinking skills. On the other hand, teachers tended to feel neutral regarding their students ability to grasp algebraic thinking skills easily: only 25% agreed or strongly agreed that their students grasped algebraic thinking skills easily while 34% disagreed or strongly disagreed with this statement.

To better understand the types of professional development teachers have recently received in math, teachers were asked if they had received professional development in teaching math or in teaching algebraic thinking skills at some point within the past five years. More than half (51%) reported receiving professional development in teaching math within the last year and 88% reported receiving professional development in teaching math within the last 5 years. On the other hand, less than a quarter (20%) reported receiving professional development in teaching algebraic thinking skills in the last year and only two-thirds total (66%) reported receiving professional development in algebraic thinking skills in the last 5 years.

These teacher survey results are displayed below in Table 4. For ease of reading, items receiving greater than 20% of teacher response have been highlighted.

Table 4.

Teacher Survey Results

	Never	Rarely	Sometimes	Often	Daily
How often do you teach math lessons that include algebraic thinking?	3 (4%)	**15 (21%)**	**29 (41%)**	**20 (28%)**	4 (6%)
	Very Poor	Poor	Fair	Good	Very Good
Rate the math skills of your current students.	0 (0%)	4 (6%)	**22 (31%)**	**40 (56%)**	6 (8%)
Rate the algebraic thinking skills of your current students.	2 (3%)	**15 (21%)**	**40 (56%)**	14 (19%)	1 (1%)
	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
I feel well prepared to teach math.	2 (3%)	1 (1%)	2 (3%)	**40 (56%)**	**27 (38%)**
I feel well prepared to teach algebraic thinking skills.	1 (1%)	6 (8%)	8 (11%)	**38 (53%)**	**19 (26%)**
My students grasp algebraic thinking skills easily.	4 (6%)	**20 (28%)**	**30 (42%)**	**16 (22%)**	2 (3%)
	Less than 3 months ago	3-12 months ago	1-5 years ago	More than 5 years ago	Never
I have received professional development in teaching math.	7 (10%)	**29 (41%)**	**26 (37%)**	6 (9%)	3 (4%)
I have received professional development in teaching students algebraic thinking skills.	1 (1%)	13 (19%)	**32 (46%)**	12 (17%)	12 (17%)

Student Demographics. A total of 1,745 students participated in this research: 351 1st grade students, 309 2nd grade students, 336 3rd grade students, 384 4th grade students, and 365 5th grade students. Students were asked to self-report their grade and if they were an English Language Learner (i.e., if they spoke a language other than English at home with their family). It is understood that there is likely some degree of error associated with young students self-reporting this information. It is likely, for example, that students may have interpreted that speaking English at home at any point (i.e., even if just a few words or only with friends or siblings) deemed a "no" answer. This demographic information is reported below in Table 5.

Table 5.

Student Demographics

	Total Students	Male	Age	English Language Learner
1st Grade	351 (20%)	175 (52%)	6.88 (*SD* = 0.44)	87 (26%)
2nd Grade	309 (18%)	149 (51%)	7.84 (*SD* = 0.43)	64 (22%)
3rd Grade	336 (19%)	167 (52%)	8.82 (*SD* = 0.45)	76 (24%)
4th Grade	384 (22%)	200 (52%)	9.85 (*SD* = 0.48)	79 (21%)
5th Grade	365 (21%)	176 (49%)	10.83 (*SD* = 0.48)	68 (19%)
All	1,745 (100%)	867 (51%)	8.96 (*SD* = 1.49)	374 (22%)

Procedures of Data Collection

Upon agreeing to have their school participate in the research study, principals were asked to recruit their teachers to participate. Principals shared with the researcher the number of teachers that agreed to participate in the study, the grade levels taught by each of these teachers, and the number of students in each of these classrooms. The researcher then prepared the box of materials and delivered it to the principal within one week. The box of materials included one packet for each participating teacher. The principal was asked to distribute the packets to the associated teacher labeled on each packet via a set of directions included with the materials: "*Thank you so much for agreeing to participate in this research! I greatly appreciate your help, and will provide you and your teachers with the results of this assessment that will hopefully inform future teaching and help your students. Please distribute each packet to the teacher whose name is on the front of the packet. Each packet contains information and directions for the teacher along with enough assessment copies for their entire class. I have included a checklist for you to monitor who has and has not returned the completed assessments. Once you have received all completed packets, please email me at nralston@uw.edu and I will stop by to pick up this box of materials. I will then score the assessments and provide you with detailed results within three weeks*."

Each packet was labeled with the teacher's name and included: 1) a letter describing the procedures of the study, 2) an informed consent form, 3) a paper/pencil teacher survey, 4) directions to administer the assessment, and 5) assessment copies for every student in their classroom. The four different versions of the assessment (i.e., A, B, AB, BA) were randomly distributed at the student level. It was made very clear to the teachers that participation was completely voluntary, and if they did not wish to participate they simply needed to do nothing with the materials provided. Teachers that agreed to participate were asked to sign the informed consent form, complete the teacher survey, and have their students complete the assessments.

Across all of the above described conditions, teachers read a set of standardized directions to their students (see administration directions in Appendix D). When students completed the assessments, teachers placed them back inside the envelope with their completed consent form and teacher survey. The teachers then returned the packet to the principal and the researcher retrieved the packets. Assessments were graded and teachers were all provided with a summary of their students' performance on the various tasks. All procedures were approved by the Institute Review Board (IRB) organization at the University of Washington in concordance with Human Subjects appropriate guidelines.

Instruments

Two different measures were utilized in this research: the teacher survey and the diagnostic assessment of algebraic thinking skills.

Teacher Survey Instrument. The teacher survey instrument contains 16 questions, 9 being content questions and 7 being demographic questions. Content questions attempted to gather information regarding teacher perceptions about their ability to teach algebraic thinking skills, the frequency of teaching algebraic thinking skills, and the professional development they

had received on teaching algebraic thinking skills. Demographic questions attempted to gather information regarding the teacher's years of experience, education levels, curriculum used, and grade level taught. The teacher survey instrument is available in Appendix E.

Teacher Survey Development. After initial development of the teacher survey tool, the survey went through several development iterations. First, five experts in the field critiqued the survey tool and revisions were made to the survey instrument in accordance with their suggestions. Second, one cognitive interview with a teacher representative of the population to be surveyed was conducted and further revisions were made in accordance with problems that arose in that cognitive interview. Third, one additional expert critiqued the survey following the new revisions, and further revisions were made. Fourth and finally, a small pilot study was conducted with 17 elementary school teachers in four elementary schools in one urban district in the Pacific Northwest. Teachers surveyed included three 1st grade teachers (18%), three 2nd grade teachers (18%), six 3rd grade teachers (35%), five 4th grade teachers (29%) and zero 5th grade teachers (0%). The majority (94%) of these teachers were female and possessed a Bachelor's Degree (18%) or a Bachelor's Degree and a Master's Degree (53%). The average number of years of teaching experience was 14.29 years (SD = 10.06). The pilot survey version was delivered as an internet web survey with the formatting kept the same when displayed as a paper/pencil survey. Survey questions were revised following the results of this pilot to reduce vagueness, induce more meaningful responses, and better discriminate answers amongst respondents (i.e., a question was removed regarding the number of minutes spent teaching math weekly, the wording of several questions was revised slightly, and the questions regarding professional development were added). Revisions were made based on this pre-testing to produce the final instrument draft attached in Appendix E.

Algebraic Thinking Diagnostic Assessment. Ten different versions of the algebraic thinking diagnostic assessment tool were developed. These 10 versions include five different levels of the assessments (i.e., one for each grade level for grades 1, 2, 3, 4, and 5) and two different versions of the assessment at each grade level (i.e., 1A, 1B, 2A, 2B, 3A, 3B, 4A, 4B, 5A, 5B). The two versions at each grade level were then combined in two different ways to create four forms at each grade level for 20 versions in all. Anchor items exist across the grade levels and across the dimensions. The items have been mapped so that anchor items can be used to link forms and to create a growth scale so that students' development of algebraic thinking can be measured over time. These anchor items can be viewed in Chapter 4, Tables 9a, 9b, and 9c.

Assessment questions were broken into three strands of algebra: 1) Modeling, which includes solving open number sentences, understanding equivalence, and work with variables; 2) Generalized Arithmetic, which includes efficient numerical calculations (i.e., simplify calculations using number relations and compensation strategies) and generalizing (i.e., utilizing mathematical properties like the commutative property, property of zero, etc.); and 3) Functions, which includes possessing the ability to recognize, describe, extend, and create patterns.

The assessment includes approximately 25-28 items, resulting in a total of approximately 27-31 points. Short-Answer items are scored with a 3-level scoring rubric (0-2). Each algebra dimension is worth approximately 3-17 points out of the total points. Scoring guides are available in Appendix D. Modeling will account for the majority of the points, with about 55% of the total score, with the remaining points divided up about equally between Generalized Arithmetic (22.5%) and Functions (22.5%). The number of items per dimension and per dimension strand on each assessment are displayed below in Table 6.

The assessment is designed to be administered in one sitting in approximately 30 minutes. The test is not a timed test so the teacher may choose to allow more than 30 minutes of time or discontinue testing at the conclusion of 30 minutes. Items are written at a reading level appropriate for an elementary school student audience; therefore the items will aim to be readable at a 2nd grade or lower readability level. Teachers may read the items to students who cannot read at a 2nd grade level.

Table 6.

Item Distribution by Dimension across Assessments

Dimension	Dimension Strand	Number of Questions	Number of Questions per Dimension	Number of Points	Number of Points per Dimension
1st & 2nd Grades					
Modeling	Solving Open Number Sentences	6	16	6	16-17
	Understanding Equivalence	6-7		7-8	
	Work with Variables	3		3	
Generalized Arithmetic	Efficient Numerical Calculations	1-2	2	1-2	3-4
	Generalizing	1		2	
Functions	Repeating Patterns	2	6	2	5-7
	Linear Patterns	2-3		2-3	
	Nonlinear Patterns	1-2		1-2	
3rd, 4th, & 5th Grades					
Modeling	Solving Open Number Sentences	5-6	14-15	5-6	15-16
	Understanding Equivalence	5		6	
	Work with Variables	4		4	
Generalized Arithmetic	Efficient Numerical Calculations	2-3	4-5	2-3	6-7
	Generalizing	2		4	
Functions	Repeating Patterns	-	6	-	6
	Linear Patterns	3		3	
	Nonlinear Patterns	3		3	

The final test specifications can be viewed in Appendix C, which displays the number of questions on each grade level's assessment, the number of points each grade level's assessment is worth, and the proportion of each grade level's assessment devoted to each algebraic thinking skill. The final item specifications can also be viewed in Appendix D, which displays the specifications for writing items measuring each algebraic thinking skill.

The detailed description of the development of the algebraic thinking assessment is described fully in Chapter 4.

Validation Procedures

A framework for validation (Kane, 2006) and the validity limitations associated with many algebra assessments currently available was previously discussed. It is now necessary to make clear the plan of action to validate this algebraic thinking assessment tool. The "need for validation derives from the scientific and social requirement that public claims and decisions be justified" (Kane, 2006, p. 17). While very similar in meaning and obviously related, validity, or validation, can have two different interpretations: 1) validation as "the development of evidence to support the proposed interpretation and uses" (i.e., building a case for, or justifying, a test's use); and 2) validation as "an evaluation of the extent to which the proposed interpretations and uses are plausible and appropriate" (i.e., examine the overall evidence for a test's use) (Kane, 2006, p. 17). This research is specifically focused on the second interpretation, in which, as the test developer, it is my responsibility to build a case for the algebraic thinking assessment's proposed diagnostic interpretations and uses. These attempts at validating such a tool are discussed fully below.

Item Analysis

In the development of any new assessment tool, it is important to conduct a thorough item analysis of all assessment items. This is done both to better understand the construct and dimensions and also to ensure there are not any problematic assessment items. This is often conducted by implementing an appropriate pilot test; however, this process should be on-going throughout the assessment's lifespan as well. The item analysis was conducted by analyzing the assessment data to generate item statistics using both Classical Test Theory and Item Response Theory.

Classical Test Theory. The data was first analyzed with Classical Test Theory (CTT; Allen & Yen, 1979) as a preliminary analysis to both better understand the construct and initially screen for potentially problematic items. CTT is an additive model based on the underlying assumption that a student's observed score (*X*) is equal to her/his true score (*T*) plus her/his error of measurement score (*E*) (Allen & Yen, 1979). Item difficulty levels (i.e., the proportion of students who answered that item correctly) were calculated. Item difficulty levels for two-point answers were calculated by taking the proportion of students who answered that item correctly and dividing by two. Although an item difficulty level is normally recommended to be within 0.3 to 0.7, because this is a diagnostic assessment, this is not a requirement for an adequate item. Item discrimination (i.e., how well a student performs on one item should predict how they score on the whole assessment) is of more use, and was examined via adjusted item-total correlations. Internal reliability coefficients (i.e., Cronbach's Alpha) were also calculated. These analyses were conducted by grade level to ensure that each set of items were appropriate for each grade level.

Item Response Theory. The use of CTT alone involves many limitations, including: student and test characteristics cannot be separated (i.e., the student's ability level is based on the difficulty of the test items, item characteristics are group-dependent, student scores are test-dependent), psychometric characteristics are test oriented instead of item oriented; the model assumes the standard error of measurement is homogenous for all students tested, etc. (Hambleton, Swaminathan, & Rogers, 1991). Because of these many limitations, this data was also analyzed using Item Response Theory (IRT; van der Linden & Hambleton, 1996), which solves many of the above-mentioned issues. Unidimensionality (i.e., that only one construct is measured by a set of items) and invariance (i.e., item parameters are not dependent on student ability and vice versa) are the major required assumptions of IRT (Hambleton et al., 1991). Although I have discussed that several dimensions may exist under the construct of algebraic thinking, IRT generally produces robust parameters despite minor violations of unidimensionality (Harrison, 1986). IRT allows for non-sample specific estimates of item parameters and therefore further strengthened arguments of the items that are used at each grade level. Item discrimination estimates were also reported via IRT. The partial credit polytomous model was used for the short answer (i.e., scored with a rubric) items.

Reliability Studies

It is often stated that, while an assessment can be reliable but not valid, it cannot be valid if not reliable. There are several aspects of reliability that, therefore, must be analyzed to ensure the interpretation of student scores on the algebraic thinking assessment tool is reliable, in that the scores the assessment provides are consistent. The following types of reliability claims were therefore examined during the validation of the algebraic thinking assessment.

Alternate Forms Reliability. Alternate forms of this diagnostic assessment are necessary for cases in which it is desired to use the assessment multiple times throughout one school year, yet to avoid any 'learning effects' that may incur from giving an identical assessment multiple times. A research study may use the alternate forms as a pretest and posttest before and after implementation of an intervention, for example, or a teacher may use the assessment to diagnose algebraic conceptions and again to determine if his or her instruction has improved learning. Each grade level has, therefore, at least two forms of the assessment. All of the assessments across the grade levels are similar except for the items progressing in difficulty moving upwards through the grade levels. Common items were used to vertically link the forms. Common persons (examinees) were used to link the alternate forms.

Collecting Alternate Forms Reliability data was therefore necessary to ensure that the scores from two comparable versions of an assessment may be interpreted similarly. Alternate Forms were developed through the pilot-testing of a large item bank and striving to develop two versions that are psychometrically equivalent. Equating the forms occurred by having students complete anchor items on both versions of the assessments. The list of anchor items may be seen in Chapter 4. Students were randomly assigned to one of four different versions of the assessment: version A, version B, the first half of version A combined with the second half of version B (AB), and the second half of version A combined with the first half of version B (BA). The four different versions were randomly assigned to students within the same classroom. The assumption that the two versions have similar observed score means and variances was tested. The items on alternate forms within a grade level were placed on the same scale using horizontal equating. Items across grade levels were then placed on the same scale using vertical equating. I chose not to use the term of Parallel-Forms here because the two versions for each grade level of

the algebraic thinking assessment created may vary slightly from one another in item difficulty and are not meant to be exactly parallel.

Inter-Rater Reliability. A potential diagnostic assessment should utilize at least a small portion of short answer items to be scored via a rubric to increase generalization through multiple measurement methods (Kane, 2006). Collecting Inter-Rater Reliability evidence was therefore necessary to determine that different raters, or scorers, interpreted and implemented the answer key, scoring criteria, and rubrics similarly. A random sample of 6% of the assessments was scored by an additional second rater trained in the scoring procedures using only the scoring rules and the rubrics. The additional rater scored the entire assessment, to ensure that similar scores are given to the dichotomous correct/incorrect items in addition to the short-answer items scored with rubrics. Percent exact agreement was calculated to analyze the inter-judge agreement between the two scorers for scoring these short-answer items. This was conducted because this assessment tool is designed to be used in the classroom by the teacher; therefore the scorers in reality will likely be trained only using the scoring criteria.

Validity Studies

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) declares validity to be the most important consideration when developing a test, and defines it as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Messick's (1989a) seminal work on validity goes on to define it as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness of inferences* and *actions* based on test scores and other modes of assessment" (p. 13, italics in original). The history of validity theory demonstrates many differing opinions and controversy regarding what counts as validity

and what is important in terms of validity, beginning with criterion-related validity, next moving to content validity, next emphasizing construct validity, and then finally moving towards a more unified approach to validity that incorporates all three as sources of evidence for validity (Kane, 2001; 2006). These three most heralded sources of evidence for validity have even been termed "something of a holy trinity representing three different roads to psychometric salvation" (Guion, 1980, p. 386). Despite many researchers separating these, some believe validity to be an integrated concept, not one that can be divided into different pieces (Messick, 1989a; AERA, APA, & NCME, 1999; Kane, 2006). Messick (1988), for example, was a huge proponent of this, claiming all validity to actually be construct validity: "all educational and psychological measurement should be construct-referenced because construct interpretation undergirds all score-based inferences – not just those related to interpretive meaningfulness but also the content- and criterion-related inferences specific to applied decisions and actions based on test scores" (p. 35). Constructs must be validated when the ability the assessment is measuring cannot be operationally defined (i.e., it is a latent variable), such as in the case of algebraic thinking skills (Cronbach & Meehl, 1955).

Messick (1989b) strove to impart that, it is the use of and interpretations of the test scores that become validated, never the test itself. Others have agreed with this theory in that it is the test score interpretations and test score uses that must be determined to be valid (Kane, 2006). Nichols and colleagues (2009) clarified this in referring to formative assessments specifically: "General validity theory appears to privilege test score interpretation over test score use… [but] validity claims with regard to formative assessment emphasize test score use over test score interpretation" (p. 15). Using Messick and Nichols' information as a framework, this research is, therefore, particularly focused on discovering that the results of this algebraic thinking

assessment tool display what they intend to (i.e., algebraic thinking skills). To ensure the interpretation of the algebraic thinking assessment scores are valid, several validity studies encompassing a variety of methods of validation were conducted. In accordance with this unified view of validity accompanied by the assessment standards put forth by the *Standards for Educational and Psychological Testing*, the sources of validity evidence collected in this research study are separated into five categories: 1) Test Content, 2) Internal Structure, 3) Response Processes, 4) Consequences of Testing, and 5) Construct-Related Evidence for Validity (AERA, APA, & NCME, 1999). Although this work focused on the following types of specific validity claims, there are many that could have potentially been tested as a starting point to validate this instrument. Future research should seek to collect other forms of evidence to validate this instrument. This work focused on collecting the following empirical evidence which will be reported in later chapters: face validity and sampling validity (test content evidence); internal reliability analysis and an examination of the internal structure (internal structure evidence), teacher surveys (consequences of testing), analysis of student results and cross-cultural comparisons (construct related evidence), and qualitative coding of student responses and "think-aloud" interview protocols (response processes evidence).

Test Content Validity Evidence. Sources of evidence for content validity include face validity and sampling validity (Allen & Yen, 1979). Face validity evidence is obtained by determining if the test appears to measure what it is supposed to measure (i.e., does this assessment appear to measure algebraic thinking?). Face validity evidence is typically collected when an expert inspects the assessment and concludes that the assessment does, indeed, measure the trait in question. In this case, this face validity evidence was collected by conducting a thorough literature review and a careful development of the assessment items. Further, the

assessment was informally examined by several measurement experts, elementary school teachers, and Dr. Linda Levi, an expert on algebraic thinking skills, prior to utilizing it in this research study.

Sampling validity evidence is obtained when it is judged that a test has included a full representative sample of the construct being tested (i.e., is the construct of algebraic thinking being fully measured?). Sampling validity evidence is typically collected when it is determined that the assessment adequately covers all dimensions of the construct (Allen & Yen, 1979). In this case this sampling validity was collected by conducting a careful examination of the assessment's representativeness of the various dimensions and developing a "test blueprint" (Lissitz & Samuelsen, 2007, p. 441). This blueprint, or test map, can be viewed Chapter 4.

Both face and sampling validity evidence are extremely important in reducing occurrences of construct underrepresentation (i.e., ensuring the assessment measures the key components of the construct) (AERA, APA, & NCME, 1999; Kane, 2006). The majority of content validity evidence can often be called subjective or even to have a "confirmatory bias", in that because the test developer often evaluates such judgments, it is common to confirm validity evidence in this area (Kane, 2006). This is another case for triangulating validity evidence through many sources.

Internal Structure Validity Evidence. According to the Standards, "analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA, APA, & NCME, 1999, p. 13). Collecting internal structure validity evidence ensures that the items act similarly to each other (i.e., that the assessment exhibits unidimensionality, or truly measures one overarching construct and not a variety of different

constructs), and generally increases as the intercorrelations amongst the items increases (Allen & Yen, 1979). Although I have suggested in my conceptual framework that three factors exist, it is likely that these factors are all highly correlated with one another, that internal consistency should remain fairly high, and that unidimensionality does indeed exist. The assessment data collected was analyzed to discover the Cronbach's alpha internal reliability coefficient, which is an estimate of the internal consistency of the instrument.

Consequences of Testing Validity Evidence. It seems obvious that the benefits of using a new assessment tool should outweigh the consequences of using that same new assessment tool (Kane, 2001). For a new assessment tool, such as the one described in this research study, collecting validity evidence on the consequences of test use can be an extremely demanding and difficult task, due to the fact that because the test is new, and no consequences currently exist, the effects of the test are therefore currently unknown, and random assignment is required for causality to be examined (Nichols & Williams, 2009). There is some controversy over whether the test user or the test developer is responsible for collecting this evidence; however, these categories often overlap, particularly when the test user is also a researcher, and both certainly need to at least be concerned about such consequences (Kane, 2006). Although test users have their responsibilities as well, test developers are expected to continue to monitor the use of the test to ensure it is used in the anticipated way, and "test developers are obligated to attempt to maximize positive consequences and minimize negative consequences" (Nichols & Williams, 2009, p. 6). In this case, it is important to take responsibility as the test developer and review potential consequences of test use and how these consequences can be minimized.

Nichols and Williams (2009) proposed a framework of conditions the test developer should consider, including: 1) breadth of responsibility (i.e., in this case the breadth is narrow

because the construct, algebraic thinking skills, is also fairly narrow), 2) distance from intended score use (i.e., the use of the test score is to make instructional decisions regarding teaching algebraic thinking skills, therefore the distance is very near and therefore the potential consequences are lower), and 3) time from test publication (i.e., the responsibility is high at time of test publication, which is now, and will grow if the test use becomes popular).

To gather initial data regarding the consequences of the use of this diagnostic assessment, all participating teachers were surveyed following assessment testing. The results of the assessment for the class in aggregate were given to the classroom teacher. The teachers then received an online web-based survey that asked a series of rating-scale and open-ended questions: 1) Do these results surprise you? Why or why not?; 2) What do you think these results mean?; and 3) What, if anything, have you done (or will you do) with this information? Teachers were also asked to rate the usefulness of the assessment information. This data will determine how the teachers are using the assessment scores, and that teachers are not using the assessment scores inappropriately (i.e., to give students' grades, to enter students' into gifted and talented programs, etc.). This data may potentially demonstrate whether use of the diagnostic tool has impacted instruction and student learning.

Construct-Related Evidence for Validity. According to the *Standards for Educational and Psychological Testing*, construct-related validity evidence includes identifying potential sources of construct irrelevant variance through the examination of how items function differently among different groups (AERA, APA, & NCME, 1999). For example, student performance should improve as students progress throughout the grade levels and progress in age (i.e., different age groups of students should perform differently from each other). Over- or under-estimations of students' true abilities may be caused by several construct-related evidence

issues, including: 1) issues related to the test or items themselves, 2) personal characteristics that may manifest themselves more or less in different groups of students, or 3) an interaction between the two (Scheuneman & Slaughter, 1991).

Cross-cultural validity study in Singapore. A second method for examining construct related validity evidence will be to examine the score patterns of students in another country in comparison to student responses in the U.S. (i.e., the comparison of different groups as group differences should likely be present). Students who have been explicitly taught algebraic thinking skills (i.e., such as in Singapore through the 'Model Method') should outperform students who have not been explicitly taught algebraic thinking skills (i.e., students in the U.S.). Further, students who outperform others in mathematical achievement tests (i.e., students in Singapore) should likely outperform students on the algebraic thinking items. These comparisons will be made to add evidence for construct-related validity. As mentioned previously, the U.S. performs in the average range on international tests of mathematics (Mullis, Martin, & Foy, 2008), and therefore the U.S. may desire to look to the many higher-performing countries for advice. Many of these higher performing countries are Asian countries, who's success has been attributed to a number of factors, including the Chinese number system, the role of culture, emphasis on mathematical learning starting in the early years, teacher beliefs, etc. (Ng & Rao, 2010). Singapore specifically consistently performs at the top tier on international tests of mathematics achievement: in 2007 Singapore significantly outperformed the U.S. in 4th grade mathematics on the TIMSS assessment (e.g., the U.S. scored 529 on 4th grade math, while Singapore scored 599 on 4th grade math) (Mullis, Martin, & Foy, 2008). This significant gap in scores appears to persist throughout the grade levels (e.g., the U.S. scored 508 on 8th grade math, while Singapore scored 593 on 8th grade math).

Unlike the U.S., Singapore operates with a centralized education system run by the Ministry of Education. Because of this, they also use a common math curriculum and common standards; but they do allow textbook choice at the school level from a list of approved textbooks (this was not the case prior to 2006). Singapore highly values mathematics education because, as a small city-state, the country does not have natural resources upon which it can rely for economic wealth (Yee & Hoe, 2009). Singapore has been proclaimed as having an extremely heavy emphasis on testing, sometimes even portrayed as an obsession, and parents highly value education (i.e., if possible students receive private tutoring beginning as early as kindergarten) (Birenbaum et al., 2005). Unlike the spiral type curriculums that exist in the U.S. the curriculum in Singapore is less repetitive and more focused. Textbooks in Singapore, for example, are often 25% as long as the textbooks in the U.S. (which often are longer than 800 pages) (Yang, Reys, & Wu, 2010). Teachers in Singapore receive relatively higher salaries and status than in the U.S.; they often have more advanced math backgrounds, and receive more extensive professional development (Birenbaum et al., 2005).

In recent years Singapore's Ministry of Education has made several changes, including increasing the emphasis on mathematical application and modeling; emphasizing highly cognitively-demanding problem-solving tasks (Eric, 2009; Yee & Hoe, 2009). Problem-solving is one of Singapore's top priorities, and their curricula even cite and embody NCTM's (1989) recommendations: "problem solving should be the central focus of the mathematics curriculum" (p. 23). The 'Model Method' was first introduced in 1983, and it involves utilizing pictorial model drawing to solve algebraic thinking problems in the primary grades, which can help students focus on what a problem represents, instead of focusing just on applying an algorithm (Cai & Moyer, 2008). The 'Model Method' was introduced with the aim of allowing primary

students to access complex algebraic thinking word problems, and is included in all approved textbooks in Singapore (Ng & Lee, 2009). Ferrucci and colleagues (2008) raved about the potential benefits of using the 'Model Method', including its use in adjusting students to learn traditional algebra, encouraging the development of improved problem-solving methods, improved ability to create representations, and that it also "encourage[s] students to construct algebraic notation in a meaningful way through their generalizations in drawing diagrams from the models and then analyzing these diagrams as instances of mathematical structures" (p. 208). Singapore therefore makes a nearly-perfect cross-cultural validation site, because of its long-history of being a top-performer internationally on mathematics assessments, speaking English in academic settings, and appearing to emphasize the learning algebraic thinking skills early through integration of the 'Model Method'.

Because Singapore has significantly outperformed the U.S. on international mathematics assessments, gaining the knowledge Singapore possesses is particularly important and will ideally lead to the improvement of algebraic thinking skills of elementary school students. Singapore was therefore selected as a site to conduct cross-cultural validity samples on this assessment tool. Because assessments themselves can be considered 'cultural artifacts', it is an important validity concern to ensure that a student's culture is not causing misinterpretation of the test items (Basterra, Trumbull, & Solano-Flores, 2010). Such concerns are especially important in the cases of English Language Learners, bilingual students, and students completing assessments developed in a different country (as in this case), and these concerns are heightened when achievement gaps exist (Basterra et al., 2010). Given Singapore's reputation for academic excellence, their use of English as the academic language, and their teaching of algebraic skills early via the 'Model Method', this was near-perfect location for a cross-cultural validation site.

For this cross-cultural validation study, a sample of 1,619 students in grades 1-6 were selected from four 'neighborhood' schools in Singapore (i.e., average, not high-performing schools or schools with a selection process). Student participants included 248 Grade 1 students (15%), 224 Grade 2 students (14%), 436 Grade 3 students (27%) , 324 Grade 4 students (20%), 297 Grade 5 students (18%), and 90 Grade 6 students (6%). Because 6th grade is not generally part of the elementary school in the U.S. and the sample of Grade 6 students was very small, the Grade 6 results were omitted from the subsequent analyses. Students completed the assessments in summer of 2011. IRT will again be utilized to analyze the assessment results with the sample of students in Singapore.

Response Processes Validity Evidence. Response process validity evidence involves the analysis of the cognitive processes students go through when answering the assessment items. Response process validity evidence is provided when these mental processes reflect the mental processes the assessment seeks to employ. Data regarding response-processes was collected in two ways. First, a detailed analysis of the open-ended responses was conducted to analyze student thinking processes. Student responses were qualitatively coded and analyzed for common thought processes and conceptions. Second, a popular method for examining response processes called the think-aloud protocol was implemented; it is described in more detail below. The codes produced through the think-aloud analysis were then connected to the codes produced through the detailed analysis of the open-ended responses to demonstrate they were highly consistent. This data was analyzed in this way to demonstrate that these diagnostic assessment items have the potential to elicit rich alternative conceptions from students. As this is the purpose of this diagnostic assessment tool, it is important to demonstrate that the data produced from such a tool has the potential to provide teachers with rich information to inform instruction.

Think-aloud protocol interviews. A think-aloud protocol study was conducted to ensure the questions measure what they are meant to measure across all grade levels and both versions of the assessment. Seventy-three students of varying abilities from two different elementary schools in King County of Washington State were selected to participate in the think aloud protocol. Grade level participants included 18 1st grade students (25%), 12 2nd grade students (16%), 11 3rd grade students (15%), 10 4th grade students (14%), and 22 5th grade students (30%). The two alternate versions of each grade level assessment were randomly assigned to 50% of the students in each grade level. In the think-aloud protocol interview students were asked to verbalize their cognitive processes when answering each of the questions. The participating students were interviewed using a think-aloud protocol, which can provide an in-depth view of how students perceive the assessment items (Ericsson & Simon, 1984; Ginsburg, 1997). In this think-aloud protocol interview students were asked to verbalize their cognitive processes when answering each of the questions. By asking the students what they were thinking when answering the assessment items or why they answered in the way they did, "the experimenter seeks to learn directly from them [the students] the underlying cognitive structure that produced the overt behavior" (Ericsson & Simon, 1984, p. 42). The information was used to ensure students truly understood the content and that the cognitive processes required of answering the items are the cognitive processes intended. The think aloud data was analyzed using a specific process. First, all think aloud interviews were audio-taped and transcribed. Second, several readings of the interview transcriptions was conducted to gather first pass "codes", or common thought processes and conceptions. Third, qualitative codes or themes were finalized and all interview transcriptions were coded to discover a pattern of themes or codes. Fourth and finally, these themes were analyzed in conjunction with the results of the response

processes analysis discovered in the large quantitative sample to look for themes across a larger range of data. The results of this process are reported in Chapter 7.

Summary

In summary, the purpose of this dissertation study was to create and validate a diagnostic assessment tool of algebraic thinking skills for students in the elementary grades. This assessment tool intends to help teachers better recognize and understand algebraic thinking conceptions, with the aim to benefit students through improved instruction and increased opportunities to experience algebra. A range of reliability and validity evidence was collected on the interpretations of the scores of this assessment. A summary of all of the studies to be conducted is displayed below in Table 7. The final assessment tool will be published for wide-spread teacher use and the student results will be used to better understand the algebraic thinking conceptions held by students in Washington State.

Table 7.

Summary of Studies Conducted

Name of Study	Focus of Study	Type of Analysis	Number of Cases
Item Analysis	To better understand the construct and identify and eliminate any problematic assessment items	Classical Test Theory and Item Response Theory for item difficulty and discrimination analyses	1,745 U.S. students 1,619 Singapore students
Alternate Forms Reliability	To determine if scores from two comparable versions of an assessment may be interpreted similarly	Horizontal equating via common persons Vertical equating via common items	1,745 U.S. students
Inter-Rater Reliability	To determine if multiple raters use the answer key, scoring criteria, and rubric similarly	Percent exact agreement	109 U.S. students (6% of assessments)
Test Content Validity Evidence (Face Validity)	To determine if the test measures what it is supposed to measure	Literature review Expert review	N/A
Test Content Validity Evidence (Sampling Validity)	To determine the test includes a full representative sample of the construct being tested	Test Blueprint / Map	N/A
Internal Structure Validity Evidence (Internal Consistency)	To determine that the items act similarly to each other	Cronbach's alpha	1,745 U. S. students 1,619 Singapore students
Internal Structure Validity Evidence (Correlations)	To determine that the dimension strands act similarly to each other	Correlations amongst the dimension strands	1,745 U.S. students 1,619 Singapore students
Construct-Related Validity Evidence (Grade Progression)	To determine the means and standard deviations of scores at each grade level progress in a sensible way as children progress through the grades	Classical Test Theory	1,745 U.S. students
Construct-Related Validity Evidence (U.S. vs. Singapore)	To examine patterns in students more explicitly taught algebraic thinking skills	Examining how items function differently in U.S. vs. Singapore students	1,745 U.S. students 1,619 Singapore students
Consequences of Testing Validity Evidence (Teacher Use)	To determine how teachers are using the assessment scores	Quantitative and qualitative analysis of teacher survey responses	81 U.S. teachers
Response Process Validity Evidence (Qualitative Coding)	To determine the common thought processes and conceptions	Quantitative and qualitative coding of answers	1,745 U.S. students
Response Process Validity Evidence (Think-Alouds)	To determine the cognitive processes students go through when answering assessment items	Qualitative coding of think aloud protocol interviews	73 U.S. students

Chapter 4: The Development of a Diagnostic Assessment for Algebraic Thinking Skills

Traditional standardized tests have their strengths (i.e., because of their expense and wide-spread use they often have many studies indicating their validity and reliability) but utilizing these tests also has many drawbacks. While the information gained from traditional tests may help in determining if a student is high or low in that latent variable, these techniques do not provide the teacher with further information regarding how to proceed (i.e., what instructional decisions to make), nor can they capture growth, as they fail to make clear how student knowledge develops over time (Pellegrino, Chudowsky, & Glaser, 2001). Further, current criticisms of widely used assessments are that large-scale tests often do not provide extensive information about student knowledge and diagnostic tests are often difficult to interpret (Roberts & Gierl, 2010). This can be demonstrated in that while 93% of K-12 teachers surveyed believed that collecting diagnostic student data was important, teachers' most commonly used diagnostic tools were self-made (Huff & Goodman, 2007). A plethora of research has demonstrated the power of the use of formative assessments on student learning. A great deal of research has also concluded, unfortunately, that many teacher-made assessments lack validity and reliability (Black & Wiliam, 1998). Because of this research, developing and using assessments that are valid and reliable, particularly for students in the elementary grades, is a practice advocated by the National Math Panel (2008). It therefore seems useful and helpful to teachers to create valid and reliable diagnostic tools for them to use in the classroom to inform instruction and improve student learning. The development of a diagnostic assessment for algebraic thinking skills for elementary school students will be described below; its design and processes attempt to create a validated instrument that overcomes many of the measurement issues and flaws associated with current assessments available.

Cognitive Diagnostic Assessment

In response to the issues with traditional tests described above, Cognitively Diagnostic Assessment (CDA) has recently arisen as an assessment method that involves the combination of cognitive science with psychometrics, with the assessment purpose of directing instructional decisions (Nichols, 1994). This is in contrast to traditional testing techniques because these methods are more likely to place heavier emphasis on the psychometrics rather than the cognitive science, while CDA seeks to join the two (Huff & Goodman, 2007). CDA is defined by Leighton and Gierl (2007) as "designed to measure specific knowledge structures and processing skills in students so as to provide information about their cognitive strengths and weaknesses" (p. 3). Further, CDA strives to produce more information than many assessments: "In contrast with reporting a small number of content-based sub-scores, typical of most current educational test score reports, the results of a CDA yield a profile of scores with specific information about a student's cognitive strengths and weaknesses" (Roberts & Gierl, 2010, p. 25).

Like other forms of assessment, however, CDA is not useful if the information is not used to change instruction. Some teachers and supporters of CDA are under the false assumption that the use of CDA will automatically improve student's academic skills. Unfortunately, if the information provided by the CDA is unused and unconnected to instruction there is little research demonstrating that the use of CDA alone will improve learning (Wiliam & Black, 1996).

Although current test theory (i.e., Classical Test Theory and Item Response Theory) is beneficial for analyzing the results of high-stakes assessments, these analysis techniques are not necessarily appropriate for diagnostic assessments designed to make instructional decisions (Nichols, 1994). Current test theory is helpful and will be used in validating such instruments, but the scores themselves mean little to teachers. Nichols (1994) describes this problem

occurring even with performance or authentic assessments in that, while the assessments may be well-made and feature beneficial psychometric properties, the "scores indicate no more than the need for additional instruction" (Nichols, 1994, p. 578). Reporting raw scores alone is, therefore, not sufficient to improve instruction when utilizing CDA. Further, many of the more traditional methods for estimating reliability and validity revolve around the variance of the test scores, while diagnostic tests may demonstrate no variance, but still be considered a satisfactory and acceptable method of measuring the construct (Nichols, 1994).

Algebraic Thinking Diagnostic Assessment Development

Nichols (1994) recognized five steps to CDA test development, which should be used in the construction of a diagnostic assessment. Going through this five-step process did not appear to be conducted by many of the current algebraic thinking assessment developers, and utilizing these steps will not only likely improve the assessment quality but also aligns well with Kane's (2006) validation framework. These five steps include: 1) substantive theory construction, 2) design selection, 3) test administration, 4) response scoring, and 5) design revision.

Substantive theory construction. During substantive theory construction the model and the theory behind the construct are developed. As recommended by Nichols (1994), the construct dimensions are defined as well as the potential alternative conceptions, weak conceptions, and common arithmetic errors students may endure. This work is presented in the construct definition section.

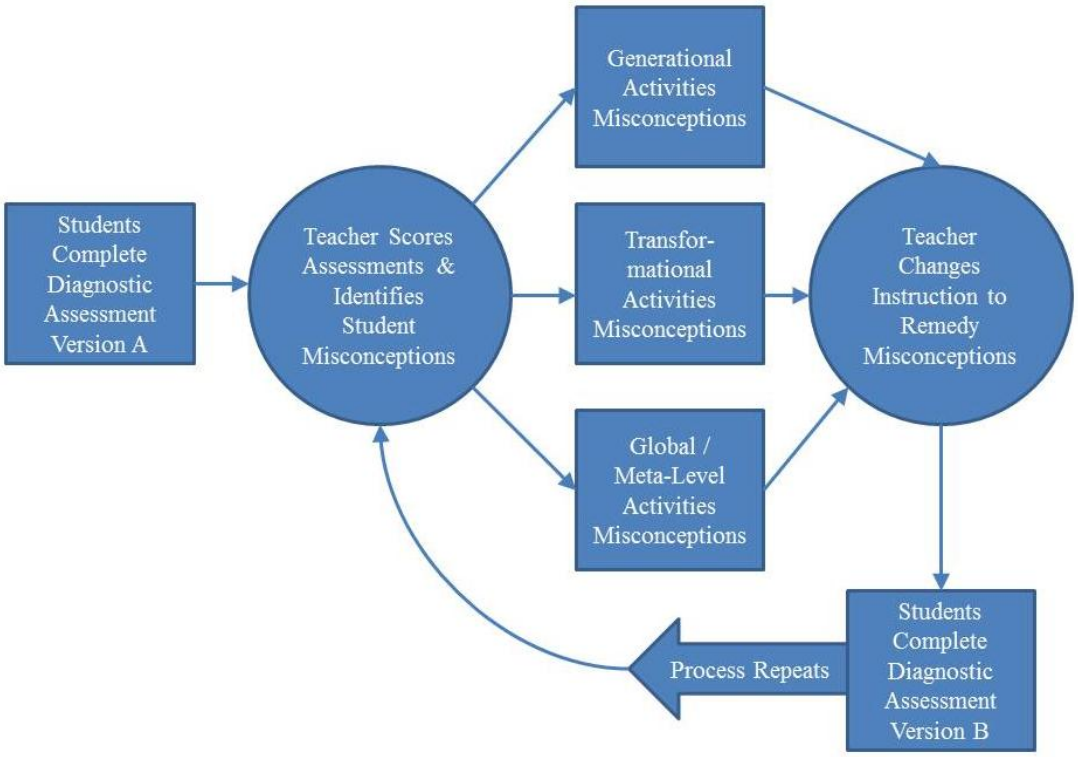
Design selection. In design selection, the methods of measurement for the diagnostic assessment are developed and the items written (Nichols, 1994). For this assessment, the test and item specifications are designed (i.e., see Appendix B and C for the Test Specifications & Item Specifications) to make explicit the exact tasks that will be required of students and so that

the test and items could be replicated by others. Multiple measurement methods (i.e., answer only, short answer, etc.) are included to strive to include a broad sampling of tasks in the domain (i.e., avoid underrepresentation) and control construct irrelevant variance (Kane, 2006).

In the development of any assessment, it is crucial to state upfront the potential interpretations and goals of the test's uses (Kane, 2006). This is in contrast to many of the previous assessments developed, which did not state the interpretive argument nor the validity argument up-front (Kane, 2006). The main interpretive argument for this assessment tool is clear. Because this assessment is meant to be used for diagnostic purposes, the design selected (i.e., test's uses) is a formative one: the information obtained from the assessment will be used to inform instruction and ideally improve student learning. This is an iterative process, and the reason for building two alternate versions of the assessment. If, for example, a student performs poorly on the equivalence items on the assessment, then the teacher should provide the student with remedial instruction in this area. If, on the other hand, a student performs well on the figural patterns items on the assessment, then the teacher does not need to provide the student with remedial instruction and may provide the student with enrichment activities in this area instead. These "if-then" statements form the backbone of any interpretive argument, and these types of instructional decisions form the backbone of any diagnostic or formative assessment (Kane, 2006; Black & Wiliam, 1998). This assessment design is detailed below in Figure 2.

Figure 2.

Algebraic Thinking Skills Assessment Design



The formulated interpretive arguments to be validated are displayed below in Table 8.

Table 8.

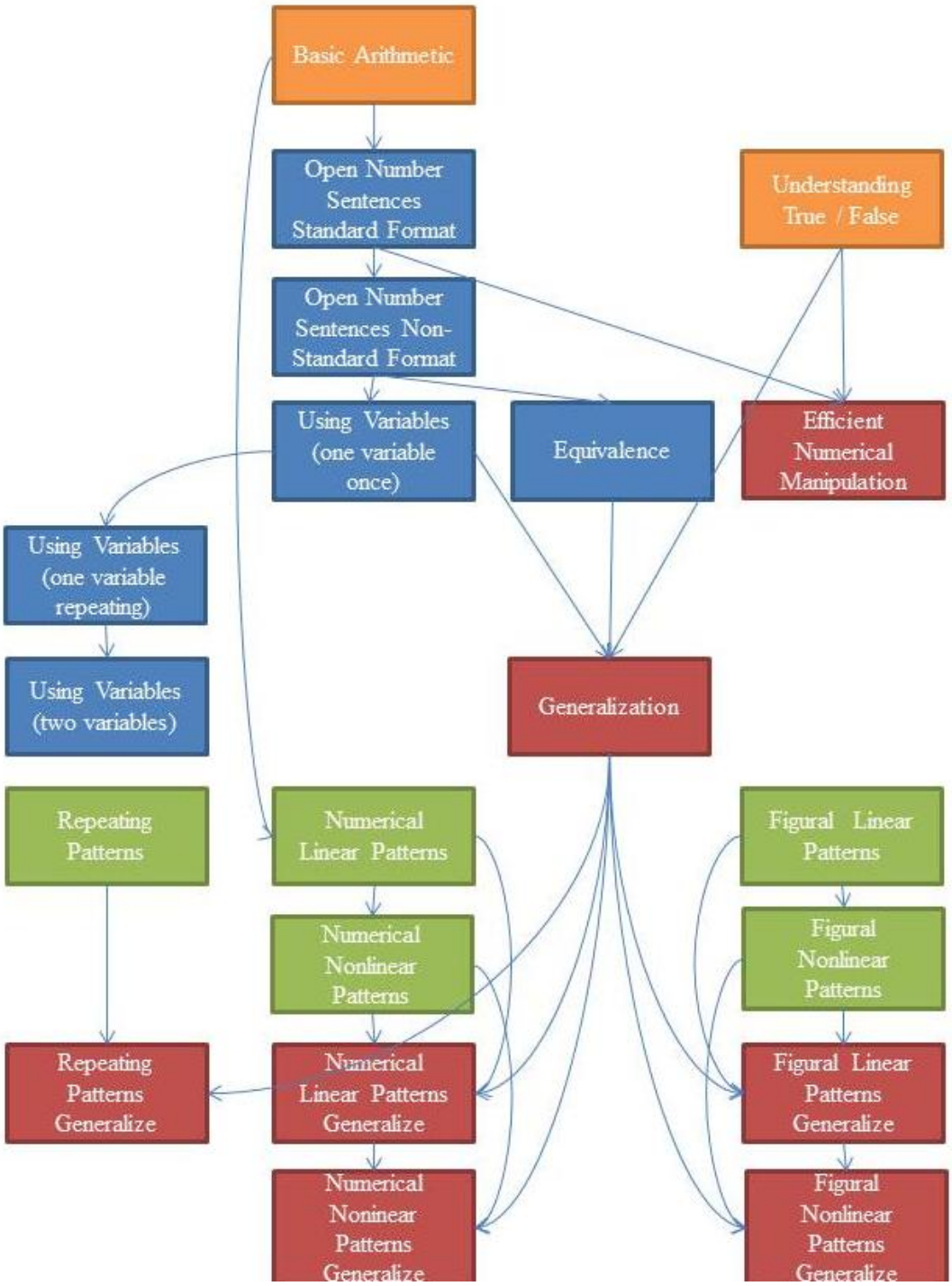
Interpretive Arguments for Diagnostic Assessment of Algebraic Thinking Skills

Interpretive Arguments Presented in If-Then Statements		Facet of Validity
If the two versions of the assessment are alternate forms…	Then there not be a statistically significant difference between the scores on the different versions.	Reliability
If the items can be scored consistently across different raters…	Then the inter-rater percent exact agreement should be at least 80% at the item level and higher at the total score level.	Reliability
If the assessment items truly test algebraic thinking skills…	Then an expert on algebraic thinking will agree that the assessment has 'face validity'.	Test Content Validity
If the assessment is representative of the universe of potential algebraic thinking assessment items…	Then a 'test blueprint' will demonstrate that a large enough set of items have been sampled from the universe.	Test Content Validity
If the examinees' responses to the assessment items are internally consistent…	Then Cronbach's alpha will be at least 0.70.	Internal Structure Validity
If the construct truly represents a unidimensional construct…	Then correlations amongst the dimension strands will all be statistically significant.	Internal Structure Validity
If the assessments truly represent the progression of skills…	Then students should improve on the skills as they progress through the grade levels.	Construct-Related Validity
If this assessment is diagnostic in nature…	Then the assessment should yield rich information about student alternative conceptions that may be used to guide instruction.	Construct-Related Validity
If the assessment is culturally valid…	Then students in countries focused on teaching algebraic thinking skills (i.e., Singapore) should perform higher on the assessment countries less focused on teaching algebraic thinking skills (i.e., the U.S.).	Construct-Related Validity
If this is a diagnostic assessment…	Then the results should be used only to modify instruction and change teaching practice / improve student learning.	Consequences of Testing Validity
If the assessment measures what it intends to measure…	Then the student cognitive processes revealed by think-aloud interviews and conception analysis will demonstrate rich algebraic thinking conceptions.	Response Process Validity

Several researchers have developed models for diagnosing student conceptions on assessments, beginning perhaps with Siegler's (1976) binary decision tree method. Taksuoka (1983), for example, recommended using diagnosis methods because "analysis of misconceptions can provide useful information in evaluating instruction or instructional materials as well as specific prescriptions for planning remediation for a student" (p. 345). This work in particular chose to incorporate the Attribute Hierarchy Method (AHM), which was used to further operationalize the assessment design through the development of the assessment map (see Figure 3) in the form of an Algebraic Skill Hierarchy (Leighton et al., 2004). AHM evolved from Tatsuoka's rule-space approach (1983, 1990), but unlike the rule-space approach AHM is designed to display the attributes required to solve test items and the order in which these attributes are needed (i.e., attribute 1 is often needed to solve items measuring attribute 2, attribute 2 is needed to solve items measuring attribute 3, etc.). In the model shown below in Figure 3, skills needed were defined more so than attributes. Developing an appropriate hierarchy is of utmost importance as it is a hypothesis as to the cognitive processes required by the assessment (Leighton et al., 2004). This model is still in draft form and has not yet been validated; future research is needed to continue this work (see Chapter 9).

Figure 3.

Algebraic Skill Hierarchy



Test administration. In test administration, all aspects of test administration are determined, including "the format of the items, the nature of the required response, the technology used to present test materials and record responses, and the environment of the testing session" (Nichols, 1994, p. 587). In accordance with measurement research, it has been determined that the assessment should be standardized, or administered in the same way across all settings and examinees. Standardizing measurement procedures strives to decrease the occurrence of construct irrelevant variance (Kane, 2006). The tool will, therefore, be accompanied by a specific set of administration directions for teachers and a set of student directions for teachers to read verbatim to students (see Appendix D). Time is not important in this assessment and, therefore, it is recommended that teachers allow enough time for all students to complete the assessment. Ideally teachers will give the students enough time for the majority of them to finish; however, it is understandable if they cannot spare more than 30 minutes. Teachers may break the test into two parts and administer it in two settings if they so desire, and this practice may be recommended for 1st and 2nd grade students who require more than 20 minutes to complete the test. Standardization has been implemented to reduce as much construct irrelevant variance as possible, but it is possible that speed may still become a trait tested by this assessment in cases where students cannot finish within 30 minutes (Kane, 2006). Researchers have found that reading ability is often a secondary dimension of mathematics ability (Jiban & Deno, 2007); therefore, it has been determined that little reading would be required in this assessment. Answer-only items were used whenever possible to simplify scoring and score interpretation for teachers. Finally, answer-only items were used instead of multiple-choice items because some research has shown that multiple-choice items draw student attention to the answer choices instead of the problem itself (Kazemi, 2002).

Response scoring. In response scoring, the method of scoring and interpreting scores is determined (Nichols, 1994). In the case of this algebraic thinking diagnostic tool, each assessment is worth a total of 24-26 points (please view the Test Specifications in Appendix B for more specificity). These points are determined simply by calculating the sum of correct answers. The majority of the questions are worth one point each, scored either correct (worth 1 point) or incorrect (worth 0 points). Item Response Theory (IRT) will be used to generate all item parameters. Item Characteristic Curves (ICC) and Test Characteristic Curves (TCC) will be calculated and utilized (van der Linden & Hambleton, 1996) to evaluate and select items for the assessment. Scale scores will be calculated using an interval scale in order to measure growth over time. Raw scores will be computed to scale scores because of the large range of difficulty of the different items on the assessment. The interval scale will be created using IRT to scale items and then compute an equal interval scale using item parameters.

Qualitative coding of conceptions will be conducted using a coding scheme previously established by other researchers for equivalence (McNeil & Alibali, 2004; Perry et al., 1988), meanings of the equal sign (Knuth et al., 2006; McNeil & Alibali, 2005a), and work with variables (MacGregor & Stacey, 1997).

A map of the assessment items was created to reduce the likelihood of construct under representativeness. Tables 9a, 9b, and 9c below displays this map of the assessment items by grade level and construct dimension. Several anchor items exist at more than one grade level. The items have been mapped so that anchor items can be used to link forms and to create a vertical (growth) scale so that students' development of algebraic thinking can be measured over time.

Table 9a.

Map of Algebraic Thinking Assessment Items: Modeling

Dimension		Assessment Version									
	Item	1A	1B	2A	2B	3A	3B	4A	4B	5A	5B
Modeling	Open Number Sentences										
	8=6+__	X									
	7=__+3		X								
	11=_-9			X		X					
	4=17-__				X						
	36=6x_						X		X	X	
	8=72/_							X			X
	8+_=15	X		X			X		X		X
	6+_=13		X		X	X		X		X	
	_-3=12	X		X			X				
	_-5=9		X		X						
	_x8=56					X		X			X
	_/8=4								X	X	
	Equivalence										
	8+4=_+5	X			X		X		X	X	
	6+7=_+4		X	X		X		X			X
	8+_=3+9		X			X					
	7+_=2+11	X									
	17+__=18_14			X			X				
	3x_=7+8								X	X	
	_-5=10-3	X			X						
	___+54=37+55							X			X
	12-5=10-__		X	X							
	3x6=7+_						X	X			X
	15+17=14+___				X	X					
	55+37=54+___								X	X	
	Equal Sign Meaning	X	X	X	X	X	X	X	X	X	X
	Using Variables										
	10-g=2	X		X		X					
	6+b=9		X		X		X				
	c+c+3=15		X		X	X			X	X	
	4+a+a=10	X		X			X	X			X
	4xn+5=21					X		X		X	
	n+n+n=n+12						X		X		X
	3+5+6=3+e	X			X	X		X			X
	7+4+5=7+e		X	X			X		X	X	
	x+y+y=10							X		X	
	x+x+y=12								X		X

Table 9b.

Map of Algebraic Thinking Assessment Items: Generalized Arithmetic

Dimension		Assessment Version									
	Item	1A	1B	2A	2B	3A	3B	4A	4B	5A	5B
Generalized Arithmetic	Efficent Numerical Manipulation										
	15+7-6	X			X						
	22+9-8		X	X							
	99+57-47=99+10					X					X
	89+44=87+46							X		X	
	99+99						X		X		
	(9x57)+57=10x57							X		X	
	3+4=7	X		X		X					
	7+6=6		X		X		X				
	8=5+13	X		X		X		X		X	
	7=2+5		X		X		X		X		X
	7=7	X		X		X		X		X	
	9=9		X		X		X		X		X
	4+0=4 / 4+0=0	X		X			X		X	X	
	9-9=0 / 9-9=9		X		X	X		X			X
	5+3=5-3	X		X		X					
	5+3=3+5		X		X						
	5x3=3x5						X		X		X
	5+(3x6)=48								X		X
	5+(3x6)=23							X		X	
	Generalization										
	a-a=0	X			X	X					
	a+0=a		X				X	X			
	ax0=a								X	X	
	ax1=a										X
	a+b=b+a			X		X			X		X
	a-b=b-a						X	X		X	

Table 9c.

Map of Algebraic Thinking Assessment Items: Functions

Dimension		Assessment Version									
	Item	1A	1B	2A	2B	3A	3B	4A	4B	5A	5B
Functions	Repeating Patterns										
	CirSquTri Near	X			X						
	CirSquTri Far	X			X						
	CiCiStar Near		X	X							
	CiCiStar Far		X	X							
	Numerical Linear Patterns										
	4,8,12	X			X		X				
	3,5,7,9		X	X							
	15,21,__		X		X	X		X		X	
	35,30,25								X		X
	72,69,66	X		X							
	Numerical Nonlinear Patterns										
	2,3,5,8				X	X			X	X	
	50,49,47			X			X	X			X
	Figural Linear Patterns										
	4,7,10 Near	X			X	X		X		X	
	4,7,10 Far					X		X		X	
	5,8,11 Near		X	X			X		X		X
	5,8,11 Far						X		X		X
	Figural Nonlinear Patterns										
	1, 4, 9 Near	X			X	X			X		X
	1, 4, 9 Far					X			X		X
	3,6,10 Near		X	X			X	X		X	
	3,6,10 Far						X	X		X	

Scoring for teachers. IRT will likely not be a method employable by teachers. Therefore it is necessary to outline what types of scores can be calculated by teachers using this assessment in their classrooms. An answer key will communicate to the teacher (or scorer) the correct answer. At least one question on every assessment is worth two points (i.e., the meaning of the equal sign tasks) and at least one question on every assessment is worth three points (i.e., the generalization task). Strand scores are calculated by simply calculating the sum for each of the three strands. Percentage correct can then be calculated by dividing the number of points the student received (total or by strand) by the total number of points possible (total or by strand).

Forms equating will be used to ensure that percent scores have the same meaning regardless of which test form is used. It is recommended that the classroom's scores then be aggregated across students to give the classroom teacher an idea as to how the classroom is performing overall. Further detail regarding specific conceptions encountered should also be detailed. In this way the teacher will know who (if any) of his or her students possess alternative algebraic thinking conceptions, which strands certain students struggle with the most, or if the majority of his or her class need further instruction in algebraic thinking.

Score meaning. The intention of using this algebraic thinking diagnostic assessment is to use the scores diagnostically, or to inform future instruction. Scores to be provided will include person parameters (i.e., Item Response Theory scores), raw scores, dimension raw scores, means, standard deviations, variance, etc. These scores are intended to be used diagnostically by teachers so teachers may make instructional decisions of teaching specific algebraic thinking standards to specific students who have not yet met those standards based on the results of the assessment. The scores are not intended to be used for any other method (i.e., to rank students, place students in groups, etc.) except diagnostically.

Design revision. In design revision, the assessment is revised before the five-stage cycle begins again (Nichols, 1994). The assessment proposed for this study has already been revised several times following pilot testing and interview protocols. Minor pretesting on the student instrument occurred in 2009 and 2010, resulting in the revised version that will be used in this study. In 2009, a pilot study was conducted with 375 elementary school students, including 53 1st grade students (14%), 54 2nd grade students (14%), 155 3rd grade students (41%), 113 4th grade students (30%), and 0 5th grade students (0%). Students completed the assessments anonymously (i.e., without putting names on papers) and, therefore, did not provide any further

demographic information. Assessment questions were revised following the results of this pilot to increase reliability (i.e., questions were added to the assessment), reduce confusion (i.e., only blank boxes were used instead of blank spaces), and induce more meaningful responses (i.e., more writing space was allowed for open-ended questions). Further, a larger number of questions was created to allow the development of assessments by grade level and alternate forms. In 2010, an interview protocol study was conducted with 52 elementary school students, including 17 1st grade students (33%), 15 2nd grade students (29%), and 20 5th grade students (38%). Students were asked to 'think-aloud' while solving problems to ensure they understood the questions being asked. Assessment questions were revised following the results of this think-aloud protocol study to further reduce confusion and make the items more age-appropriate at each of the five levels. After the current research study is completed the assessment tools will again be revised, as needed, and this revision cycle will be continual over time.

Expert Analysis. The expert opinions of Dr. Linda Levi were sought to develop content validity of the assessment tool. Dr. Linda Levi is one of the leaders in the field of algebraic thinking. She co-authored the (2003) *Thinking Mathematically: Integrating Arithmetic & Algebra in Elementary School* casebook with Thomas Carpenter and Megan Franke on which began my inquiry into the world of algebraic thinking. She has also worked on Cognitively Guided Instruction (CGI), a book and part of the Teachers Development Group nonprofit organization committed to helping teachers understand children's mathematical thinking.
Dr. Linda Levi had several points regarding the assessments, some of which were taken into account and others that were not for various reasons. Dr. Levi was concerned about the use of the box in that "you may be intending to use the box as a variable in some of your equations. If this is true, you shouldn't use the box as a place to write the answer". This suggestion was not

utilized, however, because in pilot tests students appeared confused when a blank was used since many curriculums students are used to often utilize the box as a variable. For my work I intended the box as a blank; as a space to write the answer to the problem; rather than as a variable.

Overall, Dr. Levi did agree that “many people would see the assessment that you wrote as addressing the important ideas for algebraic thinking”, which adds to the content validity of the assessment and the results that are derived.

Chapter 5: Results of the Washington State Sample

In this chapter I will describe the results obtained from the Washington State sample of students. In this sample 1,745 students from six different schools in Washington State were administered the diagnostic assessment of algebraic thinking skills. This chapter will explore the results of several analyses conducted to begin validating the interpretation of the assessment scores. First, item analyses including Classical Test Theory analyses, Item Response Theory analyses, and discrimination indices were conducted to better understand the construct. Second, reliability analyses including inter-rater reliability (i.e., to ensure the assessments can be scored similarly by multiple raters), alternate forms reliability (i.e., to ensure the various versions of the assessment are indeed alternate forms of each other), and internal structure validity evidence (i.e., to determine that the items and dimension strands act similarly to each other) are reported. Third, construct-related validity evidence in the form of how scores change as students progress up through the grade levels are reported.

Item Difficulty and Discrimination Results: Classical Test Theory and Item Response Theory Analyses

Both Classical Test Theory (CTT) and Item Response Theory (IRT) analyses were used to analyze all item statistics. CTT p values are reported for each item, which describes the proportion of students who solved that particular item correctly. An item with a p value of 0.75, for instance, indicates that 75% of those students solved that item correctly. IRT difficulty location parameters are reported for each item in addition. In general the larger the negative number, the easier the item, and the larger the positive number, the harder the item. An item with an IRT difficulty parameter of -2.35, for example, indicates a fairly easy item, while an item with an IRT difficulty parameter of 4.50 indicates a very difficult item. Although some

researchers have recommended ‘ideal’ difficulty parameters, such as CTT scores of 0.3 to 0.7, because this is a diagnostic assessment these requirements do not hold for this work. Very high scores or very low scores may need to be investigated, however, to ensure they are appropriate items for that grade level.

Both the adjusted item-total correlations and the IRT discrimination parameter were used to measure discriminability of items. Items largely had positive and statistically significant ($p <$ .001) adjusted item-total correlations, indicating that as a student performed better on that item, they also performed better on the assessment as a whole. Many of the items also exhibited IRT discrimination parameters greater than 1.00. A score of less than 1.00 indicates that the item discriminates between high and low performers less than expected for an item of that difficulty while a score of greater than 1.00 indicates that the item discriminates between high and low performers better than expected for an item of that difficulty. Because of the diagnostic assessment purpose, however, it was generally determined that all items will be retained for this assessment because of their uses diagnostically and for comparison despite being low in discriminability.

Both the mean-square infit and mean-square outfit statistics were also reported. Poor fit can often reveal multidimensionality issues, or even multidimensionality issues only at specific grade levels, therefore it is an important measure to analyze. The mean-square fit statistics show the size of the randomness; therefore 1.0 is the expected and desired value. Values less than 1.0 indicate that the observations are too predictable (i.e., the data overfit the model) while values greater than 1.0 indicate that the observations are more unpredictable (i.e., the data underfit the model). A general rule of thumb is that mean-square values of 0.5-1.5 are productive for measurement, mean-square values of 0.00-0.5 and 1.5-2.0 may be unproductive for measurement

but are generally acceptable, and mean-square values greater than 2.0 need to be investigated as they may be distorting the measurement system. Infit issues are also much more concerning than outfit issues. Out of all the items at all of the different grade levels, no items were found with an infit statistic greater than 2.0, and only four items were discovered with an outfit statistic greater than 2.0. These items included the equivalence-context definition of the equal sign for 3rd grade students, the true or false statement of 5+(3x6)=23 for 4th grade students, and the open number sentence 8=72/__ for 4th and 5th grade students. Therefore it appears that the items fit the model well, and there does not appear to be any question of multidimensionality.

Information regarding the sample sizes of students completing each item for each dimension is listed in Tables 10a and 10b.

Table 10a.

Sample Sizes for all Items

	1st Grade	2nd Grade	3rd Grade	4th Grade	5th Grade	All
Open Number Sentences						
8+___=15	177	155	172	186	183	873
6+___=13	174	154	163	198	182	871
8=6+___	177					177
7=___+3	174					174
___-3=12	176	155	172			503
___-5=9	174	154				328
11=___-9		151	162			313
4=17-___		153				153
___x8=56			163	198	182	543
36=6x___			173	186	182	541
8=72/___				198	183	381
___/8=4				186	182	368
Equivalence						
8+4=___+5	177	150	172	186	182	867
6+7=___+4	174	153	162	197	183	869
7+___=2+11	175					175
8+___=3+9	173		161			334
___-5=10-3	176	147				323
12-5=10-___	173	153				326
17+___=18+14		151	172			323
15+17=14+___		153	162			315
___+54=37+55				198	183	381
55+37=54+___				185	181	366
Meaning of the Equal Sign						
5+3=8 Equal Sign Name	154	144	160	187	183	828
5+3=8 Equal Sign Definition	152	144	158	186	183	823
2+7=5+4 Equal Sign Name	171	139	162	189	177	838
2+7=5+4 Equal Sign Definition	169	138	160	189	176	832
Work with Variables						
10-g=2	177	151	161			489
6+b=9	173	153	171			497
4+a+a=10	176	149	171	198	183	877
c+c+3=15	173	152	160	186	181	852
3+5+6=3+e	174	148	158	198	183	861
7+4+5=7+e	172	152	171	186	180	861
4xn+5=21			161	198	180	539
n+n+n=n+12			171	185	180	536
x+y+y=10				196	178	374
x+y=6						
x+x+y-12				184	183	367
x-y=3						
Numerical Patterns						
3, 5, 7, 9, 11,___	173	152				325
4, 8, 12, ___, 20, 24	173	150	172			495
15, 21, ___, 33, 39	173	152	158	198	180	861
72, 69, 66, 63, ___	172	148				320
2, 3, 5, 8, ___		150	159	183	179	671
50, 49, 47, 44, ___		150	170	196	176	692

Table 10b.

Sample Sizes for all Items

	1st Grade	2nd Grade	3rd Grade	4th Grade	5th Grade	All
True / False Statements						
3+4=7	171	150	160			481
7+6=6	173	152	171			496
8=5+13	171	150	160	198	181	860
7=2+5	173	152	171	186	183	865
7=7	171	150	160	197	180	858
6=6	171	151	171	186	183	862
4+0=4	171	150				321
4+0=0			171	185	180	536
9-9=9	173	152	160	197	183	865
5+3=5-3	170	150	160			480
5+3=3+5	172	152				324
5x3=3x5			171	185	182	538
15+7-6=15+1	170				183	353
22+9-8=22+1	171	146				317
99+57-47=99+10		151	161			312
99+99=100+100-2			170	185		355
(9x57)+57=10x57				197	178	375
89+44=87+46				197	180	377
5+(3x6)=23				197	178	375
5+(3x6)=48				185	183	368
Figural Patterns						
Shape Repeating ABCABC	170	152				322
Shape Repeating ABCABC Far	168	146				314
Shape Repeating AABAAB	170	153				323
Shape Repeating AABAAB Far	163	149				312
Linear Figure 4, 7, 10, ___	158	143	160	189	184	834
Linear Figure 4, 7, 10, ___ Far			160	187	183	530
Linear Figure 5, 8, 11, ___	175	147	164	192	178	856
Linear Figure 5, 8, 11, ___ Far			163	187	178	528
Nonlinear Figure 1, 4, 9, ___	149	135	154	191	176	805
Nonlinear Figure 1, 4, 9, ___ Far			143	181	173	497
Nonlinear Figure 3, 6, 10, ___	166	143	159	187	181	836
Nonlinear Figure 3, 6, 10, __ Far			151	184	178	513
Generalization						
a-a=0 true / false	160	145	160			465
a-a=0 explain	158	142	158			458
a+0=a true / false	174		170	189		533
a+0=a explain	172		169	188		529
ax0=a true / false				184	180	364
ax0=a explain				185	179	364
a+b=b+a true / false		147	159	192	180	678
a+b=b+a explain		145	156	190	177	668
a-b=b-a true / false			164	195	184	543
a-b=b-a explain			161	194	182	537
ax1=a true / false					178	178
ax1=a explain					176	176

Open Number Sentences

Twelve different open number sentence items were used across the different grade levels and the different versions of the assessments. The items varied on where the blank was positioned (i.e., first, second, or third position), the type of operation (i.e., addition, subtraction, multiplication, or division), and whether it was a 'traditional' item (i.e., if the answer came first) or not.

The item statistics of difficulty and discrimination from both CTT and IRT for the open number sentence items are summarized in Table 11. As you can see students were fairly proficient on the open number sentences: 79% and 82% of 1st grade students were capable of solving the most simple open number sentence items and these values neared 100% by 3rd grade. Among all the open number sentence items offered, students in lower grades (i.e., the 2nd and 3rd grades) only struggled with the non-traditional subtraction when the blank came first (i.e., 11=___-9) as an averaged p value of 0.35 whereas students in 4th grade and above struggled with the division when the blank came first (i.e., ___/8=4) as an averaged p value of 0.43.

For the two items used across all the grades, there was a statistically significant difference by grade for the item 8+__=15 (F (4, 868) = 23.586, $p < .001$) and for the item 6+__=13 (F (4, 866) = 18.456, $p < .001$). Post-hoc Tukey analyses revealed that 1st grade students significantly underperformed all other grades on both items.

The discrimination indices were quite low for the open number sentence items; however this was expected based on the high p values and the ceiling effect. Although many adjusted item-total correlations approached zero, none were significantly negative; therefore these items are appropriate from a discrimination standpoint.

Table 11.

Item Statistics for Open Number Sentence Items

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
8+___=15						
1st Grade	0.79	-2.34	0.000	1.00	1.02	0.87
2nd Grade	0.92	-2.86	0.169	0.99	1.04	1.10
3rd Grade	0.98	-3.71	0.113	0.93	1.14	1.32
4th Grade	0.99	-4.40	0.000	0.91	1.18	1.07
5th Grade	0.99	-4.58	0.000	1.10	0.86	0.05
6+___=13						
1st Grade	0.82	-2.55	0.185	0.89	1.08	1.31
2nd Grade	0.92	-3.07	0.370	1.07	0.97	0.51
3rd Grade	0.99	-5.35	0.090	0.93	1.10	1.02
4th Grade	0.99	-4.61	0.000	1.05	0.98	0.21
5th Grade	0.98	-2.99	0.222	1.07	0.79	0.56
8=6+___						
1st Grade	0.46	-0.34	0.385	1.13	0.94	0.99
7=___+3						
1st Grade	0.65	-1.37	0.386	1.13	0.94	0.86
___-3=12						
1st Grade	0.47	-0.38	0.290	0.88	1.04	1.05
2nd Grade	0.66	-0.70	0.485	1.08	0.95	0.93
3rd Grade	0.76	-0.71	-0.015	0.71	1.15	1.67
___-5=9						
1st Grade	0.40	-0.04	0.395	1.05	0.98	0.95
2nd Grade	0.62	-0.61	0.403	0.80	1.11	1.06
11=___-9						
2nd Grade	0.31	1.31	0.337	0.84	1.12	1.06
3rd Grade	0.39	1.38	0.400	0.85	1.06	1.08
4=17-___						
2nd Grade	0.50	0.11	0.197	0.53	1.17	1.78
___x8=56						
3rd Grade	0.79	-1.00	0.364	0.99	0.98	1.21
4th Grade	0.90	-1.94	0.462	1.18	0.81	0.43
5th Grade	0.91	-1.37	0.276	0.99	1.03	1.32
36=6x___						
3rd Grade	0.82	-1.21	0.215	1.00	0.97	1.13
4th Grade	0.91	-1.91	0.088	0.98	1.03	0.96
5th Grade	0.96	-2.34	0.272	1.01	1.01	0.68
8=72/___						
4th Grade	0.85	-1.42	0.264	0.86	1.07	3.76
5th Grade	0.87	-0.82	-0.039	0.70	1.26	2.19
___/8=4						
4th Grade	0.41	1.56	0.481	1.06	0.97	0.98
5th Grade	0.45	1.85	0.504	1.40	0.85	0.80

Equivalence

Ten different equivalence items were used across the different grade levels and the different versions of the assessments. The items varied on where the blank was positioned (i.e., first, second, third, or fourth position), the type of operation (i.e., addition, subtraction, multiplication, division, or mixed), and whether it was an arithmetic or relational item (i.e., whether it was simpler to use relational strategies to solve the problem than using arithmetic). The item statistics of difficulty and discrimination from both CTT and IRT for the equivalence items are summarized in Table 12. It appears students struggled much more with equivalence type items than they did with open number sentences, even when the arithmetic difficulty remained the same. Only 11-15% of 1st grade students were capable of solving equivalence problems like 8+4=___+5 correctly, and even in 5th grade only 66-69% of students could solve such items correctly.

Understandably, students struggled more with solving equivalence items that were intended to measure relational thinking skills. In these cases, using relational thinking strategies (i.e., in 17+__=18+14, 18 is 1 more than 17 so the blank must be 1 more than 14) would make solving the items much easier and simpler than fully calculating the problem. In these cases students needed to not only understand equivalence but also master either relational thinking or higher order arithmetic skills. Interestingly enough, despite this increase in complexity, 4th and 5th grade students actually solved these types of problems at the same or slightly higher rates than the simpler problem 8+4=___+5 (i.e., p values ranged from 0.53 to 0.70 for these types of problems). The difference here may lie in where the blank lies; future research is needed to discover this.

There was a statistically significant difference by grade for the item 8+4=___+5 (F (4, 862) = 45.620, $p < .001$) and for the item 6+7=___+4 (F (4, 864) = 39.660, $p < .001$). Post-hoc Tukey analyses for 8+4=___+5 revealed that 1st and 2nd grade students significantly underperformed all other grades, and that 4th grade students underperformed 5th grade students ($p < .01$). Post-hoc Tukey analyses for 6+7=___+4 revealed that although 1st and 2nd grade students did not differ significantly from each other, they both significantly underperformed 3rd, 4th, and 5th grade students.

Both CTT and IRT discrimination indices were very high for all open number sentence items. All adjusted item-total correlations were statistically significant ($p < .001$) and all IRT discrimination parameters were greater than 1.00 (i.e., the item discriminates better than expected for its difficulty level). Therefore these items all appear to be of high-quality discrimination wise.

Table 12.

Item Statistics for Equivalence Items

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
8+4=___+5						
1st Grade	0.11	2.19	0.538	1.11	0.89	0.57
2nd Grade	0.30	1.28	0.604	1.28	0.81	0.67
3rd Grade	0.56	0.46	0.481	1.39	0.84	0.75
4th Grade	0.52	0.99	0.527	1.42	0.83	0.77
5th Grade	0.69	0.59	0.544	1.34	0.80	0.77
6+7=___+4						
1st Grade	0.15	1.71	0.476	1.13	0.84	0.76
2nd Grade	0.29	1.42	0.721	1.49	0.66	0.52
3rd Grade	0.62	0.15	0.566	1.57	0.70	0.62
4th Grade	0.53	0.77	0.460	1.41	0.79	0.74
5th Grade	0.66	0.76	0.453	1.30	0.84	0.75
7+___=2+11						
1st Grade	0.11	2.18	0.636	1.17	0.80	0.48
8+___=3+9						
1st Grade	0.18	1.44	0.537	1.25	0.75	0.62
3rd Grade	0.75	-0.65	0.506	1.32	0.79	0.59
___-5=10-3						
1st Grade	0.06	2.90	0.542	1.10	0.83	0.72
2nd Grade	0.22	1.89	0.632	1.36	0.69	0.46
12-5=10-___						
1st Grade	0.13	1.89	0.588	1.27	0.69	0.40
2nd Grade	0.29	1.43	0.574	1.42	0.67	1.42
17+___=18+14						
2nd Grade	0.26	1.69	0.666	1.47	0.65	0.48
3rd Grade	0.52	0.68	0.526	1.42	0.84	0.74
15+17=14+___						
2nd Grade	0.25	1.57	0.743	1.46	0.63	0.44
3rd Grade	0.52	0.65	0.589	1.40	0.81	0.81
__+54=37+55						
4th Grade	0.53	0.81	0.545	1.48	0.75	0.76
5th Grade	0.70	0.53	0.583	1.43	0.74	0.60
55+37=54+___						
4th Grade	0.53	0.93	0.306	1.05	0.98	0.96
5th Grade	0.69	0.58	0.511	1.32	0.81	0.77

Meaning of the Equal Sign

Students were asked to not only identify the equal sign but also to provide a definition for the equal sign. Two different meaning of the equal sign items were used: one on each of the two

different versions for each grade level. The items varied whether the sample provided was in an equivalence context (i.e., 2+7=5+4) or a non-equivalence context (i.e., 5+3=8). The item statistics of difficulty and discrimination from both CTT and IRT for the meaning of the equal sign items are summarized in Table 13.

The majority of students, starting with 60% of 1st grade students and ending with 97% of 5th grade students could name the '=' symbol (i.e., the equal sign). Despite this capacity, a significantly smaller number of students could articulate the correct, relational meaning of the equal sign (i.e., the same). Many of the younger students failed to answer at all and many of the older students answered with the operational meaning of the equal sign (i.e., the answer).

There was a statistically significant difference by grade for both the equivalence (i.e., 2+7=5+4) context (F (4, 833) = 35.596, $p < .001$) and the non-equivalence (i.e., 5+3=8) context (F (4, 823) = 26.001, $p < .001$) for the equal sign identification task. Post-hoc Tukey analyses revealed that 1st and 2nd grade students significantly underperformed all other grades, and that 4th grade students underperformed 5th grade students ($p < .01$). Additionally there was a statistically significant difference by grade for both the equivalence (i.e., 2+7=5+4) context (F (4, 827) = 17.217, $p < .001$) and the non-equivalence (i.e., 5+3=8) context (F (4, 818) = 11.185, $p < .001$) for the meaning of the equal sign task. Both contexts revealed the same results: although there was not a significant difference between 1st and 2nd grade students, both 1st and 2nd grade students significantly underperformed 3rd, 4th, and 5th grade students. This may be due to literacy and writing effects rather than an actual knowledge effect.

When students were asked to produce the meaning of the equal sign, students were significantly more successful in the equivalence context (F (1, 1,664) = 7.206, $p < .001$). Although a visual analysis of the data produces the possibility that the equivalence context may

lead to more students providing a relational definition of the equal sign instead of an operational definition, a regression model revealed that the equivalence context was not a unique predictor after grade was taken into account ($p > .05$).

Although five of the adjusted item-total correlations approach zero and several are even negative, none are significantly negative and therefore it is appropriate that they be kept. It is interesting to note that while these items were significantly correlated with total scores in 1st and 2nd grade, they were largely insignificant for 3rd, 4th, and 5th grade students. Perhaps this is because these students largely knew the name of the equal sign? Although these items were better discriminators for younger students, it is important to keep such items to analyze how student knowledge changes as students age.

Table 13.

Item Statistics for the Meaning of the Equal Sign Items

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
5+3=8 Equal Sign Name						
1st Grade	0.70	-1.60	0.326	1.07	0.96	0.88
2nd Grade	0.77	-1.44	0.441	0.98	1.01	1.08
3rd Grade	0.94	-2.68	0.217	0.93	1.06	1.96
4th Grade	0.97	-3.00	0.175	0.92	1.15	1.65
5th Grade	0.96	-2.50	-0.058	0.84	1.25	1.64
5+3=8 Equal Sign Definition						
1st Grade	0.17	1.59	0.297	0.89	1.24	1.00
2nd Grade	0.24	1.81	0.396	0.70	1.30	1.25
3rd Grade	0.31	1.99	-0.005	0.35	1.48	1.47
4th Grade	0.36	2.20	0.334	0.56	1.30	1.34
5th Grade	0.34	2.94	0.153	0.47	1.39	1.55
2+7=5+4 Equal Sign Name						
1st Grade	0.60	-1.01	0.437	1.10	0.96	0.94
2nd Grade	0.72	-0.96	0.254	0.88	1.06	1.11
3rd Grade	0.88	-1.65	0.142	0.88	1.12	1.25
4th Grade	0.95	-2.77	0.140	1.04	0.92	0.69
5th Grade	0.97	-2.20	-0.003	0.92	1.13	1.50
2+7=5+4 Equal Sign Definition						
1st Grade	0.15	1.72	0.432	0.99	1.08	0.97
2nd Grade	0.25	1.81	0.466	0.71	1.36	1.23
3rd Grade	0.40	1.34	0.099	0.21	1.55	2.47
4th Grade	0.40	1.45	0.294	0.49	1.33	1.37
5th Grade	0.40	2.19	0.098	0.15	1.35	1.73

Work with Variables

Ten different work with variables items were used across the different grade levels and the different versions of the assessments. The items varied on the letter used as the variable (i.e., g, b, a, etc.), whether one or two of the same variables appeared, whether one or two different variables appeared, whether there was an "equivalence context" (i.e., 3+5+6=3+e) or not (i.e., 6+b=9), and the type of operation (i.e., addition, subtraction, multiplication, or division). The item statistics of difficulty and discrimination from both CTT and IRT for the work with variables items are summarized in Table 14.

Although students often initially struggled to solve such items in 1st and 2nd grades (i.e., with p values of 0.38 to 0.54), by 3rd grade they were fairly proficient (i.e., with p values of 0.81 to 0.88) with one variable items (i.e., 6+b=9) and by 5th grade they were fairly proficient (i.e., with p values of 0.74 and above) with both two variable items (i.e., 4+a+a=10) and variables use in an equivalence context (i.e., 3+5+6=3+e). Older students were asked to solve more complex items involving variables, and students struggled accordingly. Students surprisingly mastered (i.e., with p values of 0.47 to 0.62) the most complex problems involving two different variables (i.e., if x+y+y=10 and x+y=6 what is x and y?) fairly well, but continued to struggle throughout the grade levels (i.e., with p values of 0.20 to 0.28) when the variables occurred on both sides of the equal sign (i.e., n+n+n=n+12).

There was a statistically significant difference by grade for all items: for 4+a+a=10 (F (4, 872) = 116.821, $p < .001$), for c+c+3=15 (F (4, 847) = 72.880), $p < 001$), for 3+5+6=3+e (F (4, 856) = 69.993, $p < .001$), and for 7+4+5=7+e (F (4, 856) = 73.106), $p < .001$). Post-hoc Tukey analyses for 4+a+a=10, 3+5+6=3+e, and 7+4+5=7+e revealed that 1st and 2nd grade students significantly underperformed all other grades, while 3rd grade students significantly underperformed 5th grade students. Similarly, analyses for c+c+3=15 revealed that 1st and 2nd grade students significantly underperformed all other grades, however there were no other differences.

Like the equivalence items, the work with variables items were typically very good discriminators. Nearly all of the adjusted item-total correlations are statistically significant ($p < .001$), and nearly all of the IRT discrimination parameters are greater than 1.00 (i.e., indicating better than expected discrimination for that difficulty level). Therefore it appears appropriate to keep all of the items.

Table 14.

Item Statistics for Work with Variables

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
10-g=2						
1st Grade	0.40	-0.01	0.495	1.09	0.97	0.89
2nd Grade	0.54	0.00	0.597	1.41	0.83	0.71
3rd Grade	0.88	-1.72	0.355	1.08	0.93	0.73
6+b=9						
1st Grade	0.38	0.11	0.416	1.08	0.95	0.99
2nd Grade	0.51	0.02	0.629	1.53	0.77	0.65
3rd Grade	0.81	-1.08	0.446	1.28	0.78	0.53
4+a+a=10						
1st Grade	0.20	1.28	0.567	1.16	0.89	0.67
2nd Grade	0.34	1.16	0.570	1.17	0.91	0.80
3rd Grade	0.73	-0.49	0.525	1.33	0.79	0.65
4th Grade	0.86	-1.47	0.461	0.98	1.03	0.90
5th Grade	0.91	-1.38	0.142	1.10	0.84	0.69
c+c+3=15						
1st Grade	0.14	1.77	0.524	0.99	1.03	0.88
2nd Grade	0.32	1.16	0.566	1.35	0.79	0.60
3rd Grade	0.68	-0.20	0.560	1.29	0.84	0.72
4th Grade	0.69	0.05	0.217	1.05	0.95	1.10
5th Grade	0.80	-0.14	0.340	0.92	1.06	1.09
3+5+6=3+e						
1st Grade	0.09	2.40	0.509	1.01	0.88	1.28
2nd Grade	0.25	1.64	0.645	1.37	0.71	0.52
3rd Grade	0.56	0.52	0.623	1.49	0.78	0.71
4th Grade	0.65	0.08	0.434	1.35	0.80	0.70
5th Grade	0.74	0.28	0.531	1.17	0.87	0.85
7+4+5=7+e						
1st Grade	0.12	2.02	0.534	1.16	0.78	0.65
2nd Grade	0.25	1.74	0.659	1.37	0.69	0.79
3rd Grade	0.54	0.58	0.646	1.72	0.69	0.61
4th Grade	0.62	0.45	0.575	1.57	0.75	0.62
5th Grade	0.82	-0.26	0.482	1.25	0.78	0.62
4xn+5=21						
3rd Grade	0.50	0.82	0.628	1.53	0.78	0.69
4th Grade	0.70	-0.21	0.306	1.15	0.91	0.80
5th Grade	0.79	-0.08	0.217	1.03	0.94	1.05
n+n+n=n+12						
3rd Grade	0.20	2.54	0.280	1.11	0.93	0.77
4th Grade	0.26	2.44	0.330	1.24	0.83	0.80
5th Grade	0.28	2.87	0.208	0.84	1.09	1.25
x+y+y=10; x+y=6						
4th Grade	0.55	0.71	0.514	1.25	0.88	0.79
5th Grade	0.62	1.01	0.445	1.11	0.95	0.93
x+y+y=12; x-y=3						
4th Grade	0.47	1.25	0.562	1.33	0.87	0.79
5th Grade	0.54	1.42	0.351	0.96	1.01	1.04

Numerical Patterns

Seven different numerical pattern items were used across the different grade levels and the different versions of the assessments. The items varied on where the blank was positioned (i.e., at the end of the list of numbers or in the middle), the type of pattern (i.e., linear or nonlinear), and whether it was increasing or decreasing. The item statistics of difficulty and discrimination from both CTT and IRT for the numerical patterns items are summarized in Table 15.

As you can see for students in the younger grades mastery of the items varied widely as the value the pattern was counting by varied (i.e., students solved counting by 2's at much higher rates than counting by 4's and solved counting by 4's at much higher rates than counting by 6's). 1st and 2nd grade students, for example, solved counting by 2's patterns at p value rates of 0.63 and 0.84, respectively, while 1st and 2nd grade students solved counting by 4's patterns at p value rates of 0.42 and 0.75, respectively. As students progressed through the grades the level of mastery equalized across the different types of patterns as students reached the ceiling (100% mastery) by 4th and 5th grade. Students understandably struggled much more with nonlinear numerical patterns than they did with linear numerical patterns, however the majority of students answered correctly by 4th and 5th grades. Younger students especially found these items difficult: 2nd grade students, for example, solved nonlinear numerical pattern items correctly at p values of 0.25 to 0.27.

There was a statistically significant difference by grade for the item 15, 21, ___, 33, 39 (F (4, 856) = 71.574, $p < .001$). Post-hoc Tukey analyses revealed that 1st and 2nd grade students significantly underperformed all other grades ($p < .001$). There was also a statistically significant difference by grade for both the item 2, 3, 5, 8, ___ (F (3, 667) = 20.882, $p < .001$)

and for the item 50, 49, 47, 44, ___ (F (3, 688) = 33.071, $p < .001$). Post-hoc Tukey analyses revealed that 2nd grade students significantly underperformed all other grades, and that 3rd and 4th grade students underperformed 5th grade students ($p < .01$).

The discrimination indices indicated the numerical patterns to be fairly good discriminators. Many of the adjusted item-total correlations were statistically significant ($p < .001$) and the IRT discrimination parameters mostly hovered around 1.00 (i.e., of average discrimination for that difficulty level). It does appear discrimination capacity may decrease for older students, which makes sense as a ceiling effect is reached and p values approach 1.00. It is possible items such as 35, 30, __, 20, 15 should be removed for 5th grade students because they are too easy, however as this is a diagnostic assessment it is important for teachers to know if students are not capable of solving such items even if only 2% of 5th grade students fail to do so. Because of this diagnostic need, all items will be retained.

Table 15.

Item Statistics for Numerical Patterns

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
3, 5, 7, 9, 11, ___						
1st Grade	0.63	-1.25	0.429	0.96	1.03	0.94
2nd Grade	0.84	-1.93	0.323	0.99	1.00	1.09
4, 8, 12, ___, 20, 24						
1st Grade	0.42	-0.12	0.479	1.37	0.84	0.79
2nd Grade	0.75	-1.34	0.484	1.29	0.82	0.63
3rd Grade	0.90	-2.00	0.215	0.95	1.06	1.07
15, 21, ___, 33, 39						
1st Grade	0.28	0.67	0.494	1.18	0.90	0.76
2nd Grade	0.60	-0.46	0.490	1.02	0.98	1.01
3rd Grade	0.85	-1.43	0.322	0.92	1.06	1.09
4th Grade	0.85	-1.37	0.462	1.16	0.85	0.59
5th Grade	0.88	-0.89	0.276	1.07	0.91	0.89
35, 30, 25, ___, 15						
4th Grade	0.95	-2.64	0.140	0.92	1.11	1.41
5th Grade	0.98	-3.08	0.156	1.01	0.96	0.97
72, 69, 66, 63, ___						
1st Grade	0.26	0.84	0.551	1.23	0.85	0.72
2nd Grade	0.69	-0.79	0.511	1.33	0.81	0.70
2, 3, 5, 8, ___						
2nd Grade	0.27	1.46	0.431	0.92	1.10	0.93
3rd Grade	0.47	0.93	0.521	1.01	1.00	0.96
4th Grade	0.54	0.87	0.288	0.89	1.04	1.06
5th Grade	0.69	0.61	0.164	0.83	1.09	1.13
50, 49, 47, 44, ___						
2nd Grade	0.25	1.79	0.564	0.99	1.02	0.92
3rd Grade	0.42	1.21	0.454	1.21	0.91	0.86
4th Grade	0.53	0.79	0.403	1.01	1.00	0.98
5th Grade	0.75	0.25	0.190	0.96	1.04	0.96

Efficient Numerical Manipulation

In these item types, students were asked to look at an equation (i.e., 7=3+4 or 7+6=6) and determine whether it was true or false. Twenty different true or false statement items were used across the different grade levels and the different versions of the assessments. The items varied widely on the intended skill to be assessed (i.e., equivalence, arithmetic skills, commutative

property, property of zero, relational thinking strategies, etc.). The item statistics of difficulty and discrimination from both CTT and IRT for the meaning of the equal sign items are summarized in Tables 16a and 16b.

It appears the majority of students even beginning in 1st grade understood when traditional number sentences involving simple arithmetic were true (i.e., 3+4=7) or false (i.e., 7+6=6). They also therefore appeared to understand the meaning of the words 'true' and 'false'. Students appeared to struggle more when the problem was nontraditional even when the arithmetic difficulty remained the same (i.e., 8=5+13 or 7=7). By 4th and 5th grade, however, nearly all students had mastered both of these nontraditional formats.

Some of the true or false items, like the equivalence items designed to measure relational thinking skills, were designed to measure relational thinking skills (i.e., efficient numerical manipulation). The values were designed to be large enough to be very cumbersome (though possible) to calculate manually, yet much easier if calculated through using simplifying / efficient techniques (i.e., 15+7-6=15+1 can be solved quickly by recognizing that 7-6=1 so 15+1=16). Students across the grades struggled more with these types of items than the simpler true/false statements described previously. While 91% of 1st grade students could agree that 3+4=7 was true, for example, only 21% could say that 15+7-6=15+1 was true.

Finally, students were asked to utilize knowledge concerning the order of operations when evaluating items such as 5+(3x6)=23 and 5+(3x6)=48 (i.e., if students understood how to interpret the parentheses or not). Students solved both of these items at similar rates, and had mastered it in 4th and 5th grades with p values of 0.84 to 0.95.

There was a statistically significant difference by grade for the item 8+5=13 (F (4, 855) = 16.453, $p < .001$), for the item 7=7 (F (4, 853) = 35.250, $p < .001$), for the item 7=2+5 (F (4,

860) = 17.917, $p < .001$), and for the item 6=6 (F (4, 857) = 26.158, $p < .001$). Post-hoc Tukey analyses revealed that for 8=5+13, 1st grade students significantly underperformed 4th and 5th grade students while 2nd grade students significantly underperformed 3rd, 4th, and 5th grade students and 3rd grade students significantly underperformed 5th grade students. This item was the only item where 1st grade students actually performed better than 2nd grade students, although not by a significant amount ($p > .05$). For 7=7 and 6=6, 1st grade students significantly underperformed all other grades, while 2nd grade students underperformed 4th and 5th grade students and 3rd grade students underperformed 5th grade students, while for 7=2+5 the only significant difference was in 1st grade students underperforming all other grades. Additionally, there was a statistically significant difference by grade for the item 9-9=9 (F (4, 860) = 16.284, p <.001). Post-hoc Tukey analyses revealed that for this item, 1st grade students significantly underperformed all other grades, with no other significant differences.

Discrimination indices for efficient numerical manipulation items varied widely across the items, however adjusted item-total correlations were largely positive and IRT parameters were largely close to 1.00 (i.e., indicating an average level of discrimination for that difficulty level). Like open number sentences, it did appear that the discrimination capacity for many of these items was affected by a ceiling effect that occurs when the majority of the students solve the item correctly. These items may have also largely been less discriminatory because they were the only items affected by a guessing effect, which was quite high at 50%. It is important, however, to include easier items even on assessments for older students to ensure that they understand the concept and that you can reliably interpret their answers to more difficult items. Because of this, all items, even with less discrimination power, will be retained.

Table 16a.

Item Statistics for True / False Statements

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
3+4=7						
1st Grade	0.91	-3.65	0.163	1.07	0.84	0.96
2nd Grade	0.93	-3.02	0.177	0.94	0.99	1.75
3rd Grade	0.98	-4.01	0.094	1.02	0.88	0.94
7+6=6						
1st Grade	0.87	-2.99	0.326	1.10	0.88	0.70
2nd Grade	0.95	-3.66	0.255	1.03	0.94	0.77
3rd Grade	0.98	-3.58	0.298	1.05	0.98	0.36
8=5+13						
1st Grade	0.63	-1.25	0.185	0.54	1.20	1.63
2nd Grade	0.58	-0.17	0.103	0.29	1.31	1.40
3rd Grade	0.74	-0.57	0.129	0.39	1.43	1.55
4th Grade	0.83	-1.18	0.278	0.79	1.15	1.62
5th Grade	0.90	-1.03	0.324	0.95	1.06	1.00
7=2+5						
1st Grade	0.71	-1.73	0.346	0.98	1.00	1.22
2nd Grade	0.91	-2.83	0.073	0.95	1.03	1.22
3rd Grade	0.92	-2.16	0.344	1.00	1.02	0.80
4th Grade	0.92	-2.00	0.390	1.14	0.83	0.43
5th Grade	0.96	-2.43	0.175	0.98	1.01	1.64
7=7						
1st Grade	0.52	-0.63	0.327	0.92	1.04	1.00
2nd Grade	0.70	-0.86	0.447	1.03	1.00	0.90
3rd Grade	0.79	-0.93	0.411	1.13	0.91	0.73
4th Grade	0.92	-2.22	0.306	0.96	1.04	1.04
5th Grade	0.94	-1.70	0.248	0.97	1.05	1.01
6=6						
1st Grade	0.57	-0.91	0.417	1.26	0.91	0.81
2nd Grade	0.72	-1.19	0.177	0.75	1.15	1.21
3rd Grade	0.80	-0.92	0.399	1.07	0.96	0.83
4th Grade	0.89	-1.61	0.235	0.99	1.01	0.96
5th Grade	0.95	-2.12	0.209	1.03	1.02	0.52
4+0=4						
1st Grade	0.84	-2.74	0.085	0.97	1.04	0.94
2nd Grade	0.75	-1.16	0.262	0.81	1.14	1.13
4+0=0						
3rd Grade	0.86	-1.47	0.320	0.93	1.04	1.15
4th Grade	0.92	-1.93	0.195	0.96	1.00	1.78
5th Grade	0.93	-1.50	0.303	1.02	0.99	0.77

Table 16b.

Item Statistics for True / False Statements

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
9-9=9						
1st Grade	0.81	-2.43	0.360	1.22	0.81	0.68
2nd Grade	0.94	-3.36	0.323	1.04	0.93	0.84
3rd Grade	0.99	-4.45	0.236	1.12	0.76	0.11
4th Grade	0.95	-2.93	0.363	1.09	0.89	0.39
5th Grade	0.98	-3.40	0.000	1.10	0.77	0.31
5+3=5-3						
1st Grade	0.55	-0.79	0.383	0.84	1.07	1.04
2nd Grade	0.71	-0.90	0.285	0.80	1.13	1.07
3rd Grade	0.85	-1.46	0.337	1.09	0.95	0.67
5+3=3+5						
1st Grade	0.44	-0.20	0.320	0.86	1.06	1.04
2nd Grade	0.51	0.02	0.390	0.65	1.15	1.21
5x3=3x5						
3rd Grade	0.88	-1.65	0.248	0.90	1.08	1.49
4th Grade	0.90	-1.62	0.409	1.10	0.89	0.61
5th Grade	0.93	-1.69	0.424	1.11	0.93	0.38
15+7-6=15+1						
1st Grade	0.21	1.21	0.239	0.75	1.21	1.26
5th Grade	0.70	0.53	0.324	1.01	1.00	0.97
22+9-8=22+1						
1st Grade	0.25	0.94	0.305	0.54	1.30	1.79
2nd Grade	0.43	1.79	0.322	0.55	1.21	1.31
99+57-47=99+10						
2nd Grade	0.29	1.76	0.399	0.75	1.16	1.72
3rd Grade	0.42	1.19	0.254	0.66	1.14	1.25
99+99=100+100-2						
3rd Grade	0.42	1.21	0.079	0.72	1.09	1.26
4th Grade	0.53	0.94	0.349	0.64	1.14	1.21
(9x57)+57=10x57						
4th Grade	0.50	0.94	0.310	0.24	1.33	1.76
5th Grade	0.65	0.85	0.171	0.56	1.21	1.34
89+44=87+46						
4th Grade	0.74	-0.48	0.345	1.01	0.99	0.98
5th Grade	0.82	-0.26	0.441	1.10	0.91	0.83
5+(3x6)=23						
4th Grade	0.84	-1.25	0.180	0.82	1.14	2.01
5th Grade	0.93	-1.59	0.260	1.03	1.01	0.64
5+(3x6)=48						
4th Grade	0.91	-1.85	0.443	1.09	0.89	0.60
5th Grade	0.95	-2.12	-0.064	1.10	0.76	0.71

Figural Patterns

Students were asked to evaluate figural patterns, determine the shape that would come next, and then generalize to the shape that would come in some given placement. Six different figural pattern items were used across the different grade levels and the different versions of the assessments. The items varied on whether the item was repeating or growing, and if it was growing, whether it was growing linearly or nonlinearly. Repeating items for 1st and 2nd grade students included a far generalization component (i.e., what shape would come 18th in the pattern) while the linear and nonlinear items for 3rd, 4th, and 5th grade students included a far generalization component (i.e., how many blocks in the 10th figure in the pattern). The item statistics of difficulty and discrimination from both CTT and IRT for the figural pattern items are summarized in Tables 17a, 17b, and 17c. It appears young students may have already mastered repeating patterns and are solving them at very high rates. Additionally, nearly half of 1st grade students and more than half of 2nd grade students could generalize to the shape that would come 18th in the pattern. As the table demonstrates, students in the early grades struggled much more with linear figural patterns than they did with repeating patterns. By 3rd grade, over half of students were solving such items correctly, however.

There was a significant difference by grade for both the 4, 7, 10, ___ figural pattern (F (4, 829) = 33.021, $p < .001$) and for the 5, 8, 11, ___ figural pattern (F (4, 851) = 33.677, $p < .001$). Post-hoc Tukey analyses revealed that students in 1st and 2nd grade significantly underperformed all other grades ($p < .01$).

Generalizing to the 10th figure remained difficult for all students in all grades. There was not a statistically significant difference for generalizing by grade on the 4,7, 10, ___ figural

pattern ($p > .05$) although there was a significant difference for generalizing by grade on the 5, 8, 11, ___ figural pattern (F (2, 525) = 8.262, $p < .001$).

Surprisingly, students performed similarly on both the linear and the nonlinear patterns. In some cases students actually performed better on the nonlinear than they did on the linear figure. There was a statistically significant difference by grade for both the 1, 4, 9, ___ figural nonlinear pattern (F (4, 800) = 47.587, $p < .001$) and for the 3, 6, 10, ___ figural nonlinear pattern (F (4, 831) = 32.366), $p < .001$). Post-hoc Tukey analyses revealed that while 1st and 2nd grade students did not differ significantly from each other ($p > .05$), they significantly underperformed every other grade ($p < .001$). In addition, 3rd grade students significantly underperformed 4th and 5th grade students for 1, 4, 9, ___ ($p < .001$) and underperformed only 5th grade students ($p < .01$) for the 3, 6, 10, ___ item.

Generalizing to the 10th figure remained a struggle again throughout all grades, however the 1, 4, 9 (i.e., the pattern of squares: 1^2, 2^2, 3^2, etc.) was much easier for students to grasp when generalizing to the 10th figure than the other nonlinear pattern and both linear patterns. There was a statistically significant difference by grade for both the 1, 4, 9, ___ figural nonlinear pattern generalizing to the 10th (F (2, 494) = 14.739, $p < .001$) and for the 3, 6, 10, ___ figural nonlinear pattern generalizing to the 10th (F (2, 510) = 5.079), $p < .01$).

The figural pattern items appear to be fairly good discriminators. Although several items had IRT discrimination parameters less than 1.00 (i.e., indicating less than average discrimination for an item of that difficulty), all of the adjusted item-total correlations were positive and many were statistically significant ($p < .001$).

Table 17a.

Item Statistics for Repeating Figural Patterns

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
Shape Repeating ABCABC						
1st Grade	0.89	-3.37	0.356	1.05	0.98	0.65
2nd Grade	0.91	-2.92	0.253	0.94	1.04	1.53
Shape Repeating ABCABC Far 18th						
1st Grade	0.45	-0.20	0.275	0.79	1.09	1.07
2nd Grade	0.56	-0.21	0.386	0.79	1.10	1.10
Shape Repeating AABAAB						
1st Grade	0.96	-4.47	0.198	0.94	1.17	0.79
2nd Grade	0.95	-3.46	0.216	1.12	0.82	0.38
Shape Repeating AABAAB Far 18th						
1st Grade	0.47	-0.30	0.294	0.55	1.17	1.26
2nd Grade	0.58	-0.18	0.326	0.78	1.07	1.28

Table 17b.

Item Statistics for Linear Figural Patterns

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
Linear Figure 4, 7, 10, ___						
1st Grade	0.27	0.80	0.376	1.04	0.99	0.88
2nd Grade	0.40	0.86	0.397	0.77	1.15	1.06
3rd Grade	0.60	0.26	0.239	0.87	1.06	1.09
4th Grade	0.71	-0.02	0.414	0.86	1.09	1.11
5th Grade	0.75	0.03	0.255	0.92	1.06	1.01
Linear Figure 4, 7, 10, ___ Far 10th						
3rd Grade	0.17	2.82	0.244	0.98	1.02	0.92
4th Grade	0.21	2.90	0.351	0.99	1.01	0.93
5th Grade	0.23	3.05	0.302	1.06	0.95	0.93
Linear Figure 5, 8, 11, ___						
1st Grade	0.23	1.09	0.064	0.68	1.21	1.56
2nd Grade	0.37	0.88	0.262	0.69	1.16	1.30
3rd Grade	0.56	0.52	0.303	0.66	1.12	1.23
4th Grade	0.56	0.51	0.322	0.77	1.11	1.14
5th Grade	0.76	0.30	0.295	0.79	1.17	1.19
Linear Figure 5, 8, 11, ___ Far 10th						
3rd Grade	0.15	3.11	0.357	0.98	1.03	1.13
4th Grade	0.12	3.48	0.264	0.87	1.16	1.22
5th Grade	0.28	2.91	0.191	0.98	1.00	1.05

Table 17c.

Item Statistics for Nonlinear Figural Patterns

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
Nonlinear Figure 1, 4, 9, ___						
1st Grade	0.25	0.99	0.352	0.71	1.19	2.70
2nd Grade	0.30	1.51	0.215	0.54	1.31	1.51
3rd Grade	0.48	0.93	0.133	0.34	1.30	1.39
4th Grade	0.72	-0.42	0.216	0.80	1.12	1.25
5th Grade	0.79	0.13	0.084	0.89	1.11	1.06
Nonlinear Figure 1, 4, 9, ___ Far 10th						
3rd Grade	0.13	3.28	0.314	1.04	0.94	1.15
4th Grade	0.29	2.10	0.284	0.85	1.10	1.08
5th Grade	0.39	2.29	0.332	1.30	0.90	0.84
Nonlinear Figure 3, 6, 10, ___						
1st Grade	0.30	0.66	0.374	0.76	1.15	1.17
2nd Grade	0.40	0.71	0.475	0.77	1.13	1.12
3rd Grade	0.61	0.31	0.340	0.77	1.11	1.09
4th Grade	0.70	0.08	0.305	0.75	1.17	1.10
5th Grade	0.77	-0.07	0.344	1.00	1.01	0.91
Nonlinear Figure 3, 6, 10, ___ Far 10th						
3rd Grade	0.03	5.51	0.345	1.13	0.76	0.20
4th Grade	0.07	4.47	0.235	1.03	0.98	0.66
5th Grade	0.12	4.09	0.303	1.13	0.86	0.62

Generalization

Six different generalization items were used across the different grade levels and the different versions of the assessments. The items varied whether the item measured the additive identify property (i.e., a+0=a), the multiplicative property of zero (i.e., ax0=a), the additive inverse property (i.e., a-a=0), the commutative property of both addition and multiplication (i.e., a+b=b+a), or the multiplicative identity property (i.e., ax1=a). Students were asked to first circle whether the property was true or false. They were then asked to explain their answer choice. Explanations were rated on a 2 point scale: 0 indicating that the student does not understand (i.e., no understanding), 1 indicating that the student is on his or her way to a full coherent understanding of the property (i.e., partial understanding), and 2 indicating that the student fully

understands the property (i.e., full understanding). The item statistics of difficulty and discrimination from both CTT and IRT for the generalization items are summarized in Table 18.

It appears that students' ability to understand generalizations and their abilities to coherently explain the property increases as they progress through the grade levels. It appears the majority of students in grades 3-5 could successfully say whether a-a=0, a+0=a, and ax0=a were true or false. There was a statistically significant difference for determining whether a-a=0 was true or false (F (2, 462) = 25.644, $p < .001$) and for explaining why a-a=0 was true (F (2, 455) = 26.674, $p < .001$). The same was true for determining whether a+0=a was true or false (F (2, 530) = 21.467, $p < .001$) and for explaining why a+0=a was true (F (2, 526) = 44.724, $p < .001$). The same was also true for determining whether ax0=a was true or false (F (1, 362) = 9.240, $p < .01$), but there was no significant difference by grade when explaining why ax0=a was false. This is likely the case because only 4th and 5th grade students answered this item and there were not large differences between 4th and 5th grade students across the items.

The commutative property appeared to be more difficulty for students than the properties of zero. Although the majority of students in grades 3-5 were able to determine that a+b=b+a was true, for instance, with p values of 0.53 to 0.71, very few students were able to coherently describe why that was the case (i.e., 3% to 26% of students, depending on the grade). Scores were even lower for the problem a-b=b-a, with the majority of students not answering correctly with "not always true" until 5th grade. There was a statistically significant difference for determining whether a+b=b+a was true or false (F (3, 674) = 18.807, $p < .001$) and for explaining why a+b=b+a was true (F (3, 664) = 27.683, $p < .001$). The same is true for determining whether a-b=b-a was true or false (F (2, 540) = 16.660, $p < .001$) and for explaining why a-b=b-a was false (F (2, 534) = 19.570, $p < .001$).

The generalization items appeared to be very good discriminators. All of the adjusted item-total correlations were positive and nearly all were statistically significant ($p < .001$). The majority of the IRT discrimination parameters hovered close to 1.00 (i.e., indicating an average discrimination capacity for that item difficulty level). All items will therefore be retained.

Table 18.

Item Statistics for Generalization Items

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
a-a=0 true / false						
1st Grade	0.28	0.77	0.544	1.03	0.97	0.99
2nd Grade	0.41	0.76	0.529	1.05	0.99	0.89
3rd Grade	0.65	-0.01	0.404	0.91	1.05	1.06
a-a=0 explain						
1st Grade	0.06	2.11	0.474	1.04	0.98	0.67
2nd Grade	0.16	2.75	0.563	1.05	0.96	0.81
3rd Grade	0.29	1.97	0.356	1.02	1.00	0.91
a+0=a true / false						
1st Grade	0.36	0.27	0.094	0.44	1.23	1.46
3rd Grade	0.66	-0.05	0.284	1.00	1.00	0.96
4th Grade	0.65	0.39	0.402	1.18	0.92	0.81
a+0=a explain						
1st Grade	0.05	3.10	0.323	0.98	1.11	0.86
3rd Grade	0.28	1.98	0.362	0.92	1.09	1.15
4th Grade	0.39	1.71	0.435	0.95	1.07	0.94
ax0=a true / false						
4th Grade	0.57	0.77	0.327	0.83	1.08	1.06
5th Grade	0.72	0.44	0.355	0.94	1.02	1.20
ax0=a explain						
4th Grade	0.50	1.11	0.379	0.84	1.15	1.14
5th Grade	0.61	1.17	0.295	0.64	1.30	1.31
a+b=b+a true / false						
2nd Grade	0.31	1.21	0.315	0.57	1.25	1.53
3rd Grade	0.53	0.61	0.507	1.16	0.94	0.87
4th Grade	0.57	0.49	0.504	1.33	0.85	0.75
5th Grade	0.71	0.51	0.510	1.31	0.81	0.71
a+b=b+a explain						
2nd Grade	0.03	3.66	0.369	1.03	1.01	0.60
3rd Grade	0.10	2.67	0.369	1.13	0.90	0.66
4th Grade	0.16	4.25	0.371	1.13	0.90	0.78
5th Grade	0.26	3.24	0.373	1.01	1.02	0.92
a-b=b-a true / false						
3rd Grade	0.33	1.77	0.076	0.67	1.13	1.43
4th Grade	0.48	1.11	0.191	0.84	1.07	1.38
5th Grade	0.63	0.80	0.471	1.18	0.93	0.80
a-b=b-a explain						
3rd Grade	0.05	3.67	0.455	1.14	0.85	0.45
4th Grade	0.12	2.67	0.483	1.33	0.75	0.54
5th Grade	0.20	3.96	0.489	1.21	0.85	0.74
ax1=a true / false						
5th Grade	0.89	-0.83	0.435	1.17	0.82	0.46
ax1=a explain						
5th Grade	0.64	1.18	0.271	0.79	1.18	1.12

Construct-Related Evidence for Validity

The average standard scores were calculated for the two versions of the assessment. These are displayed below in Table 19. As students progressed through the grade levels, so did their average standard score. This indicates that as students progressed through the grade levels they also improved on their algebraic thinking skills. Because this result is expected and desired due to student growth in mathematical skills across the grades, this is construct-related evidence that this assessment of algebraic thinking skills actually measures algebraic thinking skills.

Table 19.

Average Standard Scores for Each Version of the Assessment

	Version A Mean	*Version B Mean*
1st Grade	45	43
2nd Grade	45	46
3rd Grade	51	52
4th Grade	53	54
5th Grade	53	54

Internal Structure Validity Evidence

As discussed previously, reliability is a necessary condition for any validation process. General consensus requires an acceptable reliability coefficient to be at 0.80 if making decisions about individuals based on scores (Webb, Shavelson, & Haertel, 2006), however this number may be acceptable at 0.70 based on the intention of the assessment (i.e., this assessment is intended to be diagnostic in nature, and will not be used to make high-stakes decisions). Cronbach's alpha coefficients were therefore calculated for each version of the assessment. If this number is high (i.e., greater than 0.70), it is likely that the assessment tool is internally reliable and therefore may be evidence of internal structure validity. It must be stated up-front

that because of their nature (and potential for ceiling or floor effects), diagnostic assessments will not always meet such reliability requirements. All coefficients, however, met the requirement of being at 0.70 and above. Further, of the 20 different versions of the assessment, only six versions had reliability coefficients less than 0.80. These results are displayed below in Table 20.

Table 20.

Cronbach's Alpha Reliability Coefficients for All Versions of the Assessment

	Number of Items	Version A	Version B	Version AB	Version BA
1st Grade	25	**0.85 (*n* = 74)**	0.73 (*n* = 80)	0.79 (*n* = 87)	**0.84 (*n* = 71)**
2nd Grade	26	**0.86 (*n* = 70)**	**0.82 (*n* = 65)**	**0.82 (*n* = 62)**	0.75 (*n* = 66)
3rd Grade	28	**0.83 (*n* = 68)**	0.77 (*n* = 72)	**0.83 (*n* = 69)**	**0.82 (*n* = 70)**
4th Grade	27	**0.82 (*n* = 87)**	**0.80 (*n* = 86)**	**0.86 (*n* = 88)**	**0.82 (*n* = 84)**
5th Grade	27	**0.81 (*n* = 78)**	0.71 (*n* = 85)	0.73 (*n* = 78)	**0.80 (*n* = 77)**

When analyzing the reliability coefficient, it was important to determine if there were any items that might be problematic. Items such as these would increase the reliability coefficient if deleted. A list of potentially problematic is included below in Table 21. If the reliability estimate increased by more than one hundredth of a point (i.e., from .73 to .74) upon deletion then it was included below. The last problem on every test was the nonlinear figure item; therefore these items had the most missing data and therefore are likely to be the least reliable because of missing data, not necessarily because of content. As the table demonstrates, however, the current alpha for all assessments is above 0.70 currently, which is considered acceptable for a diagnostic assessment. The only item that drastically increases the alpha if deleted is Nonlinear

Figure B on assessment 1B, however this is likely because of missing data. Therefore it has been determined that for a diagnostic assessment, all items are acceptable and will be retained.

Table 21.

Potentially Problematic Items Based on Cronbach's Alpha

Item	Grade	Assessment	Item Difficulty	Current Alpha	Alpha if Deleted
Nonlinear Figure B	1	1AB	0.14	0.79	0.82
Nonlinear Figure B	1	1B	0.13	0.73	0.82
Nonlinear Figure A	2	2AB	0.15	0.82	0.86
Nonlinear Figure A	2	2B	0.15	0.82	0.88
Linear Figure B	2	2BA	0.23	0.75	0.78
True or False: 99+57-47=99+10	2	2BA	0.45	0.75	0.80
Meaning of the Equal Sign	3	3B	0.38	0.77	0.78
Meaning of the Equal Sign	3	3A	0.28	0.83	0.84
Meaning of the Equal Sign	5	5B	0.42	0.71	0.73
Meaning of the Equal Sign	5	5AB	0.41	0.73	0.76

Correlational statistics for student grade level and the different algebraic thinking tasks on the assessment are displayed below in Table 22. By examining this table one can see that all of the tasks were statistically significantly positively correlated with each other. These results demonstrate initial internal structure validity evidence: indicating that the dimension strands act similarly to each other and that there may indeed be unidimensionality across the construct.

Table 22.

Correlation Statistics of Algebraic Thinking Dimensions

	1.	2.	3.	4.	5.	6.	7.	8.
1. Open Number Sentences	-							
2. Equivalence	.244*	-						
3. Work with Variables	.292*	.632*	-					
4. Efficient Numerical Calculation	.296*	.489*	.516*	-				
5. Generalization	.243*	.388*	.541*	.430*	-			
6. Numerical Patterns	.326*	.408*	.402*	.332*	.293*	-		
7. Figural Patterns	.176*	.238*	.208*	.199*	.123*	.314*	-	
8. Figural Patterns Generalize	.164*	.169*	.148*	.130*	.156*	.225*	.324*	-

Note. N=1,745, *p<.01

Interrater Reliability

One teacher unfamiliar with the assessment randomly scored approximately 6% (n = 109) of the assessments. Short-answer items only were scored twice; answer only items were not scored again. This teacher was given only the scoring rules developed by the researcher (see Appendix C) as this will be the likely scenario in the "real world". Percent exact agreement was calculated by adding the number of scores matches and dividing by the number of score matches plus the number of non-matches (i.e., dividing by the total number of assessments graded).

Overall percent exact agreement was 94%. Percent exact agreement for 1st grade, 2nd grade, 3rd grade, 4th grade, and 5th grade assessment items were 92%, 98%, 94%, 93%, and 96%, respectively. Overall percent exact agreement plus or minus 1 was 100%, however this is expected given that the assessment items are only scored on a 3-point rubric.

Given that the inter-rater reliability coefficient estimates were fairly high (i.e., greater than 80%), it is therefore likely that even untrained raters are reliable in scoring and using the scoring guides and rubrics.

Alternate Forms Equating

Two forms of the assessment at each grade level were developed to allow assessing throughout the school year without a 'testing effect'. In addition, these versions were combined in two different ways to enable alternate equating. The means and standard deviations of the different versions may be compared below in Table 23. As one can see, there are no statistically significant differences ($p > .05$) between Version A and Version B means and standard deviations at any grade level. It can therefore be determined that the two forms are likely, indeed, alternate forms of one another and may be used interchangeably.

Table 23.

Total Scores and Standard Deviations for Each Version of the Assessment and Overall

	Total		Version A		Version B		Version AB		Version BA	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1st Grade	9.77	4.28	10.77	4.39	10.23	4.30	8.70	4.13	10.03	4.17
2nd Grade	12.73	4.94	12.42	4.98	12.92	5.17	13.27	5.01	12.33	4.61
3rd Grade	15.24	4.77	15.17	4.97	15.13	4.38	14.71	4.69	15.93	5.03
4th Grade	15.91	4.92	16.36	4.54	16.25	4.81	14.36	4.85	16.75	5.17
5th Grade	18.37	4.51	17.20	4.83	18.17	4.04	20.91	4.00	17.36	4.17

Each assessment was then vertically equated to place each assessment on the same scale. With this method, student scores can be tracked over time, across grades, and show student growth. Equating of the diagnostic assessment of algebraic thinking skills for grades 1-5 used the Partial Credit Model (PCM). The equating was performed for algebraic thinking skills to link across the grades: grades 1, 2, 3, 4, and 5.

A comparison between the anchor items from 1st to 2nd grade, 2nd to 3rd grade, 3rd to 4th grade, and 4th to 5th was made. In performing each equating a separate analysis of the anchor

items was performed. This analysis examined the stability of the anchor item difficulties estimated in the current grade and their bank values. The first step in this analysis is to calibrate the 1st grade test form without fixing any of the anchor items. This allows all items to be estimated. The mean of the anchor item difficulties from the bank is used and a constant is added to the 2nd grade item difficulties so they have the same mean difficulty. The absolute difference between the item's bank value and "equated" 2nd grade value is computed and the item with the largest absolute difference in the PCM difficulty values greater than or equal to 0.3 is removed from consideration as an anchor item. This procedure is repeated until only those items with absolute differences less than 0.3 remain. For some tests this procedure was repeated several times. In some cases the comparisons may produce additional anchor items demonstrating instability in the item difficulties and they are not used as anchors. Rarely does this procedure require more than several rounds of comparisons.

Once the set of anchor items was determined, the data sets were calibrated again using the PCM and fixing the anchor item difficulties and step values to their bank values. A linear transformation was then used to move the equated thetas to scale scores. The formula used was: 5 * Theta + 50 = Scale Score. Tables 24a, 24b, 24c, 24d, and 24e displays these results for 1st, 2nd, 3rd, 4th, and 5th grades, respectively.

Table 24a.

Raw Score to Scale Score Relationship for the 1st Grade Assessment

1st Grade Form A				1st Grade Form B			
Raw Score	Theta	S.E.	Scale Score	Raw Score	Theta	S.E.	Scale Score
0	-6.60	1.88	17	0	-6.76	1.91	16
1	-5.24	1.11	24	1	-5.34	1.14	23
2	-4.32	0.86	28	2	-4.38	0.87	28
3	-3.68	0.75	32	3	-3.74	0.74	31
4	-3.16	0.69	34	4	-3.25	0.67	34
5	-2.72	0.64	36	5	-2.83	0.62	36
6	-2.33	0.61	38	6	-2.47	0.58	38
7	-1.98	0.58	40	7	-2.15	0.56	39
8	-1.66	0.56	42	8	-1.85	0.54	41
9	-1.36	0.54	43	9	-1.57	0.52	42
10	-1.08	0.52	45	10	-1.31	0.51	43
11	-0.81	0.51	46	11	-1.06	0.5	45
12	-0.55	0.51	47	12	-0.81	0.49	46
13	-0.30	0.5	49	13	-0.57	0.49	47
14	-0.05	0.50	50	14	-0.34	0.48	48
15	0.20	0.50	51	15	-0.11	0.48	49
16	0.45	0.50	52	16	0.12	0.48	51
17	0.70	0.50	54	17	0.35	0.48	52
18	0.95	0.51	55	18	0.59	0.49	53
19	1.22	0.52	56	19	0.83	0.50	54
20	1.50	0.54	58	20	1.08	0.51	55
21	1.80	0.56	59	21	1.35	0.53	57
22	2.13	0.59	61	22	1.65	0.56	58
23	2.51	0.64	63	23	1.98	0.60	60
24	2.98	0.73	65	24	2.38	0.67	62
25	3.60	0.87	68	25	2.90	0.78	65
26	4.59	1.16	73	26	3.71	1.05	69
27	6.07	1.93	80	27	4.98	1.85	75

Table 24b.

Raw Score to Scale Score Relationship for the 2nd Grade Assessment

2nd Grade Form A				2nd Grade Form B			
Raw Score	Theta	S.E.	Scale Score	Raw Score	Theta	S.E.	Scale Score
0	-6.62	1.88	17	0	-6.99	1.87	15
1	-5.28	1.09	24	1	-5.67	1.09	22
2	-4.39	0.83	28	2	-4.78	0.83	26
3	-3.79	0.73	31	3	-4.18	0.74	29
4	-3.31	0.66	33	4	-3.68	0.68	32
5	-2.91	0.61	35	5	-3.24	0.65	34
6	-2.56	0.58	37	6	-2.84	0.61	36
7	-2.24	0.55	39	7	-2.48	0.59	38
8	-1.94	0.53	40	8	-2.15	0.56	39
9	-1.67	0.51	42	9	-1.84	0.54	41
10	-1.41	0.5	43	10	-1.56	0.53	42
11	-1.17	0.49	44	11	-1.29	0.51	44
12	-0.93	0.49	45	12	-1.03	0.5	45
13	-0.69	0.48	47	13	-0.78	0.49	46
14	-0.47	0.48	48	14	-0.54	0.49	47
15	-0.24	0.48	49	15	-0.31	0.48	48
16	-0.01	0.48	50	16	-0.08	0.48	50
17	0.21	0.48	51	17	0.15	0.47	51
18	0.45	0.49	52	18	0.37	0.47	52
19	0.69	0.49	53	19	0.59	0.47	53
20	0.94	0.51	55	20	0.82	0.48	54
21	1.21	0.53	56	21	1.06	0.49	55
22	1.5	0.56	58	22	1.31	0.51	57
23	1.83	0.59	59	23	1.58	0.54	58
24	2.21	0.64	61	24	1.89	0.58	59
25	2.66	0.71	63	25	2.25	0.64	61
26	3.24	0.82	66	26	2.73	0.76	64
27	4.09	1.07	70	27	3.5	1.03	68
28	5.38	1.85	77	28	4.75	1.84	74

Table 24c.

Raw Score to Scale Score Relationship for the 3rd Grade Assessment

3rd Grade Form A				3rd Grade Form B			
Raw Score	Theta	S.E.	Scale Score	Raw Score	Theta	S.E.	Scale Score
0	-7.08	1.92	15	0	-5.97	1.88	20
1	-5.62	1.18	22	1	-4.63	1.1	27
2	-4.54	0.94	27	2	-3.73	0.83	31
3	-3.75	0.83	31	3	-3.14	0.72	34
4	-3.13	0.75	34	4	-2.68	0.65	37
5	-2.62	0.68	37	5	-2.29	0.6	39
6	-2.19	0.62	39	6	-1.95	0.57	40
7	-1.83	0.58	41	7	-1.64	0.54	42
8	-1.51	0.55	42	8	-1.35	0.53	43
9	-1.22	0.52	44	9	-1.08	0.51	45
10	-0.96	0.51	45	10	-0.82	0.50	46
11	-0.71	0.49	46	11	-0.57	0.50	47
12	-0.48	0.48	48	12	-0.33	0.49	48
13	-0.25	0.47	49	13	-0.09	0.49	50
14	-0.03	0.47	50	14	0.15	0.48	51
15	0.18	0.46	51	15	0.38	0.48	52
16	0.40	0.46	52	16	0.62	0.48	53
17	0.61	0.46	53	17	0.85	0.48	54
18	0.82	0.46	54	18	1.08	0.49	55
19	1.04	0.47	55	19	1.32	0.49	57
20	1.26	0.48	56	20	1.56	0.50	58
21	1.50	0.49	58	21	1.82	0.51	59
22	1.74	0.50	59	22	2.10	0.54	61
23	2.00	0.52	60	23	2.40	0.56	62
24	2.28	0.54	61	24	2.74	0.60	64
25	2.58	0.57	63	25	3.13	0.66	66
26	2.93	0.61	65	26	3.61	0.73	68
27	3.33	0.67	67	27	4.22	0.83	71
28	3.84	0.76	69	28	5.01	0.95	75
29	4.53	0.93	73	29	6.11	1.18	81
30	5.70	1.26	79	30	7.58	1.92	88
31	7.35	2.00	87				

Table 24d.

Raw Score to Scale Score Relationship for the 4th Grade Assessment

4th Grade Form A				4th Grade Form B			
Raw Score	Theta	S.E.	Scale Score	Raw Score	Theta	S.E.	Scale Score
0	-6.19	1.91	19	0	-6.06	1.90	20
1	-4.77	1.14	26	1	-4.67	1.12	27
2	-3.82	0.86	31	2	-3.74	0.85	31
3	-3.19	0.73	34	3	-3.12	0.73	34
4	-2.71	0.66	36	4	-2.65	0.66	37
5	-2.31	0.61	38	5	-2.24	0.62	39
6	-1.96	0.58	40	6	-1.88	0.59	41
7	-1.64	0.55	42	7	-1.55	0.57	42
8	-1.35	0.53	43	8	-1.24	0.55	44
9	-1.07	0.52	45	9	-0.95	0.53	45
10	-0.81	0.50	46	10	-0.67	0.52	47
11	-0.56	0.50	47	11	-0.41	0.51	48
12	-0.32	0.49	48	12	-0.16	0.49	49
13	-0.09	0.48	50	13	0.07	0.48	50
14	0.14	0.48	51	14	0.30	0.47	52
15	0.37	0.47	52	15	0.52	0.46	53
16	0.59	0.47	53	16	0.72	0.45	54
17	0.82	0.47	54	17	0.93	0.45	55
18	1.05	0.48	55	18	1.13	0.45	56
19	1.28	0.48	56	19	1.34	0.45	57
20	1.51	0.49	58	20	1.55	0.46	58
21	1.76	0.51	59	21	1.76	0.47	59
22	2.03	0.53	60	22	1.99	0.49	60
23	2.32	0.55	62	23	2.25	0.51	61
24	2.65	0.59	63	24	2.53	0.55	63
25	3.04	0.65	65	25	2.85	0.59	64
26	3.51	0.73	68	26	3.24	0.66	66
27	4.12	0.84	71	27	3.74	0.77	69
28	4.99	1.05	75	28	4.47	0.97	72
29	6.53	1.46	83	29	5.81	1.38	79
30	8.49	2.07	92	30	7.66	2.05	88

Table 24e.

Raw Score to Scale Score Relationship for the 5th Grade Assessment

5th Grade Form A				5th Grade Form B			
Raw Score	Theta	S.E.	Scale Score	Raw Score	Theta	S.E.	Scale Score
0	-5.52	1.86	22	0	-6.43	1.88	18
1	-4.22	1.07	29	1	-5.08	1.10	25
2	-3.38	0.80	33	2	-4.19	0.83	29
3	-2.84	0.69	36	3	-3.59	0.72	32
4	-2.42	0.62	38	4	-3.12	0.65	34
5	-2.06	0.58	40	5	-2.73	0.61	36
6	-1.75	0.54	41	6	-2.37	0.58	38
7	-1.47	0.52	43	7	-2.05	0.56	40
8	-1.21	0.50	44	8	-1.75	0.54	41
9	-0.97	0.48	45	9	-1.46	0.53	43
10	-0.74	0.47	46	10	-1.19	0.52	44
11	-0.52	0.46	47	11	-0.92	0.51	45
12	-0.31	0.45	48	12	-0.67	0.50	47
13	-0.11	0.45	49	13	-0.42	0.50	48
14	0.09	0.44	50	14	-0.17	0.49	49
15	0.28	0.44	51	15	0.06	0.48	50
16	0.47	0.44	52	16	0.3	0.48	52
17	0.67	0.44	53	17	0.53	0.48	53
18	0.86	0.44	54	18	0.76	0.48	54
19	1.06	0.45	55	19	0.98	0.48	55
20	1.27	0.46	56	20	1.22	0.48	56
21	1.50	0.48	58	21	1.46	0.49	57
22	1.74	0.50	59	22	1.7	0.51	59
23	2.01	0.54	60	23	1.97	0.52	60
24	2.32	0.58	62	24	2.26	0.55	61
25	2.69	0.64	63	25	2.57	0.58	63
26	3.15	0.72	66	26	2.94	0.63	65
27	3.74	0.82	69	27	3.37	0.70	67
28	4.51	0.94	73	28	3.94	0.82	70
29	5.59	1.17	78	29	4.83	1.10	74
30	7.05	1.92	85	30	6.18	1.89	81

Chapter 6: Results of the Singapore Sample and Comparisons with Washington State Sample

In this chapter I will describe the results obtained from the Singapore sample of students. In this sample 1,619 students from four different schools in the country of Singapore were administered the diagnostic assessment of algebraic thinking skills. This chapter will explore the results of several analyses conducted to begin validating the interpretation of the assessment scores. First, item analyses including Classical Test Theory analyses, Item Response Theory analyses, and discrimination indices will be conducted to better understand the construct. Second, reliability analyses including internal structure validity evidence (i.e., to determine that the items and dimension strands act similarly to each other) will be reported. Third, construct-related validity evidence in the form of how scores change as students progress through the grade levels will be reported. Fourth and finally, construct-related validity evidence in the form of comparisons of how students in Singapore performed compared to how students in the U.S. performed will be discussed.

Description of Singapore Sample

For this cross-cultural validation study, a sample of 1,619 students in grades 1-6 were selected from four 'neighborhood' schools in Singapore (i.e., average, not high-performing schools or schools with a selection process). Student participants included 248 1st grade students (15%), 224 2nd grade students (14%), 436 3rd grade students (27%), 324 4th grade students (20%), 297 5th grade students (18%), and 90 6th grade students (6%). Because 6th grade is not generally part of the elementary school in the U.S. and the sample of 6th grade students was very small, the 6th grade results were omitted from the subsequent analyses. Sample sizes for each item are displayed in Table 25. In addition, over 24 mathematics lessons were also observed and 14

teachers were interviewed to better understand how algebraic thinking skills are taught in Singapore. The sample of students from Singapore completed the assessments in summer of 2011, while the sample of students from the U.S. completed the assessments in spring of 2012.

There were several severe limitations associated with the Singapore sample that were the result of logistical issues when working in a different country. The most glaring is obviously the convenience sample. In addition, no spiraling of alternate forms occurred within the classroom, therefore there may be 'classroom' or 'school effects' on certain items because of this. In addition, because of this lack of spiraling, certain items have very small sample sizes or are missing entirely in certain grade levels. Because of these limitations, the results must be interpreted very cautiously.

Information regarding the sample sizes of students completing each item for each dimension is listed in Table 25.

Table 25.

Sample Sizes for all Items: Singapore

	1st Grade	2nd Grade	3rd Grade	4th Grade	5th Grade	All
Open Number Sentences						
8+___=15	195	224	111	288	258	1,076
___-3=12	195	224	111			530
___x8=56			324	36		360
___x6=36			111	288	258	657
Equivalence						
8+4=___+5	221		111	288	39	659
6+7=___+4	27	222	325	36	258	868
8+___=3+9	219		324			543
3x4=___x2			325	288		613
40/4=80/___				36	258	294
Work with Variables						
10-g=2	221	222				443
c+c+3=15	26		325	288	39	678
4+a+a=10		222	111	288		621
5xgxg=20			436	36	258	730
x+x+y=12; x-y=3					296	296
Numerical Patterns						
3, 5, 7, 9, 11,___	27	222				249
15, 21, ___, 33, 39	53		324	36		413
72, 69, 66, 63, 60, ___	193	222				415
35, 30, 25, ___, 15			111	288	258	657
2, 3, 5, 8, ___			325	288	297	910
50, 49, 47, 44, ___			111	36		147
Figural Patterns						
Linear Figure 4, 7, 10	205		314	36	39	594
Linear Figure 4, 7, 10 Far			303			303
Linear Figure 4, 7, 10 Far				35	39	74
Linear Figure 5, 8, 11	26	209	111	288	256	890
Linear Figure 5, 8, 11 Far		202	111			
Linear Figure 5, 8, 11 Far				284	255	539
Nonlinear Figure 3, 6, 10,	52	202	109	34	38	435
Nonlinear Figure 3, 6, 10 Far		178	108			
Nonlinear Figure 3, 6, 10 Far				31	37	68
Generalization						
a+0=a true / false	212	193	109			514
a+0=a explain	210	211	110			531
a-b=b-a true / false			318	287	257	862
a-b=b-a explain			316	287	257	860

Item Difficulty and Discrimination Results: Classical Test Theory and Item Response Theory Analyses

Both Classical Test Theory (CTT) and Item Response Theory (IRT) analyses were used to analyze all item statistics. Both the adjusted item-total correlations and the IRT discrimination parameter were used to measure discriminability of items. Examples of how to interpret such analyses was described previously. Discrimination was largely poorer for the Singapore sample than for the U.S. sample, however this may be due to the fact that Singapore students performed much better (i.e., encountered more ceiling effects) or due to the nature of the sample, which was less high-quality than that of the U.S. (i.e., items not spiraled within classrooms equally). Most of the items did have positive and statistically significant ($p < .001$) adjusted item-total correlations, however, and many exhibited IRT discrimination parameters greater than 1.00. Fit statistics were largely poorer for the Singapore sample as well. Out of all the items at all of the different grade levels, however, no items were found with an infit statistic greater than 2.0. Ten different items, however, were discovered with an outfit statistic greater than 2.0. These items were spread across the various dimensions and across the various grade levels, therefore indicating that multidimensionality is likely not an issue for the Singapore sample either. Again, because of the diagnostic assessment purpose, it was generally determined that all items will be retained for this assessment despite potentially low discrimination because of their uses diagnostically and for comparison.

Open Number Sentences: Singapore

The item statistics of difficulty and discrimination from both CTT and IRT for the open number sentence items for the Singapore sample are summarized in Table 26. Students had largely mastered all such items with the lowest p value being 0.70 across all items and grades.

There was a statistically significant difference by grade for the item 8+___=15 (F (4, 1,071) = 11.445, $p < .001$) and for the item ___-3=12 (F (2, 527) = 16.870, $p < .001$). Post-hoc Tukey analyses for 8+___=15 revealed that 1st grade students significantly underperformed all other grades. Post-hoc Tukey analyses for ___=3=12 revealed that 1st and 2nd grade students significantly underperformed 3rd grade students and that 1st grade students significantly unperformed 2nd grade students.

These high p values likely affected the items' discriminating power, given that ceiling effects were reached. The majority of the adjusted item-total correlations were actually higher in the Singapore sample than the U.S. sample, however. The only item that has a significantly negative correlation is ___x6-36 for 3rd grade students, however the p value of 0.98 (i.e., nearly at the maximum) likely accounts for this.

Table 26.

Item Statistics for Open Number Sentence Items: Singapore

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
8+__=15						
1st Grade	0.88	-4.18	0.300	0.93	1.02	1.71
2nd Grade	0.97	-3.55	0.183	0.83	1.23	3.29
3rd Grade	1.00	-	0.000	-	-	-
4th Grade	0.96	-2.33	0.120	0.88	1.22	1.56
5th Grade	0.99	-3.39	0.090	0.98	1.04	0.87
___-3=12						
1st Grade	0.70	-2.42	0.436	0.95	0.99	1.27
2nd Grade	0.89	-1.61	0.237	0.75	1.22	2.46
3rd Grade	0.89	0.17	0.276	1.08	0.96	0.61
___x8=56						
3rd Grade	0.82	-1.17	0.442	0.92	1.11	0.92
4th Grade	0.97	-1.80	0.130	1.04	0.99	0.45
___x6=36						
3rd Grade	0.98	-1.86	-0.141	0.84	1.10	3.94
4th Grade	0.98	-2.98	0.193	0.93	1.07	2.73
5th Grade	0.96	-2.06	0.256	1.03	0.94	0.72

Equivalence: Singapore

The item statistics of difficulty and discrimination from both CTT and IRT for the equivalence items for the Singapore sample are summarized in Table 27. With p values of 0.33 for 8+4=___+5 and 0.22 for 6+7=___+4, it is clear that 1st grade students struggle with such equivalence items. Despite this, however, it appears that the majority of 2nd, 3rd, 4th, and 5th grade students have mastered these items with p values of 0.82 and above. These results were mirrored by the analysis of differences by grade.

There was a statistically significant difference by grade for the item 8+4=___+5 (F (3, 655) = 133.998, $p < .001$) and for the item 6+7=___+4 (F (4, 863) = 28.100, $p < .001$). Post-hoc Tukey analyses for 8+4=___+5 and 6+7=___+4 revealed that 1st grade students significantly underperformed all other grades.

Like the U.S. sample, the discrimination indices for the Singapore sample indicate that the equivalence items are very good discriminators. Nearly all of the adjusted item-total correlations are statistically significant ($p < .001$) and all of the IRT discrimination parameters are 1.00 or higher (i.e., indicating better than average discrimination for that difficulty level). Even though the adjusted item-total correlation for 40/4=80/__ for 4th Grade is negative, the IRT discrimination is above 1.00, therefore this item is acceptable and will be retained.

Table 27.

Item Statistics for Equivalence Items: Singapore

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
8+4=___+5						
1st Grade	0.33	0.13	0.673	1.38	0.74	0.70
3rd Grade	0.96	-1.12	0.315	1.10	0.87	0.38
4th Grade	0.88	-0.92	0.403	1.10	0.89	0.63
5th Grade	1.00	-	0.000	-	-	-
6+7=___+4						
1st Grade	0.22	2.25	0.717	1.57	0.50	0.31
2nd Grade	0.82	-0.73	0.488	1.08	0.90	0.88
3rd Grade	0.83	-1.33	0.533	1.13	0.83	0.96
4th Grade	0.94	-1.04	0.307	1.11	0.90	0.38
5th Grade	0.93	-1.43	0.283	1.02	0.96	1.01
8+___=3+9						
1st Grade	0.36	-0.03	0.706	1.49	0.68	0.60
3rd Grade	0.91	-2.25	0.500	1.11	0.75	1.74
3x4=___x2						
3rd Grade	0.83	-1.30	0.518	1.09	0.89	1.10
4th Grade	0.95	-2.05	0.394	1.14	0.82	0.30
40/4=80/___						
4th Grade	0.94	-1.04	-0.009	1.07	0.93	0.51
5th Grade	0.93	-1.29	0.219	1.00	1.05	0.74

Work with Variables: Singapore

The item statistics of difficulty and discrimination from both CTT and IRT for the work with variables items for the Singapore sample are summarized in Table 28. Like open number sentences, it appears students in Singapore had largely mastered work with variables items. P

values of 0.62 for 1st grade students were the lowest of all p values across even more difficult items across the grade levels.

There was a statistically significant difference by grade for the item 10-g=2 (F (1, 441) = 78.513, $p < .001$) with 1st grade students significantly underperforming 2nd grade students with p values of 0.62 and 0.94, respectively. There was a statistically significant difference by grade for the items 4+a+a=10 (F (2, 618) = 7.818, $p < .001$) and c+c+3=15 (F (3, 674) = 7.167, $p < .001$), however post-hoc Tukey analyses for did not reveal any differences between the grades at the required significance level ($p < .001$).

The adjusted item-total correlation discrimination indices for the work with variables items are largely significantly positive and the IRT discrimination parameters are largely greater than 1.00. One correlation (i.e., c+c+3=15 for 5th Grade) is negative however it is not negligibly different from zero and it is likely due to a ceiling effect (i.e., the p value is 0.97).

Table 28.

Item Statistics for Work with Variables Items: Singapore

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
10-g=2						
1st Grade	0.62	-1.74	0.536	1.00	1.00	1.07
2nd Grade	0.94	-2.47	0.273	1.07	0.89	0.53
c+c+3=15						
1st Grade	0.70	-1.25	0.524	1.31	0.83	0.58
3rd Grade	0.82	-1.24	0.605	1.25	0.76	0.47
4th Grade	0.97	-1.78	0.217	1.01	0.87	6.70
5th Grade	0.97	-1.88	-0.027	0.92	1.09	1.11
4+a+a=10						
2nd Grade	0.86	-1.20	0.594	1.28	0.66	0.34
3rd Grade	0.98	-1.86	0.283	1.09	0.90	0.20
4th Grade	0.93	-1.74	0.409	1.15	0.79	0.47
5xgxg=20						
3rd Grade	0.70	0.22	0.549	1.27	0.82	0.61
4th Grade	0.94	-1.04	0.130	1.12	0.87	0.36
5th Grade	0.96	-1.95	0.393	1.11	0.83	0.38
x+x+y=12; x-y=3						
5th Grade	0.80	0.18	0.248	0.92	1.06	1.16

Numerical Patterns: Singapore

The item statistics of difficulty and discrimination from both CTT and IRT for the numerical patterns items for the Singapore sample are summarized in Table 29. In general, Singapore students were very competent with numerical patterns. 1st and 2nd grade students easily mastered counting by 2's patterns with p values of 0.93 and 0.94, respectively, while older 3rd, 4th, and 5th grade students easily mastered more difficult nonlinear patterns with p values ranging from 0.83 to 0.94. 1st and 2nd grade students did seem to struggle more with more difficult numerical patterns such as counting by 6's (p value of 0.58) or counting backwards by 3's (p values of 0.44 and 0.76). These results are mirrored in the differences by grade analyzed below.

There was a statistically significant difference by grade for the item 15, 21, ___, 33, 39 (F (2, 410) = 16.325, $p < .001$). Post-hoc Tukey analyses revealed that 1st grade students significantly underperformed 3rd and 4th grade students. There was also a statistically significant difference by grade for the item 72, 69, 66, 63, 60, ___ (F (1, 413) = 48.253, $p < .001$), with 1st grade students significantly underperforming 2nd grade students. No other items showed significant differences by grade.

Like the open number sentence items for both the U.S. and Singapore and like the U.S. numerical pattern items, the numerical pattern items for Singapore were not highly discriminatory. Results varied widely however these results should be interpreted cautiously because of the ceiling effects that occurred with so many of the items. Because these types of items are often prerequisites for more difficult types of skills, and because this is a diagnostic assessment, it is important to retain such items.

Table 29.

Item Statistics for Numerical Pattern Items: Singapore

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
3, 5, 7, 9, 11, ___						
1st Grade	0.93	-2.91	-0.146	0.85	0.93	3.06
2nd Grade	0.94	-2.37	0.362	1.00	1.01	0.78
15, 21, ___, 33, 39						
1st Grade	0.58	-0.30	0.490	1.25	0.88	0.71
3rd Grade	0.85	-1.48	0.129	0.63	1.37	1.75
4th Grade	1.00	-	0.000	-	-	-
72, 69, 66, 63, 60, ___						
1st Grade	0.44	-0.72	0.497	0.88	1.09	0.99
2nd Grade	0.76	-0.16	0.401	1.03	0.97	0.92
35, 30, 25, ___, 15						
3rd Grade	1.00	-	0.000	-	-	-
4th Grade	0.98	-3.16	0.074	0.98	0.97	1.35
5th Grade	0.99	-3.81	0.039	0.98	1.06	0.58
2, 3, 5, 8, ___						
3rd Grade	0.83	-1.30	0.393	0.89	1.16	0.93
4th Grade	0.86	-0.83	0.327	0.98	1.01	0.95
5th Grade	0.92	-1.05	0.177	0.91	1.09	1.46
50, 49, 47, 44, ___						
3rd Grade	0.90	0.06	-0.063	0.75	1.19	2.55
4th Grade	0.94	-1.04	-0.271	0.91	0.99	2.34

Figural Patterns: Singapore

The item statistics of difficulty and discrimination from both CTT and IRT for the figural patterns items in the Singapore sample are summarized in Table 30.

Younger students in Singapore seemed to struggle much more with linear patterns when they were viewed figurally rather than numerically. 1st grade students solved the linear figural problems 4, 7, 10, ___ and 5, 8, 11, ___ with the low p values of 0.32 and 0.27, respectively, for example.

There was a statistically significant difference by grade for both the items 4, 7, 10, ___ (F (3, 590) = 44.168, $p < .001$) and 5, 8, 11, ___ (F (4, 885) = 21.114, $p < .001$). Post-hoc Tukey analyses revealed that 1st grade students significantly underperformed 3rd, 4th, and 5th grade

students on the 4, 7, 10, ___ item. Post-hoc Tukey analyses revealed that for the item 5, 8, 11, ___, 1st grade students significantly underperformed all other grades, and 2nd and 4th grade students significantly underperformed 5th grade students. Finally, there was also a statistically significant difference by grade for the nonlinear figural pattern item 3, 6, 10, ___ (F (4, 430) = 23.217, $p < .001$), with 1st and 2nd grade students significantly underperforming all other grades.

The discrimination indices for the figural pattern items are largely variable across the items, however the majority of the adjusted item-total correlations are positive and statistically significant ($p < .001$). There are a few items with negative correlations or correlations close to zero, but none of the negative correlations are significantly different from zero. Further, again, as with many of the Singapore items, ceiling effects were prevalent and for one item a floor effect may be occurring.

Table 30.

Item Statistics for Figural Pattern Items: Singapore

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
4, 7, 10, ___						
1st Grade	0.32	0.01	0.423	0.84	1.14	0.99
3rd Grade	0.70	-0.20	0.408	0.82	1.13	1.13
4th Grade	0.92	-0.58	-0.081	0.89	1.15	1.01
5th Grade	0.87	0.01	0.067	0.81	1.11	2.90
4, 7, 10, ___ Far 7th						
3rd Grade	0.38	1.87	0.411	0.98	1.01	1.02
4, 7, 10, ___ Far 20th						
4th Grade	0.51	2.23	-0.010	0.43	1.19	1.42
5th Grade	0.62	1.84	0.130	0.49	1.24	1.47
5, 8, 11, ___						
1st Grade	0.27	-0.78	0.648	1.53	0.60	0.43
2nd Grade	0.64	0.71	0.376	0.96	1.03	1.03
3rd Grade	0.83	0.80	0.300	1.09	0.93	0.80
4th Grade	0.75	0.37	0.326	0.89	1.07	1.11
5th Grade	0.89	-0.78	0.288	1.03	1.00	0.75
5, 8, 11, ___ Far 7th						
2nd Grade	0.40	2.10	0.357	1.16	0.95	0.84
3rd Grade	0.65	2.01	0.353	1.14	0.94	0.87
5, 8, 11, ___ Far 20th						
4th Grade	0.26	3.57	0.362	1.02	0.98	0.93
5th Grade	0.51	1.87	0.236	0.93	1.03	1.04
3, 6, 10, ___						
1st Grade	0.21	1.38	0.260	1.16	0.87	0.60
2nd Grade	0.55	1.23	0.520	1.36	0.85	0.72
3rd Grade	0.81	0.99	0.157	0.93	1.11	0.88
4th Grade	0.88	-0.09	-0.063	0.88	1.12	1.10
5th Grade	0.87	0.11	0.541	1.30	0.69	0.37
3, 6, 10, ___ Far 10th						
2nd Grade	0.21	3.26	0.196	0.88	1.09	1.05
3rd Grade	0.38	3.49	0.375	1.32	0.85	0.75
3, 6, 10, ___ Far 20th						
4th Grade	0.13	4.81	-0.079	0.81	1.08	2.23
5th Grade	0.57	2.20	0.528	1.69	0.67	0.59

Generalization

The item statistics of difficulty and discrimination from both CTT and IRT for the generalization items are summarized in Table 31. Overall, of all the algebraic skills measured, generalization seemed to be the most difficult skill for students in Singapore. Even the easier

item a+0=a, for example, which has a chance of guessing correctly of 50%, saw only 57% of 1st grade students and 72% of 2nd grade students answering correctly. Further, when asked to explain 'why' the student knew it was true or false, p values were quite low at 0.08, 0.28, and 0.52 for 1st, 2nd, and 3rd grade students, respectively. Older students performed better when asked to say whether the more difficult item a-b=b-a was true or false: 69%, 75%, and 91% of 3rd, 4th, and 5th grade students, respectively, reported that this was false. Students again struggled greatly when asked to explain 'why' though, with p values of 0.13, 0.19, and 0.26 for 3rd, 4th, and 5th grade students, respectively.

There was a statistically significant difference by grade for the item a+0=a true or false (F (2, 320) = 10.441, $p < .001$). Post-hoc Tukey analyses revealed that 1st grade students significantly underperformed 2nd and 3rd grade students and 2nd grade students significantly underperformed 3rd grade students. There was also a statistically significant difference by grade when students were asked to explain the item a+0=a (F (2, 343) = 17.890, $p < .001$), with 1st and 2nd grade students significantly underperforming 3rd grade students.

There was a statistically significant difference by grade for the item a-b=b-a true or false (F (2, 859) = 3.704, $p < .001$). Post-hoc Tukey analyses revealed that 3rd and 4th grade students significantly underperformed 5th grade students. There was also a statistically significant difference by grade when students were asked to explain the item a-b=b-a (F (2, 857) = 19.495, $p < .001$), with 3rd grade students significantly underperforming 5th grade students.

Unlike issues with some of the other dimension strands, the generalization items all appear to be largely high quality discriminators – at least in terms of highly positive adjusted item-total correlations.

Table 31.

Item Statistics for Generalization: Singapore

Item	P value by CTT	Difficulty Parameter by IRT	Adjusted item-total Correlation	Discrimination Parameter by IRT	Infit Mean-Square	Outfit Mean-Square
a+0=a true / false						
1st Grade	0.50	0.50	0.377	0.54	1.33	1.54
2nd Grade	0.72	0.30	0.290	0.74	1.21	1.16
3rd Grade	0.91	0.00	0.263	1.05	0.97	0.65
a+0=a explain						
1st Grade	0.08	3.48	0.336	0.95	1.07	0.89
2nd Grade	0.28	2.61	0.378	0.99	0.99	1.32
3rd Grade	0.51	2.78	0.172	0.71	1.19	1.24
a-b=b-a true / false						
3rd Grade	0.69	-0.09	0.291	0.61	1.27	1.44
4th Grade	0.75	0.33	0.175	0.65	1.27	1.39
5th Grade	0.91	-1.04	0.285	1.03	1.01	0.63
a-b=b-a explain						
3rd Grade	0.13	5.16	0.401	1.10	0.92	0.91
4th Grade	0.19	5.43	0.416	1.08	0.94	0.88
5th Grade	0.26	4.16	0.349	1.12	0.92	0.86

Internal Structure Validity Evidence

Cronbach's alpha coefficients were calculated for the Singapore version of the assessment at each grade level to demonstrate internal structure validity evidence. The alpha coefficients were much lower for the Singapore assessments than for the U.S. assessments, however a great deal of revision occurred after the Singapore assessment was given. The number of items was increased which likely led to the increased reliability. Further, the ceiling effects seen with many items earlier (i.e., with nearly all students answering correctly) also likely led to lower reliability coefficients in Singapore than in the U.S. Assessment versions were only included in the reliability analysis when sample sizes exceeded 100. Of the five assessments with sample sizes larger than 100, four of these assessments had alpha coefficients at 0.70 and above. Only the 5th grade assessment, with a reliability coefficient of 0.63, should potentially be examined further for reliability issues. These results are displayed below in Table 32.

Table 32.

Cronbach's Alpha Reliability Coefficients for the Singapore Assessments

	Number of Items	Reliability Coefficient
1st Grade	16	0.87 (*n* = 177)
2nd Grade	18	0.79 (*n* = 167)
3rd Grade	19	0.83 (*n* = 254)
4th Grade	18	0.72 (*n* = 275)
5th Grade	17	0.63 (*n* = 253)

Correlational statistics for student grade level and the different algebraic thinking tasks on the assessment are displayed below in Table 33. By examining this table one can see that all of the tasks were statistically significantly positively correlated with each other. These results provide additional confirmation to the findings of the U.S. sample: these results demonstrate initial internal structure validity evidence indicating that the dimension strands act similarly to each other and that there may indeed be unidimensionality across the construct.

Table 33.

Correlation Statistics of Algebraic Thinking Dimensions: Singapore

	1.	2.	3.	4.	5.	6.	7.	8.
1. Open Number Sentences	-							
2. Equivalence	.365*	-						
3. Work with Variables	.399*	.494*	-					
4. Efficient Numerical Calculation	.277*	.306*	.303*	-				
5. Generalization	.178*	.270*	.244*	.242*	-			
6. Numerical Patterns	.333*	.487*	.394*	.303*	.228*	-		
7. Figural Patterns	.290*	.434*	.382*	.298*	.348*	.369*	-	
8. Figural Patterns Generalize	.160*	.293*	.240*	.219*	.270*	.231*	.515*	-

Note. N=1,529, *p<.01

Construct-Related Validity Evidence: Comparisons of Singapore to U.S. Results

The item statistics of difficulty and discrimination from both CTT and IRT for the items in the modeling, functions, and generalized arithmetic dimensions are summarized in Tables 34, 35, and 36, respectively. This comparison analysis only involved items which are exactly the same across the two samples. Several generalization items that were not exactly the same will be discussed but not officially compared because the items were not identical across the two samples.

Open Number Sentences. Although Singapore students appeared to outperform U.S. students on open number sentence items in 1st and 2nd grade, by 3rd grade any differences between the two samples were non-existent or negligible. For example, for the item 8+__=15, students from Singapore outperformed students in the U.S. in 1st grade and in 2nd grade, however the rigorous significance levels required by this work were not met (i.e., the *F* statistics were less than .05 but not less than .001). There were no significant differences at 3rd grade, 4th grade, or 5th grade. For the more difficult item ___-3=12, students from Singapore significantly outperformed students in the U.S. in 1st grade and 2nd grade, but not in 3rd grade. It therefore appears that students in the U.S. and students in Singapore have similar knowledge when it comes to open number sentences.

Equivalence. Unlike what was the case for the common open number sentence items, Singapore students significantly outperformed U.S. students on common equivalence items across 1st through 5th grades. For example, for the item 8+4=__+5, students from Singapore significantly outperformed students in the U.S. in 1st grade, 3rd grade, 4th grade and 5th grade. The data related to 6+7=__+4 must be interpreted cautiously because the Singapore sample sizes were very small for 1st and 4th grades. For this item students from Singapore did not

significantly outperformed students in the U.S. in 1st grade, but they did at every other grade level. Further, the effect sizes for equivalence items were quite large, ranging from 0.26 to 0.53, meaning that between 26% and 53% of the variance of the item is explained by the country effect.

Work with Variables. As was the case with equivalence items, Singapore students outperformed U.S. students on common work with variables items across all the grades, although the differences became less apparent as the grades increase. It is important to again note sample sizes in the Singapore sample (i.e., for 1st grade students completing c+c+3=15 the sample size is only 26); as discussed previously because of logistical issues the alternate forms of the assessment were not spiraled and distributed equally as they were in the U.S. For example, for the item 10-g=2, students from Singapore significantly outperformed students in the U.S. in 1st grade and in 2nd grade. For the more difficult item c+c+3=15, students from Singapore outperformed students in the U.S. across the grade levels, although the effect sizes drastically decreased after 1st and 2nd grades, where they were 0.45 and 0.54, respectively (i.e., between 45% and 54% of the variance was accounted for by the country effect).

Table 34.

Comparisons of Modeling Common Items: Singapore to U.S.

Item	P value by CTT: Singapore	P value by CTT: U.S.	*F*	Effect Size
8+___=15				
1st Grade	0.88 (*SD* = 0.33)	0.79 (SD = 0.41)	5.042	0.12
2nd Grade	0.97 (SD = 0.16)	0.92 (SD = 0.28)	6.341	0.13
3rd Grade	1.00 (SD = 0.00)	0.98 (SD = 0.15)	2.634	0.10
4th Grade	0.96 (SD = 0.20)	0.99 (SD = 0.10)	3.782	0.09
5th Grade	0.99 (SD = 0.11)	0.99 (SD = 0.07)	0.451	0.03
___-3=12				
1st Grade	0.70 (SD = 0.47)	0.47 (SD = 0.50)	20.482*	0.23
2nd Grade	0.89 (SD = 0.31)	0.66 (SD = 0.47)	32.166*	0.28
3rd Grade	0.89 (SD = 0.31)	0.76 (SD = 0.43)	7.672	0.16
8+4=___+5				
1st Grade	0.33 (SD = 0.47)	0.11 (SD = 0.31)	29.386*	0.26
2nd Grade	-	0.30 (SD = 0.46)	-	-
3rd Grade	0.96 (SD = 0.19)	0.56 (SD = 0.50)	65.723*	0.48
4th Grade	0.88 (SD = 0.32)	0.52 (SD = 0.50)	93.387*	0.41
5th Grade	1.00 (SD = 0.00)	0.69 (SD = 0.47)	17.623*	0.27
6+7=___+4				
1st Grade	0.22 (SD = 0.42)	0.15 (SD = 0.36)	0.844	0.07
2nd Grade	0.82 (SD = 0.39)	0.29 (SD = 0.46)	144.911*	0.53
3rd Grade	0.83 (SD = 0.37)	0.62 (SD = 0.49)	96.396*	0.41
4th Grade	0.94 (SD = 0.23)	0.53 (SD = 0.50)	23.243*	0.30
5th Grade	0.93 (SD = 0.25)	0.66 (SD = 0.48)	46.243*	0.33
10-g=2				
1st Grade	0.62 (SD = 0.49)	0.40 (SD = 0.49)	19.694*	0.22
2nd Grade	0.94 (SD = 0.24)	0.54 (SD = 0.50)	106.460*	0.47
c+c+3=15				
1st Grade	0.70 (SD = 0.47)	0.14 (SD = 0.35)	49.626*	0.45
2nd Grade	0.86 (SD = 0.34)	0.32 (SD = 0.48)	151.029*	0.54
3rd Grade	0.82 (SD = 0.38)	0.68 (SD = 0.47)	13.020*	0.16
4th Grade	0.97 (SD = 0.53)	0.69 (SD = 0.46)	33.693*	0.26
5th Grade	0.97 (SD = 0.16)	0.80 (SD = 0.40)	7.042	0.17

* $p < .001$

Numerical Patterns. Table 35a below displays the comparative results of Singapore students vs. U.S. students on the common numerical patterns items. Unlike equivalence and work with variables, the results for linear numerical patterns are more consistent with those of open number sentences: although Singapore students significantly outperformed U.S. students in the early grades on more difficult linear items, there were no differences by country for the

counting by 2's pattern in 1st and 2nd grade and no differences by country on the more difficult linear items in 3rd and 4th grades. Unlike linear numerical patterns, however, Singapore students did significantly outperformed U.S. students on nonlinear numerical patterns across the grades. The effect sizes for the items 2, 3, 5, 8, ___ and 50, 49, 47, 44, ___, for example, ranged from 0.29 to 0.48, indicating that between 29% and 48% of the variance can be explained by the country effect.

Figural Patterns. Table 35b below displays the comparative results of Singapore students vs. U.S. students on the common figural patterns items. Unlike many of the other items, for the linear figural pattern item 4, 7, 10, __, students from Singapore did not significantly outperform students in the U.S. in any of the grades, however for the very similar item 5, 8, 11, ___, students from Singapore significantly outperformed students in the U.S. in 2nd, 3rd, 4th, and 5th grades. The far generalization items could not be compared statistically because while "generalize to the 10th figure" was used with U.S. students, Singapore students were asked to either generalize to the 7th or to the 20th figure, depending on the age. It is interesting to note that Singapore students still outperformed U.S. students, likely at significant levels, even when they were asked to generalize to the 20th figure and U.S. students were only asked to generalize to the 10th figure.

Again, similar to the linear figures above, for the nonlinear figure pattern 3, 6, 10, ___, students from Singapore did not significantly outperform students in the U.S. in any of the grades, however some grades (i.e., 2nd and 3rd grades) were very close but were not because of rigorous statistical requirements (i.e., that the significant level be less than .001 instead of .05). Again, comparisons could not be made on the far generalization items; however it may be still be noted that students in Singapore outperformed students in the U.S. at all grade levels.

Table 35a.

Comparisons of Functions Numerical Patterns Common Items: Singapore to U.S.

Item	P value by CTT: Singapore	P value by CTT: U.S.	*F*	Effect Size
3, 5, 7, 9, 11, ___				
1st Grade	0.93 (SD = 0.27)	0.63 (SD = 0.48)	9.598	0.21
2nd Grade	0.94 (SD = 0.24)	0.84 (SD = 0.37)	9.057	0.15
15, 21, ___, 33, 39				
1st Grade	0.58 (SD = 0.50)	0.28 (SD = 0.45)	17.234*	0.27
2nd Grade	-	0.60 (SD = 0.49)	-	-
3rd Grade	0.85 (SD = 0.36)	0.85 (SD = 0.36)	0.000	0.00
4th Grade	1.00 (SD = 0.00)	0.85 (SD = 0.36	6.374	0.16
2, 3, 5, 8, ___				
3rd Grade	0.83 (SD = 0.38)	0.47 (SD = 0.50)	77.771*	0.37
4th Grade	0.86 (SD = 0.33)	0.54 (SD = 0.54)	76.096*	0.37
5th Grade	0.92 (SD = 0.28)	0.69 (SD = 0.47)	45.104*	0.29
50, 49, 47, 44, ___				
3rd Grade	0.90 (SD = 0.30)	0.42 (SD = 0.50)	83.040*	0.48
4th Grade	0.94 (SD = 0.23)	0.53 (SD = 0.50)	23.626*	0.31

* $p < .001$

Table 35b.

Comparisons of Functions Figural Patterns Common Items: Singapore to U.S.

Item	P value by CTT: Singapore	P value by CTT: U.S.	*F*	Effect Size
4, 7, 10, ___				
1st Grade	0.32 (SD = 0.47)	0.27 (SD = 0.45)	0.859	0.05
2nd Grade	-	0.40 (SD = 0.49)	-	-
3rd Grade	0.70 (SD = 0.46)	0.60 (SD = 0.49)	5.192	0.10
4th Grade	0.92 (SD = 0.28)	0.71 (SD = 0.45)	6.684	0.17
5th Grade	0.87 (SD = 0.34)	0.75 (SD = 0.43)	2.715	0.11
4, 7, 10, ___ Far 7th / 10th / 20th				
1st Grade	0.27 (SD = 0.45)	-	-	-
2nd Grade	-	-	-	-
3rd Grade	0.38 (SD = 0.49)	0.17 (SD = 0.38)	-	-
4th Grade	0.51 (SD = 0.51)	0.21 (SD = 0.41)	-	-
5th Grade	0.62 (SD = 0.49)	0.23 (SD = 0.43)	-	-
5, 8, 11, ___				
1st Grade	0.27 (SD = 0.45)	0.23 (SD = 0.42)	0.207	0.03
2nd Grade	0.64 (SD = 0.48)	0.37 (SD = 0.48)	27.846*	0.27
3rd Grade	0.83 (SD = 0.38)	0.56 (SD = 0.50)	23.096*	0.28
4th Grade	0.75 (SD = 0.44)	0.56 (SD = 0.50)	18.329*	0.19
5th Grade	0.89 (SD = 0.31)	0.76 (SD = 0.43)	12.744*	0.17
5, 8, 11, ___ Far 7th / 10th / 20th				
2nd Grade	0.40 (SD = 0.49)	-	-	-
3rd Grade	0.65 (SD = 0.48)	0.15 (SD = 0.36)	-	-
4th Grade	0.26 (SD = 0.44)	0.12 (SD = 0.33)	-	-
5th Grade	0.51 (SD = 0.50)	0.28 (SD = 0.45)	-	-
3, 6, 10, ___				
1st Grade	0.21 (SD = 0.41)	0.30 (SD = 0.46)	1.385	0.08
2nd Grade	0.55 (SD = 0.50)	0.40 (SD = 0.49)	8.287	0.15
3rd Grade	0.81 (SD = 0.40)	0.61 (SD = 0.49)	12.221	0.21
4th Grade	0.88 (SD = 0.33)	0.70 (SD = 0.46)	4.871	0.15
5th Grade	0.87 (SD = 0.34)	0.77 (SD = 0.42)	1.704	0.09
3, 6, 10, ___ Far 10th / 20th				
2nd Grade	0.21 (SD = 0.41)	-	-	-
3rd Grade	0.38 (SD = 0.49)	0.03 (SD = 0.16)	-	-
4th Grade	0.13 (SD = 0.34)	0.07 (SD = 0.26)	-	-
5th Grade	0.57 (SD = 0.50)	0.12 (SD = 0.32)	-	-

* $p < .001$

Generalization. Table 36 below displays the comparative results of Singapore students vs. U.S. students on the common generalization items. For the generalization item a+0=a, students from Singapore did not significantly outperform students in the U.S. in 1st grade, however they did in 2nd and 3rd grade for marking the item as true or false. Students in Singapore

only significantly outperformed students in the U.S. when explaining why it was true or false in 3rd grade. The effect sizes for these significant grades ranged from 0.28 to 0.30, meaning that approximately 30% of the variance could be accounted for by the country effect. It is important to note that p values for the explain item were low in both countries, however.

For the generalization item a-b=b-a, 3rd, 4th, and 5th grade students from Singapore significantly outperformed students in the U.S. in determining whether a-b=b-a was true or false. Only 3rd grade students in Singapore, however, significantly outperformed 3rd grade students in the U.S. in explaining why this was true or false. Again, p values for both countries were quite low for the explaining items.

Table 36.

Comparisons of Generalization Common Items: Singapore to U.S.

Item	P value by CTT: Singapore	P value by CTT: U.S.	*F*	Effect Size
a+0=a true / false				
1st Grade	0.50 (SD = 0.51)	0.32 (SD = 0.48)	3.497	0.13
2nd Grade	0.72 (SD = 0.45)	0.41 (SD = 0.49)	33.885*	0.30
3rd Grade	0.91 (SD = 0.29)	0.66 (SD = 0.48)	24.202*	0.28
a+0=a explain				
1st Grade	0.08 (SD = 0.24)	0.06 (SD = 0.17)	0.512	0.05
2nd Grade	0.28 (SD = 0.38)	0.16 (SD = 0.28)	9.871	0.17
3rd Grade	0.51 (SD = 0.43)	0.28 (SD = 0.37)	23.014*	0.28
a-b=b-a true / false				
3rd Grade	0.69 (SD = 0.47)	0.33 (SD = 0.47)	62.917*	0.34
4th Grade	0.75 (SD = 0.43)	0.48 (SD = 0.50)	41.496*	0.28
5th Grade	0.91 (SD = 0.29)	0.63 (SD = 0.48)	57.865*	0.34
a-b=b-a explain				
3rd Grade	0.13 (SD = 0.22)	0.05 (SD = 0.15)	14.800*	0.17
4th Grade	0.19 (SD = 0.25)	0.12 (SD = 0.21)	12.0258	0.16
5th Grade	0.26 (SD = 0.28)	0.19 (SD = 0.27)	5.401	0.11

* $p < .001$

Summary

It is clear from the analyses comparing student results in Singapore to students results in the U.S. that students in Singapore largely, and mostly statistically significantly, outperformed students in the U.S. This can be viewed as construct-related validity evidence for this diagnostic assessment because students in Singapore routinely outperform students in the U.S. on common international mathematics tests (e.g., the TIMSS). Further, it appears the Singapore education system may emphasize algebraic thinking skills at earlier grades with the inclusion of the 'Model Method' in their curriculum. The fact that students in Singapore significantly outperformed students in the U.S. on many of the more algebraic tasks (i.e., equivalence, work with variables, figural patterns, etc.) indicates initial construct evidence that the interpretation of the assessment scores may be valid.

Chapter 7: Results of the Think Aloud Protocols & Conception Analysis

In this chapter I will describe the results obtained from the think-aloud interviews conducted with 73 students in Washington State coupled with the conception analysis of the 1,745 students in Washington State administered the assessments. Qualitative coding combined with quantitative descriptive statistics will be reported. This data was analyzed in this way to demonstrate that these diagnostic assessment items have the potential to elicit rich alternative conceptions from students. As this is the purpose of this diagnostic assessment tool, it is important to demonstrate that the data produced from such a tool has the potential to provide teachers with rich information to inform instruction.

Description of the Think-Aloud Interview Student Sample

Seventy-three students of varying abilities from two different elementary schools in King County of Washington State were selected to participate in the think aloud protocol. Grade level participants included 18 1st grade students (25%), 12 2nd grade students (16%), 11 3rd grade students (15%), 10 4th grade students (14%), and 22 5th grade students (30%). The two alternate versions of each grade level assessment were randomly assigned to 50% of the students in each grade level. In the think-aloud protocol interview students were asked to verbalize their cognitive processes when answering each of the questions.

Validation Purpose of the Think Aloud Study

The results of these think aloud interviews are combined with the assessment data discussed previously when analyzing common student conceptions. If problems are not reported here that have been discussed previously it is because no common conceptions other than the correct answer became apparent. Across this data, alternative conceptions which 10% or more of the student sample seemed to possess are bolded for ease of reading and interpretation. This

analysis of the think aloud interviews in conjunction with the coding of the student results provide validity evidence because they justify that the assessment results can generate diagnostic information useful to teachers that captures students' alternative conceptions. Further, many of the alternative conceptions captured are consistent with the results discovered by other researchers. Finally, it was discovered that the think aloud results mirrored the results of the large assessment study, which provides evidence that the assessment results can provide similar results to conducting lengthy one-on-one interviews with students.

Open Number Sentences

For the most part, students solved open number sentences correctly. Therefore there are not many alternative conceptions in this subarea. However, when solving the open number sentence 8=6+__, many students appeared to misconstrue the item as 8+6=___, with answers of 14 to indicate this. These results mirrored results found by other researchers demonstrating students often struggle with nonstandard open number sentences when the answer comes first (Lindvall & Ibarra, 1980). The think aloud interviews with students also revealed this alternative conception. For example, during an interview one student said that "8 equals 6+14 because 8+6=14" while other students described their different adding strategies (i.e., counting on) to reach 14. The most common answers to the 8=6+__ item are displayed below in Table 37.

Table 37.

Common Answers to the Open Number Sentence 8=6+__

	Blank	2 (correct answer)	14
1st Grade	**18 (10%)**	82 (46%)	**41 (23%)**

Second, when students solved the item ___-3=12, several students appeared to misconstrue the item as either 12-___=3 or as ___+3=12, as many students answered with 9. Other students may have simply added wrong, believing that 14-3=12. Again, these results mirrored the alternative conceptions discovered by others (Lindvall & Ibarra, 1980). The most common answers to the __-3=12 item are displayed below in Table 38.

Table 38.

Common Answers to the Open Number Sentence __-3=12

	Blank	9	14	15 (correct)
1st Grade	**20 (11%)**	**26 (15%)**	16 (9%)	83 (47%)
2nd Grade	4 (2%)	**21 (12%)**	10 (6%)	103 (57%)
3rd Grade	1 (1%)	**26 (15%)**	4 (2%)	131 (76%)
All	25 (5%)	**73 (14%)**	30 (6%)	317 (60%)

Equivalence

Alternative conceptions were highly prevalent for equivalence items. These alternative conceptions have been well documented by other researchers (e.g., Carpenter et al., 2005) and this work supported their findings for the most part. Some students were confused by the format and had not seen problems like this before. Many of these students likely simply left the answer blank. As you can see, 11% of 1st grade students left the item blank, with this percentage decreasing as the students increased in age. One 2nd grade student, for example, said, "Hmm, I can't figure this out. This one's kind of hard because it has two things – two signs." Supporting other researchers' work (Carpenter, Franke, & Levi, 2003), many students believed that you should add the first two numbers and obtain 12 (8+4=12). One 1st grade student, for example, said "I counted on, 8+4=12+5". A 4th grade student, for example, said, "8+4= blank + 5, that'd be 12. Because you don't really have to add this one [points to the 5], this really doesn't matter". Others thought you should add all three numbers to obtain 17 (8+4+5=17). One 2nd grade

student, for example, said "8+4 is 12. Why is there a plus 5 on the end? [pause] 12 and 5 is… 17. So it equals 17?" One 5th grade student, for example, said, "8+4=12+5, is that right? Wait… is this like… add those [points to 8 and 4], and then add these? [points to 5 and erases 12] I think that would be… I was thinking 17. Because you add the 5".

Students produced similar conceptions when asked to solve the alternate item 6+7=__+4. Again, many students believed that one should just add the first two numbers together to obtain 13 (6+7=13). One 5th grader, for example, said, "So like 6+7 I just do that in my head, and that's 13". Many others believed you should instead add all three numbers together to obtain 17 (6+7+4=17). One 3rd grader, for example, said, "You add the 7 plus 6 and then you add the 4". A very small number believed in making the numbers add to 7 (3+4=7). For the most part, the results were consistent across the think aloud interviews and the assessments, and were consistent with the findings of other researchers. The most common answers to the equivalence problems 8+4=__+5 and 6+7=__+4 are displayed below in Table 39.

Table 39.

Common Answers to the Equivalence Items

	Correct Answer	Blank	Add to equals 1 number	Equals 2nd number	Add first 2 numbers	Add all 3 numbers
8+4=___+5	7	-	3	4	12	17
1st Grade	19 (9%)	**25 (11%)**	4 (2%)	12 (5%)	**50 (23%)**	**31 (14%)**
2nd Grade	45 (29%)	14 (9%)	6 (4%)	2 (1%)	**61 (40%)**	**17 (11%)**
3rd Grade	97 (56%)	6 (3%)	1 (0%)	0 (0%)	**60 (35%)**	1 (1%)
4th Grade	96 (52%)	0 (0%)	1 (1%)	0 (0%)	**65 (35%)**	14 (8%)
5th Grade	125 (69%)	0 (0%)	1 (1%)	0 (0%)	**40 (22%)**	9 (5%)
All	382 (42%)	45 (5%)	14 (2%)	14 (2%)	**276 (30%)**	72 (8%)
6+7=___+4	9	-	3	7	13	17
1st Grade	26 (13%)	**20 (10%)**	8 (4%)	5 (3%)	**59 (30%)**	**21 (11%)**
2nd Grade	45 (28%)	8 (5%)	5 (3%)	1 (1%)	**49 (30%)**	15 (9%)
3rd Grade	100 (61%)	1 (1%)	1 (1%)	1 (1%)	**36 (22%)**	10 (6%)
4th Grade	105 (53%)	2 (1%)	1 (1%)	2 (1%)	**68 (34%)**	11 (6%)
5th Grade	121 (66%)	0 (0%)	0 (0%)	0 (0%)	**44 (24%)**	9 (5%)
All	397 (44%	31 (3%)	15 (2%)	9 (1%)	**256 (28%)**	66 (7%)

Meaning of the Equal Sign

As discussed previously, many students tend to have an operational views of the equal sign (i.e., viewing it as meaning the answer comes next), instead of a relational view of the equal sign (i.e., viewing it as meaning the same as) (Asquith et al., 2007). The results discovered by other researchers were mirrored in this work: approximately 8% of students across the elementary grades held a relational view of the equal sign when used in the context 5+3=8. This number increased to 18% when the equal sign appeared in an equivalence context like 5+3=4+4. Conversely, 42% and 29% of students, depending on the context, held an operational view of the equal sign. These results were mirrored in the think aloud sample. One 4th grade student, for example, said that the equal sign means "the sum of two numbers", while another 4th grade students said it means "what is the answer", and a 3rd grade student said the equal sign means "that the answer, the number after it is the answer to whatever the problem is". The percentages

of students writing a relational definition versus an operational definition when asked the meaning of the equal sign are displayed below in Table 39.

Table 40.

Common Answers to the Meaning of the Equal Sign Items

	Definition Answer	Definition Same
5+3=8 No context		
1st Grade	**31 (20%)**	10 (7%)
2nd Grade	**43 (30%)**	13 (9%)
3rd Grade	**70 (44%)**	15 (9%)
4th Grade	**96 (52%)**	19 (10%)
5th Grade	**108 (59%)**	8 (4%)
All	**348 (42%)**	65 (8%)
5+3=4+4 Context		
1st Grade	**31 (18%)**	9 (5%)
2nd Grade	**35 (25%)**	17 (12%)
3rd Grade	**41 (26%)**	43 (27%)
4th Grade	**74 (39%)**	39 (21%)
5th Grade	**63 (36%)**	40 (23%)
All	**244 (29%)**	148 (18%)

Work with Variables

Students faced alternative conceptions when solving problems involving variables. These items were similar to the open number sentence items, except a letter was used in place of the blank. Younger students were much more likely to be confused. One 1st grade student, for example, said, "I can't figure these out. The letter ones I can't do." Other students thought that the 'g' in 10-g=2 was actually a 9: "9? Because they kind of look like 9's." Another student thought 'g' was a 6 because "if you turn the g upside down it's going to be a 6". Others simply replaced the answer with numbers from the problem like 2 or 10. Others still thought that the letter stood for a specific word, like 'b' for box or 'g' for Grady (i.e., the students' name). Some students who weren't sure about the problem, though, were still able to talk themselves through it successfully. One 3rd grade student, for example, said: "What? Okay, that's kind of weird. I'm going to skip it. Wait, I think I know, you have to find out what the g is, 10 minus what would

equal 2… so it is 8." It's clear that students held a wide variety of different conceptions about what the variable could stand for, especially in cases where they hadn't been exposed to such problems prior. These findings supported the findings of previous researchers (MacGregor & Stacey, 1997).

Findings were very similar for the alternative work with variables item 6+b=9. Some students understood immediately: for example a 2nd grader said, "Oh, that's kind of like algebra where there's like an x and you try and figure out what it is!" Again, some students who were unfamiliar with problems such as these were able to work them out on their own. One 1st grade student, for example, said, "I'll skip it. Wait, I think I know how to do this one. 6 plus what number equals 9… b must be 3!" Others just listed numbers in the problem such as 6 or 9: "9? Because it's [written] right there."

Students continued to experience alternative conceptions as the work with variables items became more difficult. Students were next asked to solve items with two identical variables such as c+c+3=15. Many of the younger students skipped this item: one 1st grader, for example, commented: "I don't know, can I just skip this one?", while a 2nd grader said, "But I don't know what these [points to 'c's] stand for" and a 3rd grader said, "This one's a little too hard for me". Other students were a little confused about having two variables in one problem. Many students put 12, for example, which is what c+c is. One 3rd grade student, for example, said, "Wait, for both c's or for only one c?" while a 5th grade student said "15 minus 3 is 12".

Students experienced similar issues with the very similar alternative item 4+a+a=10. Many younger students, for example, simply left this item blank. One 2nd grader, for example, simply said, "This is too hard" while a 4th grader said, "I'm going to skip that". Others answered with a number from the problem (i.e., 4 or 10). One 2nd grader, for example, said "4 plus 'a' plus

'a' equals 10 so 'a' equals 10". Others simply added 4 and 10 together to get 14, while others failed to understand the meaning of two variables and answered with 6. One 3rd grader, for example, simply said that the answer was 6 because, "4 plus 6 equals 10". Students in 5th grade, however, had largely grasped this concept: "So 4+a+a is 10 so 10 minus 4 is 6 and 6 divided by 2 is 3 so it would have to be 3". A small number of students believed the answer to be 1 because in a sometimes-used numerical-alphabetic "code", a=1. The most common answers to the work with variables items 10-g=2, 6+b=9, c+c+3=15, and 4+a+a=10 are displayed below in Table 41.

Table 41.

Common Answers to the Work with Variables Items

	Blank	Number in problem	G is the nth letter)	Correct	Number looks like a letter	Number in problem)	Add or subtract numbers
10-g=2	-	2	7	8	9	10	12
1st Grade	**30 (17%)**	**29 (16%)**	7 (4%)	71 (40%)	6 (3%)	9 (5%)	5 (3%)
2nd Grade	15 (8%)	17 (9%)	8 (4%)	82 (42%)	2 (1%)	1 (1%)	4 (2%)
3rd Grade	3 (2%)	6 (4%)	4 (2%)	141 (83%)	0 (0%)	2 (1%)	1 (1%)
All	48 (9%)	52 (10%)	19 (3%)	294 (54%)	8 (1%)	12 (2%)	10 (2%)
6+b=9	-	6	2	3	-	9	15
1st Grade	**26 (15%)**	6 (3%)	7 (4%)	65 (37%)	-	**25 (14%)**	7 (4%)
2nd Grade	**25 (16%)**	4 (3%)	12 (8%)	78 (51%)	-	**21 (14%)**	1 (1%)
3rd Grade	5 (3%)	3 (2%)	6 (3%)	139 (80%)	-	9 (5%)	3 (2%)
All	**56 (11%)**	13 (3%)	25 (5%)	282 (56%)	-	**55 (11%)**	11 (2%)
c+c+3=15	-	3	-	6	-	15	12
1st Grade	**37 (21%)**	**21 (12%)**	-	25 (14%)	-	15 (9%)	12 (7%)
2nd Grade	**33 (21%)**	**21 (14%)**	-	48 (31%)	-	12 (8%)	13 (8%)
3rd Grade	7 (4%)	12 (7%)	-	108 (66%)	-	4 (2%)	11 (7%)
4th Grade	6 (3%)	15 (8%)	-	125 (67%)	-	4 (2%)	15 (8%)
5th Grade	3 (2%)	8 (4%)	-	144 (79%)	-	1 (1%)	12 (7%)
All	**86 (10%)**	77 (9%)	-	450 (52%)	-	36 (4%)	63 (7%)
4+a+a=10	-	4	1	3		10	6
1st Grade	**35 (18%)**	8 (4%)	10 (5%)	35 (18%)	-	**26 (14%)**	17 (9%)
2nd Grade	12 (7%)	3 (2%)	8 (4%)	51 (28%)	-	**20 (11%)**	17 (9%)
3rd Grade	3 (2%)	3 (2%)	3 (2%)	125 (69%)	-	8 (4%)	14 (8%)
4th Grade	3 (1%)	1 (0%)	2 (1%)	170 (83%)	-	2 (1%)	10 (5%)
5th Grade	1 (1%)	1 (1%)	0 (0%)	167 (90%)	-	1 (1%)	6 (3%)
All	54 (6%)	16 (2%)	23 (3%)	548 (58%)	-	57 (6%)	64 (7%)

Numerical Patterns

The majority of students solved numerical patterns correctly, and of those who did not answer correctly, the majority of these answered with an arithmetic mistake (i.e., answering with 15 instead of 16 for 4, 8, 12, ___). Some younger students were confused by these pattern items though. One 1st grader, for example, pointed to the commas and said, "What are these for? If I knew what these were supposed to be, if it was plus or minus, I'd be able to figure out." Several potential alternative conceptions were revealed during the think-aloud interviews, however they did not occur in any great numbers in the large sample. For example, some students believed the patterns to be repeating patterns (i.e., 3, 5, 7, 9, 11, __; the answer is 3 because the pattern repeats). Others believed that the value the pattern is 'counting by' is whatever the first number in the sequence was (i.e., 3, 5, 7, 9, 11, __; the answer is 14 because the pattern is counting by 3's). Others also thought that the answer should just be the number that comes next: one 1st grader said: "It goes 3, 5, 7, 9, 11, and then 12 because 12 comes next". When solving the nonlinear numerical pattern 50, 49, 47, 44, __, for example, several students believed the next number was 43 "because we're counting backwards".

Figural Patterns

For these items, students were asked to both write the number of blocks or dots that would come next as well as draw the next figure in the pattern. While a large number of students both wrote the correct number and drew the correct figure (38%), an oddly identical percentage (38%) of students wrote the incorrect number and the incorrect figure (sometimes with the same two wrong answers and sometimes with two different wrong answers). What is most interesting, however, is the remaining 24% of students who did something else. Sometimes students left one (or both) of these questions blank and sometimes students answered one correctly and not the

other. For example, 6% of students wrote the correct number of blocks yet drew an incorrect figure, while an additional 6% of students wrote the incorrect number of blocks yet drew the correct figure. Although the majority of students answered consistently, it is interesting and important to think about the reasoning of such students.

Very similar findings were discovered for the nonlinear figural patterns. While approximately 41% of students both wrote the correct number and drew the correct figure, approximately 36% of students both wrote the incorrect number and drew the incorrect figure. Similarly, about 4% of students wrote the correct number yet drew an incorrect figure, while about 6% of students drew the correct figure yet wrote the incorrect number. These results are displayed in Table 42 below.

Table 42.

Matching of the Figure Drawn to the Numerical Answer for the Figural Patterns Items

	Blank Number	Incorrect Number	Correct Number	Total
Linear Figural Patterns				
Blank Figure	101 (6%)	18 (1%)	7 (0%)	126 (7%)
Incorrect Figure	66 (4%)	660 (38%)	106 (6%)	831 (48%)
Correct Figure	14 (1%)	102 (6%)	671 (38%)	787 (45%)
Total	181 (10%)	780 (45%)	784 (45%)	1,745 (100%)
Nonlinear Figural Patterns				
Blank Figure	96 (6%)	21 (1%)	8 (0%)	125 (7%)
Incorrect Figure	92 (5%)	633 (36%)	63 (4%)	788 (45%)
Correct Figure	18 (1%)	104 (6%)	709 (41%)	831 (48%)
Total	206 (12%)	758 (43%)	776 (44%)	1,744 (100%)

As for the students who answered the number portion of the problem incorrectly, a wide variety of different conceptions appeared to be occurring. Students often believed that it was a repeating pattern instead of a growing pattern; in that they drew the same figure as the 1st figure in the pattern. One 1st grade student, for example, said: "I thought 4 because it goes 4, 7, 10, and then 4". Other students believed that the next figure should just have one more block than the

most recent block (i.e., 11). For example, one 2nd grade student said: "That one has 10, and so that one has 11".

Very similar findings were discovered for the alternate item form 5, 8, 11, __. Similarly, many students believed that the pattern was just adding one additional block to the most recent figure (i.e., 12). For example, one 1st grade student said: "Because every time it gets taller so this one must be taller by a block so if that one is this tall and I add just another block it makes it… 12". Again, like the other problem, some students believed it was a repeating pattern so the next figure should look identical to the 1st figure. One 2nd grade student, for example, said: "Does it have to be small like that one because that's what I'm going to try to do. So 5 blocks."

Students were also asked to find the next figure in growing nonlinear figural patterns. Results were similar to those for other figural patterns. For the item 1, 4, 9, ___, students also tended to make the shape in a 3x4 rectangle, instead of the 4x4 square that was expected. This was a conception only related to this particular answer, and was held by 15%, 17%, 16%, 9%, and 7% of 1st Grade, 2nd Grade, 3rd Grade, 4th Grade, and 5th Grade students, respectively. These alternative conceptions may simply be because, as other researchers have discovered, few students have had opportunities to experience figural patterns (Warren & Cooper, 2008). The most common answers to the figural linear problems 4, 7, 10, __ and 5, 8, 11, ___ and the figural nonlinear problems 1, 4, 9, __ and 3, 6, 10, __ are displayed in Table 43 below.

Table 43.

Common Answers to the Figural Pattern Items

	Blank	Same as 1st number	Same as 3rd number	The next number	One less or more than correct	Correct	Add 1st number
4, 7, 10	-	4	10	11	12	13	14
1st Grade	**52 (31%)**	7 (4%)	4 (2%)	10 (6%)	**17 (10%)**	37 (22%)	14 (8%)
2nd Grade	**37 (23%)**	4 (2%)	4 (2%)	6 (4%)	**23 (14%)**	48 (30%)	7 (4%)
3rd Grade	11 (7%)	2 (1%)	3 (2%)	5 (3%)	**23 (14%)**	88 (52%)	10 (6%)
4th Grade	6 (3%)	1 (1%)	3 (2%)	4 (2%)	**24 (13%)**	123 (65%)	11 (6%)
5th Grade	5 (3%)	1 (1%)	1 (1%)	5 (3%)	12 (6%)	126 (68%)	13 (7%)
All	**111 (13%)**	15 (2%)	15 (2%)	30 (3%)	**99 (11%)**	422 (48%)	55 (6%)
5, 8, 11	-	5	11	12	15	14	16
1st Grade	**35 (19%)**	14 (8%)	8 (4%)	13 (7%)	10 (6%)	28 (16%)	7 (4%)
2nd Grade	**16 (10%)**	13 (8%)	3 (2%)	11 (7%)	**20 (13%)**	41 (27%)	7 (5%)
3rd Grade	11 (6%)	4 (2%)	3 (2%)	10 (6%)	13 (8%)	76 (45%)	15 (9%)
4th Grade	5 (3%)	3 (2%)	2 (1%)	9 (5%)	**20 (10%)**	98 (51%)	14 (7%)
5th Grade	3 (2%)	0 (0%)	1 (1%)	4 (2%)	11 (6%)	120 (67%)	9 (5%)
All	70 (8%)	34 (4%)	17 (2%)	47 (5%)	74 (8%)	363 (41%	52 (6%)
1, 4, 9	-	1		10	15	16	10
1st Grade	**56 (33%)**	5 (3%)	3 (3%)	10 (6%)	5 (4%)	26 (15%)	10 (9%)
2nd Grade	**40 (26%)**	2 (1%)	1 (1%)	5 (3%)	9 (8%)	30 (19%)	5 (4%)
3rd Grade	15 (9%)	1 (1%)	0 (0%)	7 (4%)	10 (6%)	63 (38%)	7 (4%)
4th Grade	8 (4%)	0 (0%)	1 (1%)	4 (2%)	7 (4%)	127 (65%)	4 (2%)
5th Grade	4 (2%)	0 (0%)	0 (0%)	0 (0%)	2 (1%)	129 (72%)	0 (0%)
All	**123 (14%)**	8 (1%)	5 (1%)	26 (3%)	33 (4%)	375 (43%)	26 (3%)
3, 6, 10	-	3	10	11	14	15	13
1st Grade	**41 (23%)**	8 (4%)	5 (3%)	10 (6%)	14 (8%)	32 (18%)	10 (6%)
2nd Grade	**23 (15%)**	5 (3%)	4 (3%)	2 (1%)	10 (7%)	48 (31%)	14 (9%)
3rd Grade	13 (8%)	1 (1%)	4 (2%)	3 (2%)	15 (9%)	82 (48%)	9 (5%)
4th Grade	4 (2%)	1 (1%)	2 (1%)	3 (2%)	13 (7%)	120 (63%)	14 (7%)
5th Grade	2 (1%)	0 (0%)	2 (1%)	0 (0%)	12 (6%)	124 (67%)	11 (6%)
All	83 (9%)	15 (2%)	17 (2%)	18 (2%)	64 (7%)	406 (46%)	58 (7%)

Generalization

Students were asked to determine if generalization statements were true or false and then explain why they knew the generalization statement was true or false. For each item, written responses were qualitatively coded. For many of these items, especially for younger students, it was clear that many students did not understand the question or the meaning of the letters. The number of students who either left the item blank or wrote something to the jist of 'I don't know'

was quite high. Some students believed that 'a' stood for 1 because of the alphabetic / numeric code some students learn (i.e., a is 1, b is 2, c is 3, etc.). One 3rd grade student, for example, said: "Everybody kind of thinks that 1 is the first letter of the alphabet".

For example, for the item a-a=0, 68% of 1st grade students either left the item blank or wrote 'I don't know', while an additional 16% wrote something that was incoherent or incomprehensible. One 2nd grade student, for example, said: "I don't know what a-a=0 means". Many students struggled with not understanding the use of a variable (i.e., not understanding what 'a' is). Other students went as far as to understand the meaning of variables but not in how they applied to this scenario. For example, a 3rd grader said: "Because 'a' could be different, a different number than they say because algebra they don't actually have like certain numbers for certain letters".

Similarly to a-a=0, students were also asked to describe why a+0=a was true or not always true. Some students had some early notions about the property of zero: understanding that zero adds nothing. One 1st grade student, for example, said: "Always true because zero is, well you just add nothing, so a plus 0 is always a". Many students remained confused about the entire concept, however. One 3rd grade student, for example, said: "I don't know this one. Are the a's supposed to be zeros or…?" Several students were able to come up with a generalized correct statement. One 4th grade student, for example, said: "Because anything plus zero equals that same number".

Students were also asked to explain early understandings of the commutative property of addition, by explaining why a+b=b+a was true. Like the other problems, many students were confused and answered with "I don't know" when interviewed. More students believed in the alphabet/numeric code. One 2nd grade student, for example, said: "Well a+b would be 3 because

a would be the first in the alphabet and b would be the second". Others still simply didn't know what the letters were. One 4th grade student, for example, said: "We don't know what a and b stands for". Others still seemed to understand at least in terms of one specific example. One 4th grade student, for example, said: "That actually is always true because say a was 9 and b was 7, then 9 plus 7 would be 18, no 17, and 7 plus 9 would be 17 again".

Conversely, students were also asked to explain why the commutative property does not apply to subtraction, when they were asked to explain why a-b=b-a was false. This item seemed to be much more difficult for students to explain than the properties of zero. Some students were confused or believed the statement to be true. One 3rd grade student, for example, said: "Always true because they're the same but backwards". Many students did not understand the possibility of a negative numbers but still had early understandings. One 4th grade student, for example, said: "Because I know that if this was say 7 and this was 6, 7 minus 6 would be 1 and you couldn't do 6 minus 7 so it's not always true". Other students did understand that negative numbers could exist. One 3rd grade student, for example, said: "Not always true because 2 minus 1 is 1 but 1 minus 2 is impossible to do, it's -1". When responding to the a-b=b-a problem, three 3rd grade students, six 4th grade students, and twenty-five 5th grade students (thirty-four students total) mentioned "negative" numbers explicitly. When responding to both a+b=b+a and a-b=b-a problem, five 5th grade students actually referred specifically to the commutative property.

Older students were also asked to apply the property of zero to multiplication problems through understanding that the item ax0=a was false. Some students appeared to be confused in that they provided a generalized statement correctly yet marked the problem as true, therefore it would be interesting to see what students would have said if the problem posed had been ax0=0 instead. One 4th grade student, for example, marked ax0=a as true but then said: "Always true

because everything times zero is zero". Some students understood it in terms of providing an example. One 4th grade student, for example, said: "Not always true because if this was a 4 and that was 0 it always has to end up as zero". Many students did appear to fully understand the generalized statement however. One 4th grade student, for example, said: "That's not true at all because in existence anything times zero equals zero. That's something I learned in 3rd grade".

Older students were also asked about the property of 1 in multiplication when asked to explain why ax1=a is always true. Many students could explain generally why this was the case. One 5th grade student, for example, said: "Everything that's timesed by 1 always ends up with itself so whatever it is, 1, 2, 3, or 4, it will always come out as itself". Another 5th grade student, for example, said: "Because a variable or number times 1 always is the same".

Finally, it is also interesting to note that, across all of the different generalization problems (i.e., a+0=a, a-a=0, a+b=b+a, a-b=b-a, etc.), approximately 40 students believed the problem it referred to a different, previous question: seven 1st grade students, eleven 2nd grade students, ten 3rd grade students, eleven 4th grade students, and one 5th grade students. A 2nd grade student, for example, believed that in the problem a-a=0, the a was equal to 3 because "these are 3's because I checked on it" and showed the problem on the previous page which was 4+a+a=10 so the a equaled 3.

Many of the alternative conceptions match up with what other researchers have discovered: that students may fail to recognize important features of properties, or that students may fail to apply information they know about numbers when variables are used instead (Baroody & Gannon, 1984; Schifter, 1999). The most common response types for generalization items a-a=0, a+0=a, a+b=b+a, a-b=b-a, ax0=a, and ax1=a, as determined through this coding process, are displayed below in Table 44.

Table 44.

Generalization Coded Responses

	Blank	I don't know	Incoherent Statement	Problem Specific Answer	We don't know what a is	Includes examples only	Generalized Correct Statement
a-a=0				a is zero			
1st Grade	**101 (59%)**	15 (9%)	**27 (16%)**	4 (2%)	9 (5%)	5 (3%)	5 (3%)
2nd Grade	**62 (40%)**	**16 (10%)**	**21 (14%)**	9 (6%)	**18 (12%)**	10 (6%)	13 (8%)
3rd Grade	**23 (14%)**	12 (7%)	**28 (17%)**	8 (5%)	**28 (17%)**	13 (8%)	50 (30%)
All	**186 (38%)**	43 (9%)	**76 (15%)**	21 (4%)	**55 (11%)**	28 (6%)	68 (14%)
a+0=a				0 means nothing			
1st Grade	**102 (57%)**	10 (6%)	**42 (23%)**	3 (2%)	10 (6%)	4 (2%)	2 (1%)
3rd Grade	14 (8%)	15 (9%)	**44 (25%)**	13 (8%)	**24 (14%)**	13 (8%)	32 (18%)
4th Grade	**22 (12%)**	14 (7%)	**35 (18%)**	15 (8%)	**19 (10%)**	**20 (11%)**	55 (29%)
All	**138 (25%)**	39 (7%)	**121 (22%)**	31 (6%)	**53 (10%)**	37 (7%)	89 (16%)
a+b=b+a							
2nd Grade	**70 (46%)**	**18 (12%)**	**31 (20%)**	-	6 (4%)	7 (5%)	16 (10%)
3rd Grade	**21 (13%)**	**18 (11%)**	**44 (27%)**	-	15 (9%)	**16 (10%)**	46 (28%)
4th Grade	**27 (14%)**	14 (7%)	**42 (22%)**	-	17 (9%)	**27 (14%)**	63 (32%)
5th Grade	15 (8%)	2 (1%)	**28 (15%)**	-	13 (7%)	**30 (16%)**	87 (48%)
All	**133 (19%)**	52 (8%)	**145 (21%)**	-	51 (7%)	**80 (12%)**	212 (31%)
a-b=b-a				Small can't subtract big			
3rd Grade	**21 (17%)**	12 (7%)	**53 (31%)**	10 (6%)	13 (8%)	**17 (10%)**	33 (19%)
4th Grade	**36 (18%)**	**23 (12%)**	**41 (21%)**	**22 (11%)**	**26 (13%)**	**31 (16%)**	19 (10%)
5th Grade	17 (9%)	7 (4%)	**34 (18%)**	**24 (13%)**	**18 (10%)**	**42 (23%)**	42 (23%)
All	**83 (15%)**	42 (8%)	**128 (23%)**	**56 (10%)**	**57 (10%)**	**90 (16%)**	94 (17%)
ax0=a				a is zero			
4th Grade	15 (8%)	11 (6%)	**26 (14%)**	10 (5%)	13 (7%)	**21 (11%)**	89 (48%)
5th Grade	10 (5%)	4 (2%)	17 (9%)	8 (4%)	11 (6%)	**26 (14%)**	102 (56%)
All	25 (7%)	15 (4%)	**43 (12%)**	18 (5%)	24 (7%)	**47 (13%)**	191 (52%)
Ax1=a							
5th Grade	8 (4%)	1 (1%)	**21 (12%)**	-	9 (5%)	**33 (18%)**	108 (60%)

Chapter 8: Consequences of Testing Validity Evidence

In this chapter I will describe the results obtained from the follow-up teacher survey (i.e., completed by the teacher after the results of the assessment were provided to the teacher). These analyses will be conducted to investigate validity evidence for the consequences of testing (i.e., what consequences occurred because of the use of this diagnostic assessment).

Teachers were asked to complete a follow-up survey after receiving the results of the assessment. Of the 82 participating teachers, 49 completed the follow-up survey, for a response rate of 60%. Because these survey items were Likert in nature and therefore an ordinal scale of measurement, it is not appropriate to report means. Frequencies of each response are therefore reported below in Table 45.

As one can see in Table 45 below, only 12% of the teachers had used their students' results at the time of the survey. Because of the timing of the assessments; the majority were given to students during the last three weeks of school during June; therefore it was not surprising that the majority of teachers said they had not used their students' assessment results. This was expected and an unfortunate consequence of the timing of the study. Future research should examine how this hopefully changes if the assessment is given at a different time of the year. On the other hand, however, the majority of teachers said that they did, in fact, plan to use their assessment results and that they indeed found the assessment results helpful. For example, 69% of teachers planned to use their students' assessment results and a whopping 74% of teachers found the assessment results helpful. Using a Chi-square test, it was discovered that the percentage of teachers that planned to use their students' assessment data was significantly higher than expected, $\chi^2(6, n = 48) = 35.417, p < .001$. A Chi-square test also discovered that the

percentage of teachers that found the assessment results helpful was also significantly higher than expected, $\chi^2(6, n = 49) = 27.714, p < .001$.

In regards to teaching algebraic thinking skills, for example, one teacher said, "I will try to introduce different formats of finding unknown numbers in addition equations" while another said, "I plan to use the results intentionally to teach they concepts they didn't understand". In regards to teaching equivalence specifically, one teacher said, "I talked with my students about how the equal sign is like the middle of a see-saw that has the same weight on each side" while another said "I now plan on doing a lesson about the meaning of the equal sign". Therefore it seems this survey data can be used as early evidence that the diagnostic assessment has the potential to be useful and informative for teachers, and that there do not appear to be negative consequences for using this diagnostic assessment.

Table 45.

Teacher Follow-Up Survey: Frequency Distribution

	1 Strongly Disagree	2	3	4 Neither Agree nor Disagree	5	6	7 Strongly Agree
I have used my students' assessment results.	**12 (25%)**	**13 (27%)**	6 (13%)	**11 (23%)**	3 (6%)	1 (2%)	2 (4%)
I plan to use my students' assessment results.	1 (2%)	2 (4%)	3 (6%)	9 (19%)	**10 (21%)**	4 (8%)	**19 (40%)**
I have found the assessment results helpful.	1 (2%)	2 (4%)	3 (6%)	7 (14%)	**13 (27%)**	**16 (33%)**	7 (14%)

Chapter 9: Discussion, Implications, & Limitations

In this chapter I will discuss the results obtained from this study. These results were reported in the chapters above. Further, the implications of these results will be discussed, as well as the related limitations of this study.

Psychometric Analysis of the Assessment

First, it appears that the interpretation of the results of this diagnostic assessment of algebraic thinking skills for elementary school students may be valid for diagnosing student algebraic thinking alternative conceptions. Through a thorough literature review, well-planned assessment map, the creation of item specifications and test specification, review of the assessments by an expert, etc., it appears that early evidence shows validity for the results of the tool (i.e., that the assessment actually measures algebraic thinking skills).

Item Analysis. Although CTT and IRT statistics revealed several items to be either too difficult or too easy for that particular grade level, given that this is a diagnostic assessment these results seem reasonable. Given that this is a diagnostic tool, however, it appears important to retain such items. Also, the majority of the items appeared to be fairly good discriminators. Further, the majority of items common across vertical versions of the assessment showed increased knowledge by students with age, which supports construct validation claims that algebraic thinking skills improve with age. It does appear that the assessment difficulty, for the most part, does not change substantially as students progress through the grades. This is likely to be expected because of the large number of common items used throughout the grade levels.

Internal Structure Validity Evidence. First, correlations across the dimension strands were statistically significant, indicating that this likely is a unidimensional construct. Second, internal reliability estimates showed the consistency of the tool to be quite high, demonstrating

internal structure validity evidence. Across the 20 versions of the assessments, 70% held internal reliability estimates at 0.80 and above, and the lowest estimate was 0.71, which can still be deemed acceptable by diagnostic standards. Third and finally, alternate equating demonstrated the different versions of the assessments at each grade level to be alternate forms of each other and capable of being used interchangeable, while inter-rater reliability revealed that the items can likely be scored consistently across different raters.

Validity Analyses: Patterns of Student Knowledge

The results of the Washington State sample of the assessment revealed that it does appear that students are experiencing alternative conceptions in algebraic thinking that need to be remedied. In terms of construct-related validity evidence, it appears that the diagnostic assessment items do have the potential to elicit rich alternative conceptions from students. These conceptions were evident in the large sample of students and confirmed through the think-aloud protocols.

Although students had largely mastered open number sentences, many students experienced alternative conceptions with the majority of the other concepts. It does appear that the majority of the learning appears to occur between 1st to 3rd grades. On the majority of the items 1st and 2nd grade students significantly underperformed all other grades. However there were usually no other differences. Although there may appear to be a positive linear relationship in most of the skills, these gains were small between 3rd, 4th, and 5th grades. Sometimes this happened because of a ceiling effect (i.e., the students had mastered the material and had reached at least 80% mastery) such as the case of open number sentences and numerical linear patterns, but in other cases (equivalence) a ceiling is clearly never met. Skills are discussed in more detail below.

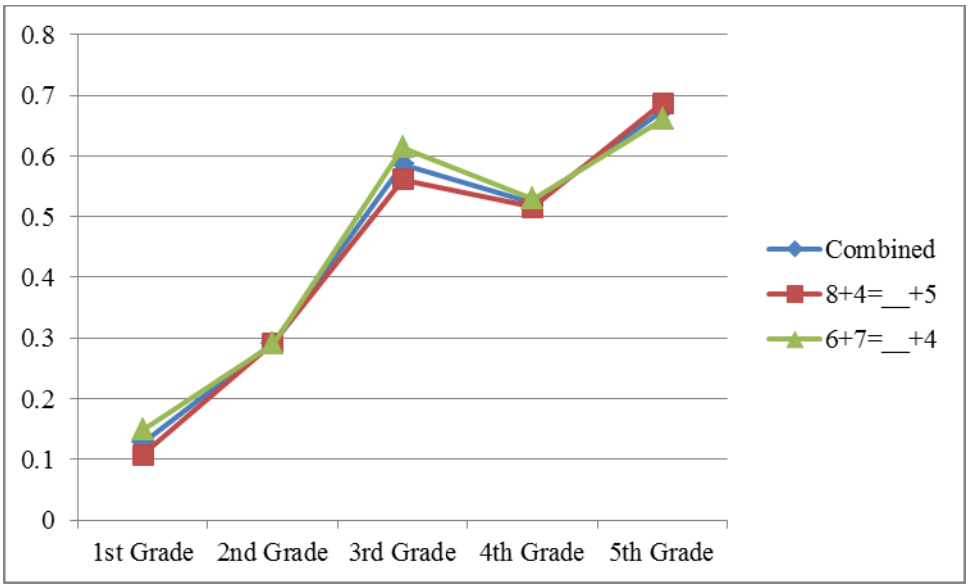
Equivalence. The analysis of the obtained results regarding equivalence specifically discovered two key findings. First, findings corroborated the findings of other researchers (Carpenter et al., 2005; Mann, 2004; McNeil, 2007) determining that students often experience alternative conceptions in the area of equivalence. Although struggles persisted throughout the elementary grades, the numbers of students experiencing these alternative conceptions are much lower than found by other researchers, however. Carpenter and colleagues (2005), for example, found that less than 10% of students in grades 1-6 answered equivalence items correctly. This research found that these numbers were not so low even for 1st grade students. Table 1 below displays these findings. Approximately 11-15% of 1st grade students, 29% of 2nd grade students, 56-61% of 3rd grade students, 52-53% of 4th grade students, and 66-69% of 5th grade students solved equivalence problems like 8+4=__+5 correctly. Clearly, work is needed to improve the conceptions held by students in this area, but these results fail to describe the bleak picture sometimes appearing in the work of others.

The second finding involves failure to corroborate the 'u-shaped' distribution found by other researchers (McNeil, 2007). McNeil (2007) hypothesized that equivalence knowledge declined between the ages of 7 and 9 and began improving again between the ages of 9 and 11. These findings were discovered with a sample size of 87 total students and replicated by a second study with 35 students (McNeil, 2007). Although the current findings did not show perfect linearity, they did show a general linear relationship with students improving from 1st grade (i.e., approximately age 7), through 3rd grade (i.e., approximately age 9) and continued improvement through 5th grade (i.e., approximately age 11). The correlation between solving problems like 8+4=__+5 correctly was in fact statistically significant ($p < .01$) with both age ($r = .351$) and grade ($r = .396$). A one-way analysis of variance (ANOVA) revealed that 1st and 2nd

grade students performed statistically significantly ($p < .001$) differently from all other grades. In addition, 4th grade students performed statistically significantly ($p < .001$) lower than 5th grade students. These findings are further supported by Figure 4 below.

Figure 4.

Proportion of Students Solving Equivalence Items Correctly by Grade



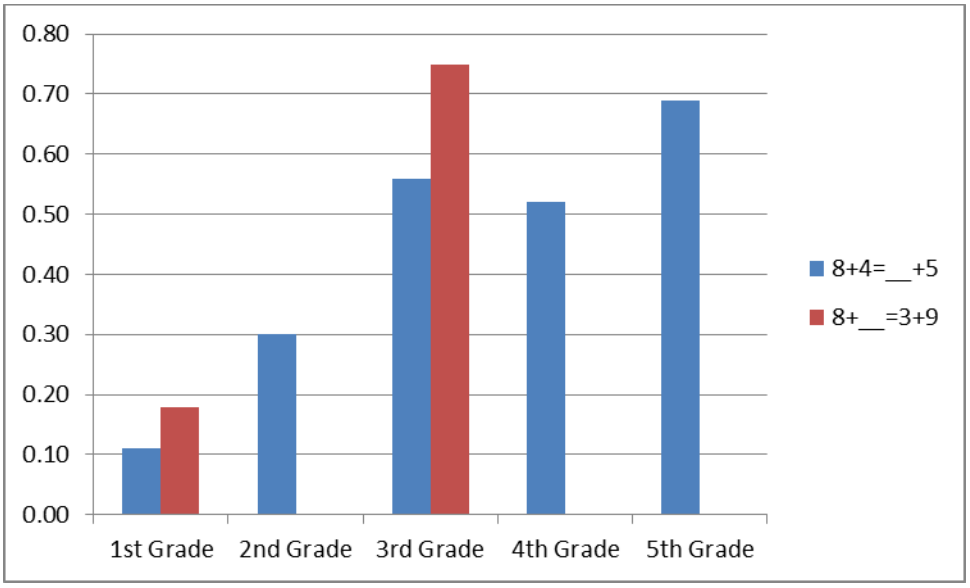
The third finding involves adding to the common alternative conceptions experienced by students when solving equivalence items. Carpenter, Franke, and Levi (2003) believe that three different alternative conceptions exist: the answer is, using all the numbers (i.e., changing the number sentence), and extending the problem. The findings of this study support these researchers' findings. Although less common, several other alternative conceptions also appear to be prominent, however, especially for the youngest students (i.e., 1st and 2nd grade students). These alternative conceptions are often problem dependent. The first of these, for example, is in the case of 6+7=__+4. Some students tended to ignore the first 6 and instead recognize that 7=3+4 (i.e., their final answer was 6+7=3+4). For the problem 8+4=__+5, on the other hand, some students ignored the 4 and recognized that 8=3+5 (i.e., their final answer was 8+4=3+5)

while other students ignored both the 8 and the 5 and recognized that 4=4 (i.e., their final answer was 8+4=4+5). In addition, a very large number of younger students simply left these items blank. These additional conceptions are important to remember when working to remedy such issues in young students.

Fourth, it appears that it may matter where the blank lies. The typical equivalence problem with the blank coming immediately after the equal sign seems to lead students to misunderstanding the meaning of the equal sign more than if the blank is in some other location. Although future research is needed to ensure this is an accurate statement, as this concept was not the primary focus of this research, there are some hints that this may be the case. A very small example is displayed below in Figure 5, in which students in 1st and 3rd grade performed better on items in which the blank was in the 2nd position rather than the 3rd position. Clearly, more research is needed in this area, but this is early evidence that the blank in the 3rd position may encourage students to use such alternative conceptions.

Figure 5.

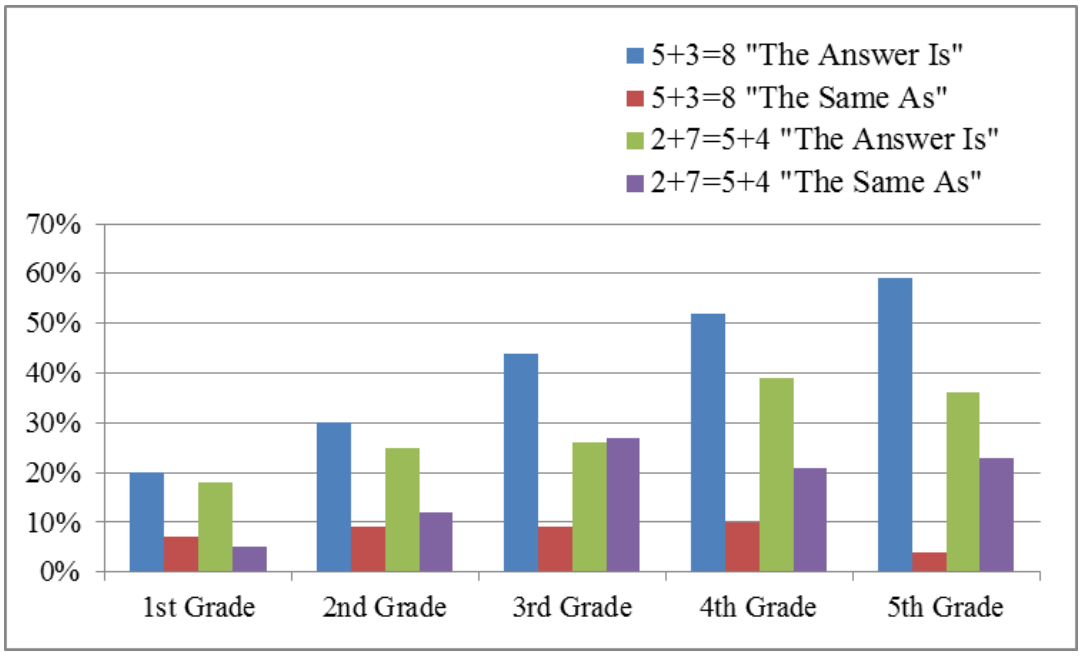
Proportion of Students Solving Different Equivalence Items Correctly



Fifth, other researchers (McNeil et al., 2006) have discovered that student understanding of the meaning of the equal sign may depend on whether the problem posed features an equivalence context or not. This research rather insubstantially bolstered this finding by discovering a statistically significant effect "by context", in that students in the equivalence context were more likely to provide a relational view of the equal sign – indicating it to be "the same as" rather than "the answer is". These results are displayed below in Figure 12. A regression model, however, revealed that the equivalence context was not a unique predictor of student equivalence knowledge after grade was taken into account. There was a small yet significant (r = .208) correlation between holding a relational view of the equal sign and solving the equivalence items like 8+4=___+5 correctly. This value was expected to be larger.

Figure 6.

Proportion of Students Understanding the Meaning of the Equal Sign by Context

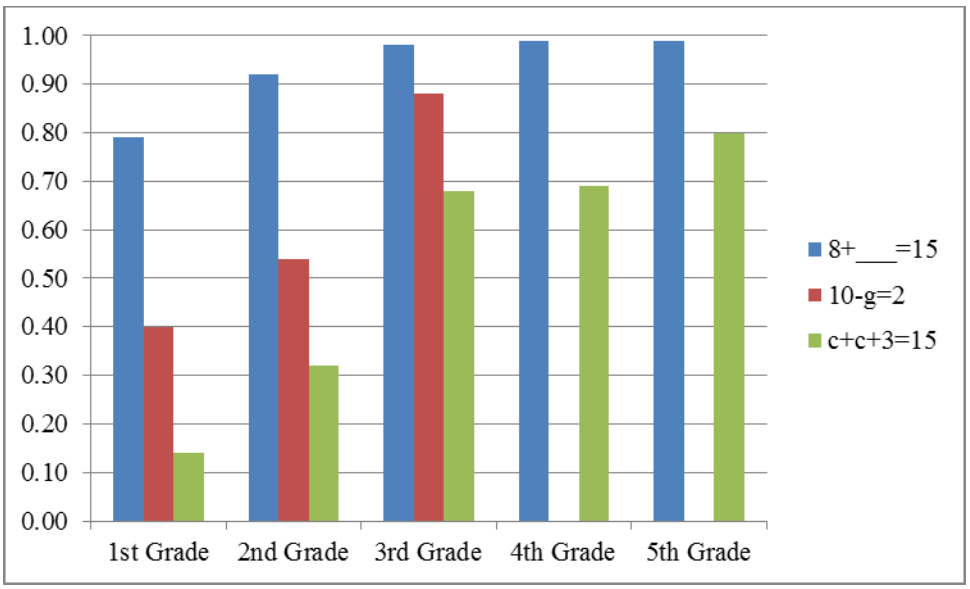


Work with Variables. Students in the U.S. also tended to experience alternative conceptions when solving items with variables. Although these items were very similar to open

number sentences, which they performed quite highly on, students struggled much more when a letter was used instead of a box or blank. This result is highlighted below in Figure 13. This may be because students in the U.S. do not face many opportunities to experience letters as variables. Many students simply reused numbers from the problem in the answer or even utilized the numeric / alphabetic code students sometimes learn (i.e., a=1, b=2, c=3, etc.). When problems became more complex and two identical variables were seen (i.e., c+c+3=15), students did not seem to understand that when the variable is the same and used more than once the same number will replace both letters.

Figure 7.

Proportion of Students Solving Open Number Sentences and Variables Items Correctly



It does appear from these results above in Figure X that students in the U.S. might benefit from more opportunities to experience variables.

Generalizing. Many students, but especially younger students, simply did not understand the generalizing items. This is understandable, and is one of the most difficult items that should likely be removed from 1st and 2nd grade assessments in the future. This may be

likely related to the fact that many students didn't understand the work with variables items as discussed above. This is likely a prerequisite skill to understanding that the letters in the generalization actually stand for numbers, and actually can stand for any number. Many students simply left this item blank or circled "I don't know". Although interesting diagnostic information was still obtained, these items are perhaps not appropriate until 3rd grade.

Figural Patterns. Like generalizing, students likely have not had many opportunities to experience figural patterns. Although numerical patterns and repeating shape patterns are common occurrences in elementary schools, and students performed highly accordingly, utilizing growing figural patterns is a more complex algebraic thinking skill not commonly utilized. Students' common experience with repeating patterns was demonstrated when many students believed the figural pattern to be a repeating pattern as well. Others simply left this item blank, while others still had the beginnings of understanding and simply failed to notice all the different parts of the shape that were growing. This is unfortunate given that researchers (Warren & Cooper, 2008) have discovered that students may benefit from more opportunities to experience figural patterns in elementary classrooms.

Validity Analyses: Singapore vs. U.S. Student Comparisons

There were several key findings regarding comparing the results of students in Singapore to those of students in the U.S. First, it appears that students in Singapore outperformed students in the U.S. on the algebraic thinking skills diagnostic assessment overall. This supports construct-related validity claims because the assessment should reveal group differences when the groups are different (i.e., such as in the case in students in the U.S. versus students in Singapore). These group differences may because of Singapore's general outperformance in all mathematical areas, because of the country's use of 'The Model Method', or some other reason

altogether. Teachers interviewed as part of this study indicated the competitive nature of Singaporeans, providing additional support outside of the classroom, and a "packed" syllabus as the most popular reasons for Singapore's success. Interestingly enough, several teachers were not sure as to the reason; they thought the education system lacked uniqueness. These themes resulting from the teacher interviews are displayed below in Table 46.

Table 46.

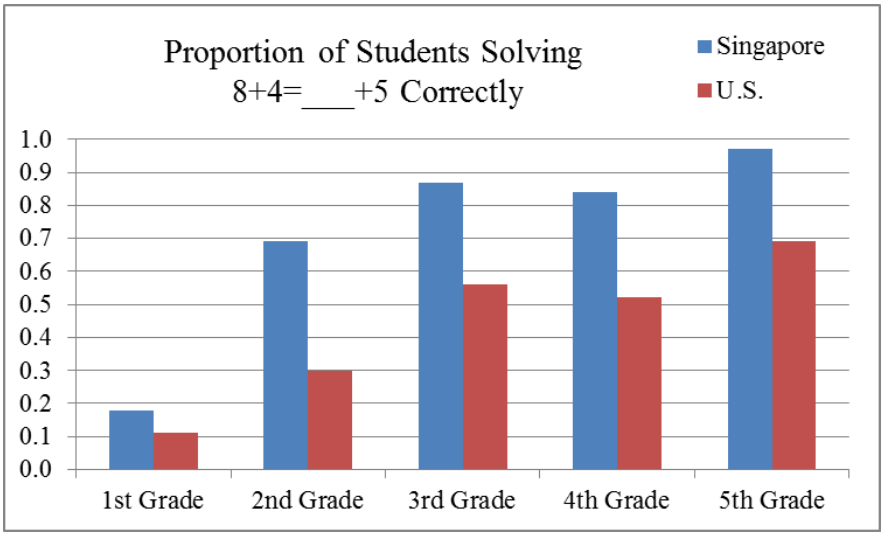
Teacher Interview Responses to Why does Singapore perform so well on the TIMSS?

Theme	Teacher Statements
Competitive Environment	"Have you heard of something called "*kiasu*"? This is the idea that we don't want to miss out, so the people here are competitive. There are no natural resources here so people are our resources." "I think in Singapore many parents emphasize education, it's a slightly competitive environment in schools." "Students in Singapore are driven – they really want to perform highly on tests." "There is a strong competitive environment, many families and parents who are well-educated may emphasize the importance of learning."
Additional Support	"Parents want students to succeed so badly that they get them tutoring, teach concepts earlier than they're supposed to, or whatever it takes." "This is a meritocracy society, so parents are anxious over their child's grades and so they give them extra classes and groom them." "Many parents in my opinion may sort of provide pupils with additional support. Tutors and tuitions and so on." "Many of them have home support, either tutors, tuition, enrichment activities, so I don't think it's just what goes on in the classroom… it wouldn't be a fair statement for me to stay that my students are doing well because of what I teach in the class."
Syllabus	"Is it because of the syllabus? I find that our syllabus is very tight. We have so many topics to cover and not enough time to slow down to go deeper. So perhaps our syllabus, the sums that our children go through are of a higher standard, perhaps that's why they can do better." "There is a lot of rote learning." "I think it comes down to drill and practice, right? Drill and practice." "I think our syllabus is pretty packed. So you can say that almost everyday you go back home there's homework to do. So I guess it's our syllabus."
Control	"The U.S. is too big. In Singapore there is just the MOE [Ministry of Education], they have much control, and there is just one NIE [university to train teachers], so it is easy to make changes because it is so small and it is all funded by the same people. In the U.S. there are too many stakeholders."
Test Scores Aren't Everything	"High test scores may not represent everything. Scoring very high on an exam does not mean I will be able to be creative in that field." "The education system has definitely played a part, but again you know, I'm not sure what exactly you're referring to as the system being successful because if you're referring to success and how is our success international, but I mean if you are saying they have done well on the TIMSS, why have they done well on the TIMSS, then my next question is who has sat for the TIMSS. So if you have done well on the TIMSS what does that show about the pupil's ability and mastery and what does it not show."
Not Sure	"Actually to be honest I don't see anything special about the way we teach, so how come can we perform well?" "I don't know, I have no idea."

Second, it appears that students in Singapore may experience some of these same alternative conceptions in 1st and 2nd grade regarding equivalence that students in the U.S. experience. Unlike in the U.S., however, these alternative conceptions appear to have all but disappeared by 3rd grade. The results of student answers to the equivalence item 8+4=__+5 in both Singapore and the U.S. are displayed below in Figure 14.

Figure 8.

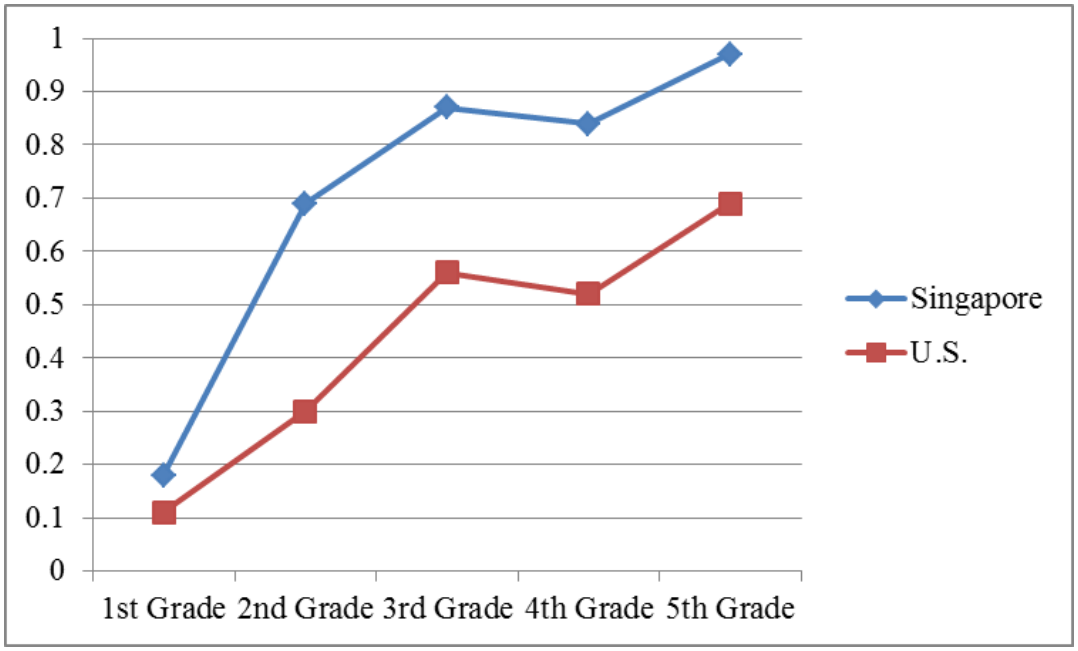
Proportion of Students Solving Equivalence Items Correctly in Singapore and the U.S.



Third, again, like in the U.S., the Singapore results did not corroborate the 'u-shaped' distribution found by other researchers (McNeil, 2007). Interestingly enough, an identical pattern occurred both in the U.S. and in Singapore: with student results dropping slightly in between 3rd and 5th grade instead of a perfectly linear pattern. These results are displayed in Figure 9 below.

Figure 9.

Pattern of Students Solving Equivalence Items Correctly in Singapore and the U.S.



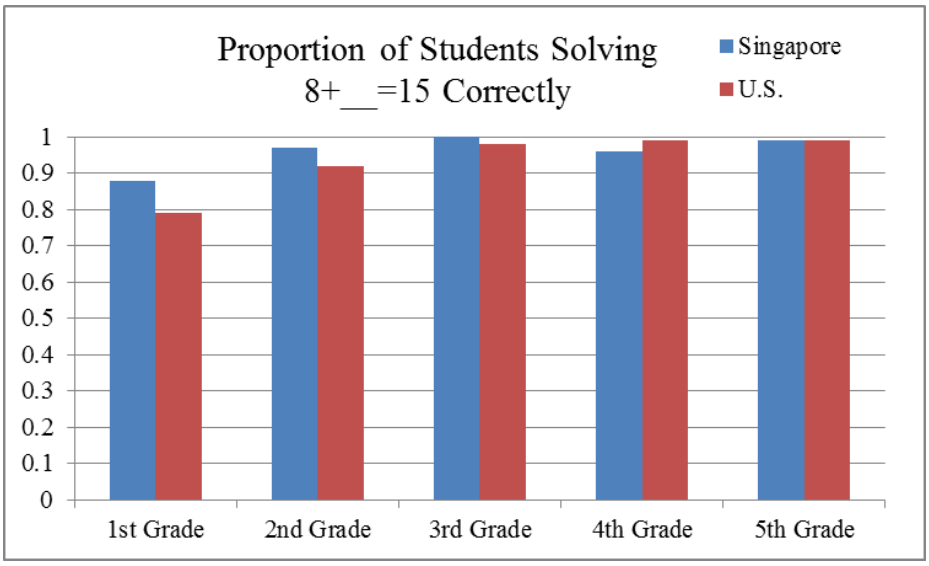
Fourth, my findings were similar to those discovered by the only other cross-country comparison study that investigated the algebraic thinking skills in middle school students. Li and colleagues (2008) conducted an international investigation of student equivalence ability with 6th grade students in China (n = 145) and the U.S (n = 105). While 98% of the Chinese students solved equivalence problems such as 6+9=a+4 proficiently only 28% of U.S. students were able to do so. Both Li and colleagues' (2008) study and this study found that the comparison country greatly outperformed U.S. students in terms of equivalence knowledge.

Fifth, it is interesting to note where the differences lie between the two countries. For instance, while Singapore and U.S. students performed similarly on open number sentences and linear numerical patterns (i.e., problems that are strongly arithmetic in nature), Singapore students largely and significantly outperformed U.S. students on other items such as equivalence, nonlinear numerical patterns, etc. An example of this is displayed below in Figure 16. This

raises the question, what is it about these tasks that causes this? Are these skills emphasized more in Singapore than they are in the U.S.? Does use of the "model method" come into play? Do these skills better prepare students to succeed in middle school, high school, and beyond? Do these skills better prepare students to succeed in "official" algebra courses? These are important questions that should be explored in further research.

Figure 10.

Proportion of Students Solving Open Number Sentence Items Correctly in Singapore and the U.S.



Validity Analyses: Teachers' Perception of Diagnostic Values of the Assessment

The results from this work show that it appears teachers may be able to use the data from this tool to remedy algebraic thinking alternative conceptions in their elementary classrooms. Although because of the timing of the administration of the assessment (i.e., in the last month of the year) most teachers had not used the results at the time of the teacher survey, the majority of the teachers did find the results helpful. The majority of teachers also intended to use the results, either in terms of teaching to concepts not mastered by their students, re-teaching the meaning of

the equal sign, introducing variables instead of just boxes, etc. More research is needed in this area to ensure teachers really do use the data in the ways they mentioned. Future research should conduct studies such as these at the beginning of the year to allow teachers time to use the results in their current classrooms.

Limitations of the Study

There are several major limitations associated with this research that must be considered. First, the overall sample was too small, too narrow, and not diverse enough. Unlike the majority of research discovered in the literature review, the U.S. sample utilized a stratified random sample of schools. Although this is a major benefit, unfortunately a cluster sample was also utilized to reduce costs. Cluster problems arise because students are nested within teachers which are nested within schools, and students within one teacher are more likely to be similar to each other just as students within one school are more likely to be similar to each other. A cluster sample can result in decreased precision; however this issue can be reduced if the sample size of clusters is large. In this case the sample size of clusters was quite small ($n = 6$), therefore this is definitely a major limitation to consider. Further, the sample of schools was too narrow in that it only sampled schools in King and Pierce counties of Washington State. Future studies should sample students across Washington State if not nation-wide, and should also emphasize the inclusion of students in rural areas. The Singapore sample was also limited in similar and additional ways. Unlike the Washington State sample, the Singapore sample was a convenience sample. A random sample was unfortunately not possible due to logistic issue associated with conducting research internationally. Like the Washington State sample, the results of the Singapore sample were also clustered across only four schools, therefore the same issues associated with a cluster sample also apply, however in this case the sample was not stratified to

ensure students of varying achievement levels were included. Because Singapore utilized ability grouping by school, it is unknown if the four participating schools were representative of the general population or not. Although they were considered 'neighborhood' schools and did not have a selection process to attend the schools, it is unknown if these results are either overestimating or underestimating actual abilities of students in Singapore. Because of these many limitations, future studies in both the U.S. and Singapore should utilize larger, randomized samples across a larger number of schools.

Second, the various versions of the assessment in Singapore were unfortunately not evenly or randomly distributed within classrooms as they were in the Washington State sample. This was due to logistical issues again associated with the lack of control that occurred when conducting research internationally. This issue accentuates the cluster issues previously discussed because of the fact that Singapore not only ability-groups by school but also ability-groups by classroom. Because of this ability grouping it cannot be known if this is a representative sample of all Singaporean ability levels. Logistical issues such as these need to be overcome prior to conducting future research.

Third, it is important to note that it is likely that the statistics reported in this study may be overestimating student ability. Because of the method used to deal with missing data (i.e., once a pattern of missing data was reached all data was coded as not reached and removing it from subsequent analysis). Had this data been coded as incorrect instead of not reached it may have shown lower student abilities. It is therefore likely that individual item values are likely at least slightly underestimated. As a more conservative method for dealing with missing data, however, it was determined important to err on the side of being conservative (i.e., overestimating student ability) instead of being too liberal (i.e., underestimating student ability).

Fourth, it is unknown if the teachers administered the assessment correctly. Because teachers administered the assessment instead of the researcher, it is unclear if the teacher used the standardized administration directions, if they allowed students enough time to finish, and how much they did or did not help students with the assessment items during the administration. It is also unknown if students were 'primed' for the assessment; in that if teachers saw the assessments and decided to teach a few of the concepts prior to administering the assessment. Future studies should improve standardization strategies by having the researcher administer the assessment.

Fifth and finally, one must be conservative in labeling this assessment one measuring algebraic thinking. Although many measurement procedures have been followed to collect construct evidence, it is difficult to measure such a complex and abstract skill through a paper / pencil assessment alone. A short, quick, paper / pencil assessment was desired in order to be convenient and easy for teachers to implement, however this method does have drawbacks in the types of information that may be collected and these drawbacks must be considered.

Implication for Future Research and Practices

There are many implications for future research arising from this work. First, it is clear from the limitations section that future research should definitely use wider, larger, and more randomized samples across a greater geographic area and should include more schools, teachers, and students. If cluster-sampling continues to be used, a greater number of clusters should be used, or else better less-biased sampling techniques such as stratified or simple random sampling should be used. Further, although other aspects of reliability claims such as test-retest samples could not be conducted at this time because of logistic issues, this is certainly an interesting and

useful avenue for future research. Researchers should also investigate the use of the alternate forms with a time delay.

Second, it appears very important to continue collecting construct-related validity evidence. Future research should investigate how alternative conceptions held by students in elementary schools change over time as students enter official algebra courses in middle and high schools. Does possessing these alternative conceptions in fact affect student learning of algebra in secondary schools? A longitudinal study tracking students across the grade levels should be conducted to investigate this further. Construct validity evidence should also be investigated further through the implementation of an intervention. Students should be given the assessment as a pre-test, given an intervention designed to teach algebraic thinking skills, and then given the assessment as a post-test. If evidence of construct validity can be collected, then students should improve on the assessment after receiving intervention on the subject. Finally, construct validity evidence should also be investigated by conducting Differential Item Functioning (DIF) studies, in comparing how various sub-groups of students performed on the diagnostic assessment. These studies should include DIF comparisons of Singapore to U.S. students, ELL to non-ELL students, male to female students, etc.

Third, future research should explore further the use and consequence validity claims of this assessment based on empirical evidence rather than self-report. Will the use of such an assessment tool alone improve student algebraic thinking skills? An experimental design where teachers are randomly assigned to either giving the assessment tool or not could be used to investigate this. Or, in contrast, is teacher professional development needed in tandem with use of the tool to truly impact student learning? How do such alternative conceptions develop and

what are the best methods for remedying them? These questions and more should be investigated through empirical evidence in the future.

Fourth and finally, several conceptual issues should be investigated further. First, it is necessary to test the attributes underlying the alternative conceptions with statistical models specific for CDA. To support validity claims, CDA software should be used in the future to extract trait scores based on the patterns of responses to the items. Further, future research is needed to validate the Algebraic Skill Hierarchy (see Figure 4 in Chapter 4). This figure is merely a hypothetical framework at this time that should be tested in future work.

It is clear that the learning of algebraic thinking skills in elementary schools may be a crucial aspect so succeeding in algebra in secondary schools. Although the link hasn't been directly discovered yet, and future research is certainly needed in this area, it appears that experiencing alternative conceptions in elementary schools could be affecting student learning of algebra in secondary schools.

Now that the diagnostic assessment tool has revealed that elementary school students are, in fact, experiencing alternative conceptions, the next question may be 'what can be done about it?' Some researchers have found that professional development can help teachers improve student outcomes. Several researchers, for example, have investigated the results of professional development (which likely increases teacher pedagogical content knowledge) on student algebraic thinking skills. Jacobs and colleagues (2007), for example, investigated the effects of an algebraic thinking based professional development program (based on Carpenter et al., 2003) on 180 teachers of grades 1-5 and their associated 3,735 students, and found that teachers who received the professional development significantly increased their pedagogical content knowledge of algebraic thinking skills. Further, not only did teacher knowledge improve but

their subsequent students performed significantly better on tests of equivalence (but not better on work with variables or generalization) than students whose teachers did not receive the professional development. Students in 1st grade whose teachers received the professional development performed significantly better on simplifying tasks than students whose teachers did not receive the professional development; however no differences existed at the other grade levels. Carpenter, Levi, Berman, and Pligge (2005) found similar results in professional development they provided to 15 elementary school teachers and their accompanied students. Prior to the professional development, less than 10% of the students in grades 1 to 6 could successfully answer the problem 8+4=__+5. Following teacher professional development, in which teachers were taught lessons and strategies to implement to help students better understand equivalence, 66% of students in grades 1 and 2, 72% of students in grades 3 and 4, and 84% of students in grades 5 and 6 could successfully answer the problem 8+4=__+5.

Several researchers have also conducted interventions specific to improving algebraic thinking skills and have discovered more direct associations. Powell and Fuchs (2010), for example, found that instruction focusing on equivalence improved equivalence skills in 3rd grade students with mathematics difficulty, while Carpenter and Levi (2000) found success by providing an intervention to students focused on analyzing true and false number sentences. Rittle-Johnson and Alibali (1999), on the other hand, implemented conceptual and procedural interventions with 4th and 5th grade students regarding equivalence knowledge and discovered that conceptual and procedural knowledge regarding equivalence influence each other (i.e., increased conceptual understanding can increase procedural knowledge and increased procedural understanding can increase conceptual knowledge), although their findings indicated that increasing conceptual knowledge may reap more benefits.

These findings demonstrating success interventions have on increasing student algebraic thinking skills have been replicated using an intervention on variables (Kuchemann, 1981; Booth, 1984), an intervention focused on comparing symbols (Hattikudur & Alibali, 2010), an intervention on figural patterns (Warren & Cooper, 2008), and interventions emphasizing conceptual knowledge (Matthews & Rittle-Johnson, 2008; Perry, 1991). Other hypothesized interventions for algebraic thinking skills include allowing students opportunities to evaluate the relationships between numbers (i.e., using the greater than, less than, and equal to symbols) (McNeil et al., 2006), introducing the equivalence topic non-symbolically first prior to moving to symbolic representations (Sherman & Bisanz, 2009), providing more experiences with the 'operations on both sides' context (McNeil et al., 2006), introducing the concept of equivalence as a balancing scale such as by utilizing a see-saw (Mann, 2004), and utilizing the phrase "is the same as" instead of "equals" when students read number sentences (Van de Walle, 2010).

Conclusions

In conclusion, it appears that the results of the diagnostic assessment of algebraic thinking skills can lead to valid and reliable score interpretations for diagnosing student alternative conceptions. This work provided empirical evidence for the assessment tool regarding reliability, test content validity, internal structure validity, construct-related validity, consequences of testing validity, and response process validity.

Use of the tool revealed that a large number of students in Washington State appear to be struggling with algebraic thinking skills. Many such students are experiencing alternative conceptions in various areas, including equivalence, that would benefit from remediation. Although students in Singapore outperformed students in the U.S. on the majority of algebraic thinking items, it did appear that students in Singapore also experienced alternative conceptions

in algebraic thinking in specific areas, particularly in equivalence in the early grades. Results of a teacher survey revealed that the majority of participating teachers not only found the assessment results helpful but also planned to use the assessment results to remedy alternative conceptions. These preliminary results demonstrate that this tool may have the potential to help teachers better understand the algebraic thinking alternative conceptions experienced by their students, which will ideally lead to teachers remedying such alternative conceptions, which will ideally lead to better algebraic thinking skills, which will ideally lead to better algebra results in official courses in secondary schools.

Appendix

Appendix A: Factors Influencing Algebraic Thinking

In the following section I will discuss some of the factors that predict or are impacted by algebraic thinking skills or mathematical ability in general. These factors should be discussed as they are extremely crucial in terms of making validation claims for the algebraic thinking assessment tool. Throughout the following section, it will be generally inferred that predictors of general mathematical ability will also predict algebraic thinking skills. General mathematical ability likely influences algebraic thinking skills and vice versa. A student who cannot perform basic addition tasks, for example, obviously is not capable of performing addition tasks featuring equivalence tasks. These factors that influence and are influenced by algebraic thinking will be organized into a nomological net, which may be viewed below in Figure 11. Developing a nomological net is important in work towards a potential diagnostic assessment because, as stated by Cronbach and Meehl (1955) in their seminal piece, "to validate a claim that a test measures a construct, a nomological net surrounding the concept must exist" (Cronbach & Meehl, 1955, p. 291).

Demographic Variables

There are many student demographic variables that predict student mathematics knowledge. These variables are not causal, however, but merely predictive, and none of these variables could be manipulated by a researcher or changed in any way. These demographic variables impacting mathematics achievement include socioeconomic status, gender, ethnicity, number of unexcused absences, student health conditions, being an English Language Learner (ELL), parent's educational expectations, parent education, (Byrnes & Wasick, 2009; Okpala et al., 2000; Gottfried, 2009; Garcy, 2009; Han & Bridglall, 2009). Because these variables are not

manipulatable, however, ‘demographic variables’ are only included in the nomological net as unimportant variables.

Working Memory, Processing Speed, & General Intelligence

Higher levels of working memory, including the central executive, the phonological loop, and the visuospatial sketch pad; processing speed; and general intelligence have long been associated with higher mathematics ability (Geary, 2011). Because these variables are not manipulatable, however, they are only included in the nomological net as unimportant variables.

Reading Ability

Several researchers have hypothesized that reading ability strongly predicts mathematics ability (Jiban & Deno, 2007). Others have stated that mathematics ability is a multidimensional construct, with reading ability being the second construct (Walker, Zhang, & Surber, 2008). This occurs at least within the methods of measuring mathematics ability, because many mathematics achievement tests require a substantial level of reading ability to fully understand the question being asked. Walker and colleagues (2008) discovered that this multidimensionality caused underestimation of mathematics ability in cases where reading ability was low. Other researchers have discovered that students perform significantly worse on mathematics items with reading levels above the students’ grade level (Lamb, 2010). Further, it also seems likely that mathematics ability may also predict reading ability, as student academic skills may overarch both reading and mathematics ability. Therefore this variable will be represented in the nomological net as a two-way variable; meaning it is both influenced by and causal of algebraic thinking (represented by a dotted line because the relationship is with mathematics, not algebraic thinking). This multidimensionality of the construct should be minimized in a potential diagnostic assessment by minimizing the reading requirements.

Self-Efficacy

Pajares and Graham (1999) discovered that a student's mathematics self-efficacy (i.e., confidence to succeed in mathematics tasks) was a statistically significant predictor of mathematics ability for 6th grade students. Fast and colleagues (2010) also examined the math self-efficacy of upper elementary school students. They found that those students who thought their classroom environments were challenging, caring, and mastery-oriented self-reported higher levels of math self-efficacy and that higher math self-efficacy predicted mathematics ability. Therefore this variable was represented in the nomological net as causing algebraic thinking (represented by a dotted line because the relationship is with mathematics, not algebraic thinking).

Math Curriculum

A plethora of research has been conducted on a variety of math curriculums. It is clear that using a specific math curriculum may have the potential of improving student mathematics abilities. Waite (2000), for example, analyzed the effects of the *Everyday Mathematics* curriculum on mathematics ability of elementary school students and discovered a statistically significant positive effect, however the What Works Clearinghouse determined this effect was not statistically significant but was still an 'important positive effect'. Similarly, DiLeo (2007) analyzed the effects of the *Odyssey Math* intervention (i.e., a 90 minute per week supplement to the core program) on mathematics ability of 5th grade students and found a statistically significant positive effect on the Pennsylvania System of School Assessment (PSSA). Again, the What Works Clearinghouse determined this effect was not statistically significant but was 'substantively important'. A plethora of other research in this area shows positive, mixed, and no effects based on the curriculum and research design utilized. Because of this research this

variable is displayed in the nomological net as causing algebraic thinking (represented by a dotted line because the relationship is with mathematics, not algebraic thinking).

Math Intervention

It appears that using particular mathematics interventions may also improve student mathematics abilities. In terms of general mathematics ability, Baker, Gersten, & Lee (2002) conducted a synthesis on the current research available on teaching mathematics to low ability students across the grade levels. Interventions that best improved student math ability included supplying teachers and student with student performance data, dispensing feedback to parents of student data, utilizing peer tutors, and using explicit instruction when teaching mathematics. Several researchers have also conducted interventions specific to improving algebraic thinking skills and have discovered more direct associations. Powell and Fuchs (2010), for example, found that instruction focusing on equivalence improved equivalence skills in 3rd grade students with mathematics difficulty, while Carpenter and Levi (2000) found success by providing an intervention to students focused on analyzing true and false number sentences. Rittle-Johnson and Alibali (1999), on the other hand, implemented conceptual and procedural interventions with 4th and 5th grade students regarding equivalence knowledge and discovered that conceptual and procedural knowledge regarding equivalence influence each other (i.e., increased conceptual understanding can increase procedural knowledge and increased procedural understanding can increase conceptual knowledge), although their findings indicated that increasing conceptual knowledge may reap more benefits.

These findings demonstrating success interventions have on increasing student algebraic thinking skills have been replicated using an intervention on variables (Kuchemann, 1981; Booth, 1984), an intervention focused on comparing symbols (Hattikudur & Alibali, 2010), an

intervention on figural patterns (Warren & Cooper, 2008), and interventions emphasizing conceptual knowledge (Matthews & Rittle-Johnson, 2008; Perry, 1991). Other hypothesized interventions for algebraic thinking skills include allowing students opportunities to evaluate the relationships between numbers (i.e., using the greater than, less than, and equal to symbols) (McNeil et al., 2006), introducing the equivalence topic non-symbolically first prior to moving to symbolic representations (Sherman & Bisanz, 2009), providing more experiences with the 'operations on both sides' context (McNeil et al., 2006), introducing the concept of equivalence as a balancing scale such as by utilizing a see-saw (Mann, 2004), and utilizing the phrase "is the same as" instead of "equals" when students read number sentences (Van de Walle, 2010). Because of this research this variable is displayed in the nomological net as causing algebraic thinking skills.

Teacher Education & Experience

While it seems obvious that the teacher the student has would play a large role in mathematics ability and therefore algebraic thinking skills, research findings in this area have been surprisingly mixed. Goddard, Hoy, and Woolfolk (2000) have begun work on measuring teacher collective efficacy (i.e., the beliefs that teachers can positively affect student ability) and have found it to be positively associated with elementary school student mathematics ability. Further, utilizing a state-wide administrative dataset, Kukla-Acevedo (2009) found a relationship between the undergraduate GPA of the teacher (i.e., overall, math education, and math) and course hours (i.e., math and math education) to predict the mathematics ability of 5^{th} graders. Okpala and colleagues (2000) found that teachers with Master's Degrees and teachers with 10 years of teaching experience or more significantly predicted higher mathematics ability of 4^{th}

grade students. Because of this research this variable is displayed in the nomological net as causing algebraic thinking.

Teacher Pedagogical Content Knowledge

Although Shulman (1986) first termed the phrase pedagogical content knowledge (PCK) and recognized it as a distinctive domain of knowledge needed by teachers, Ball, Thames, and Phelps (2008) have since developed a theoretical framework that elaborates on this idea. In Ball and colleagues' framework, PCK is separated into knowledge of content and teaching (KCT) and knowledge of content and students (KCS). KCT was defined as "knowledge that combines knowing about teaching and knowing about mathematics" (Ball et al., 2008, p. 401) and includes designing effective instruction, sequencing concepts appropriately, and choosing suitable examples. KCS was defined as "knowledge that combines knowing about students and knowing about mathematics" (Ball et al., p. 401) and includes anticipating common conceptions and misconceptions, understanding what concepts will cause students confusion, and knowing what students will find motivating and appealing. This research is particularly concerned with KCS, in discovering the knowledge teachers have of their student knowledge of algebraic thinking topics (e.g., knowing if students have misconceptions surrounding algebraic thinking) and if knowing about such misconceptions through a diagnostic assessment tool will enhance learning. It has been made fairly clear throughout this document that teachers need this specialized knowledge about teaching algebraic thinking in the elementary grades to address misconceptions held by elementary students, as it is known that teachers sometimes fail to recognize the misconceptions their students possess. Although much research has been conducted on PCK in a variety of content areas including English (Grossman, 1989), Social Studies (Wilson & Wineburg, 1988), and Mathematics (Ball, 1990; Ma, 1999), very few research studies exist

focusing specifically on algebraic thinking. Asquith and colleagues (2007), have begun work in this area, discovering that while teachers were fairly capable of predicting middle school variable knowledge, they were largely unable to correctly predict student knowledge concerning the equal sign.

Stephens (2006) has also begun work in this area by investigating the knowledge of student algebraic thinking alternative conceptions held by pre-service teachers. Through qualitative interviews with 30 pre-service elementary teachers, Stephens showed teachers five algebraic thinking tasks that teachers might pose to students to explore algebraic thinking topics, asked them reasons teachers would have for posing such tasks, and interviewed them on strategies students might employ when solving the tasks. Examples of tasks included "What number goes in the __? 37+54=__+55" and "The solution to the equation 2n+15=31 is n=8, what is the solution to the equation 2n+15-0=31-9?" (p. 257). Results varied depending on the task, but Stephens found that many of the pre-service teachers failed to recognize the tasks as having the potential to address algebraic thinking alternative conceptions. This study concluded that it is possible pre-service teachers need further PCK development in the area of algebraic thinking skills. Further, Stephens (2008) also discovered that the majority of these pre-service teachers held a very limited view of algebraic thinking, interpreting algebra to mean solely symbol manipulation. Although research has yet to be conducted analyzing the explicit link between teacher pedagogical content knowledge of algebraic thinking and student skill, Hill, Rowan, and Ball (2005) have conducted research on whether teacher PCK affected elementary school student general mathematics ability. They discovered a statistically significant relationship between teacher mathematical PCK and 1st and 3rd grade student math ability gains.

Professional Development

Several researchers have investigated the results of professional development (which likely increases teacher pedagogical content knowledge) on student algebraic thinking skills. Jacobs and colleagues (2007), for example, investigated the effects of an algebraic-thinking based professional development program (based on Carpenter et al., 2003) on 180 teachers of grades 1-5 and their associated 3,735 students, and found that teachers who received the professional development significantly increased their pedagogical content knowledge of algebraic thinking skills. Further, not only did teacher knowledge improve but their subsequent students performed significantly better on tests of equivalence (but not better on work with variables or generalization) than students whose teachers did not receive the professional development. Students in 1st grade whose teachers received the professional development performed significantly better on simplifying tasks than students whose teachers did not receive the professional development; however no differences existed at the other grade levels. Carpenter, Levi, Berman, and Pligge (2005) found similar results in professional development they provided to 15 elementary school teachers and their accompanied students. Prior to the professional development, less than 10% of the students in grades 1 to 6 could successfully answer the problem 8+4=__+5. Following teacher professional development, in which teachers were taught lessons and strategies to implement to help students better understand equivalence, 66% of students in grades 1 and 2, 72% of students in grades 3 and 4, and 84% of students in grades 5 and 6 could successfully answer the problem 8+4=__+5.

Because of this research this variable (pedagogical content knowledge) is displayed in the nomological net as causing algebraic thinking.

Factors that Student Mathematical Knowledge Predicts

Several major factors are caused by algebraic thinking. As discussed previously, it is inferred that what is caused by mathematical ability in general would also be caused by algebraic thinking.

Science Ability

Although others have discovered the positive relationship between mathematics ability and science ability for high school and college students (Gustin & Corazzo, 1994; Wang, 2005), Maerten-Rivera and colleagues (2010) have discovered this relationship occurs as early as the elementary level (i.e., 5th grade students). Therefore this variable is displayed as caused by algebraic thinking skills in the nomological net (represented by a dotted line because the relationship is with mathematics, not algebraic thinking).

Algebraic Thinking Skills in the Later Grades

As discussed previously, depriving students of opportunities to experience algebraic thinking in the elementary grades may make learning algebra later (e.g., in middle and high schools) more difficult (Smith, diSessa, & Roschelle, 1994). Therefore official algebra knowledge in the later grades is likely to be caused by algebraic thinking in the elementary grades. This is represented in the nomological net by a dotted line because this is an assumption.

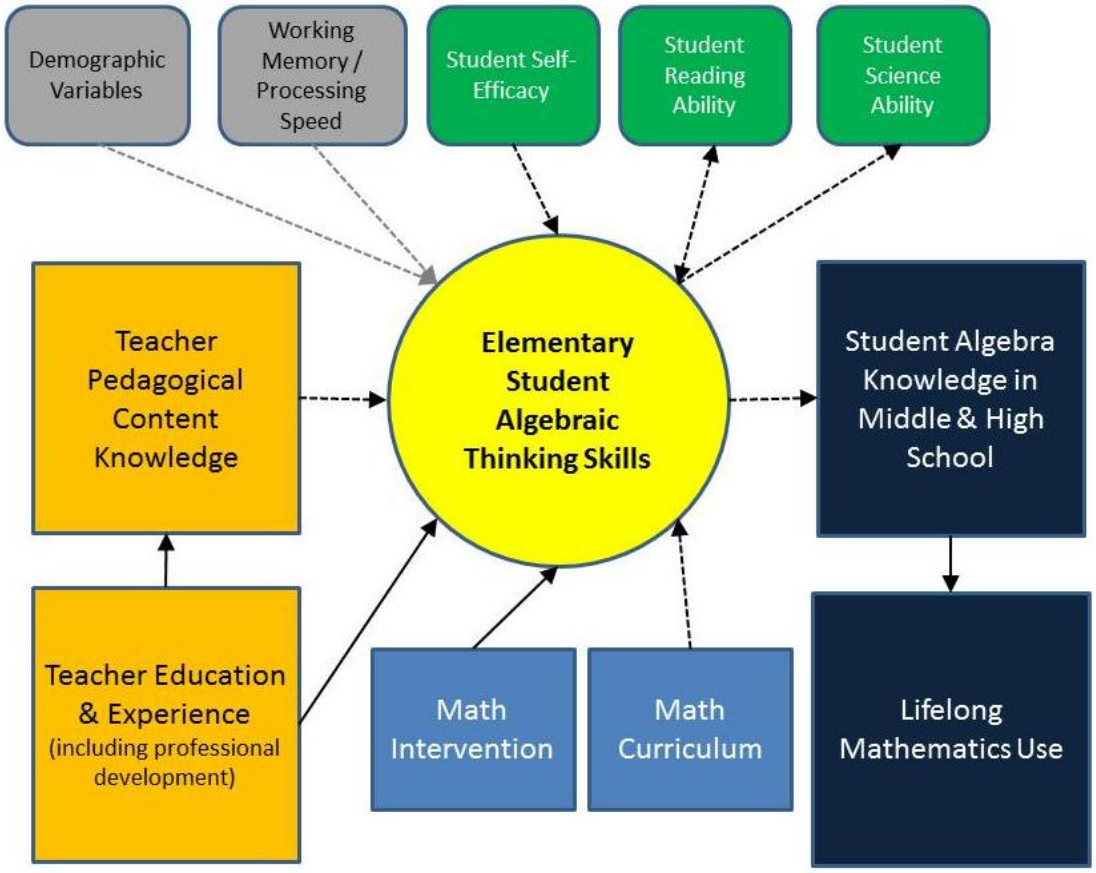
Lifelong Mathematics Use

In the United States algebra is often an important determinant in students' mathematical lives; in that often it is after their experience with 'official' algebra that students determine if they would like to continue to pursue math (Mason, 2008). As discussed previously, the National Mathematics Advisory Panel (2008) demonstrated that students who successfully complete Algebra II coursework in high school may be more likely to graduate from college and

earn higher incomes. Because of this research this variable is shown to be caused indirectly by algebraic thinking in the nomological net in Figure 11 below.

Figure 11.

Nomological Net



Appendix B: Test Specification for the Student Algebraic Thinking Skills Assessment
Grades 1 – 5
2012

I. TEST PURPOSE

The purpose of this test is to measure the level of algebraic thinking skills that elementary school students (i.e., grades 1, 2, 3, 4, and 5) in Washington State have reached. Teachers should use the results of this test diagnostically. The test should be given at the beginning of each school year so the new teacher can plan effective opportunities for students to experience algebraic thinking on an as needed basis.

Assessment questions will be broken into three strands of algebraic thinking: 1) Modeling, which includes solving open number sentences, understanding equivalence, and work with variables; 2) Generalized Arithmetic, which includes efficient numerical calculations (i.e., simplify calculations using number relations and compensation strategies) and generalizing (i.e., utilizing mathematical properties like the commutative property, property of zero, etc.); and 3) Functions, which includes possessing the ability to recognize, describe, extend, and create patterns. The Washington State Learning Standards and the National Common Core Standards document that these algebraic thinking standards are required at a variety of grade levels (i.e., grades 1 to 5). These are detailed below. Despite these general grade level standards, all three topics will be assessed all five grade levels in varying degrees of sophistication and complexity.

Algebraic Thinking Dimension	Dimension Tasks	Washington State Grade Level & Standard	Common Core Grade Level & Standard
Modeling (M)	Solving open number sentences	2nd: "Solve equations in which the unknown number appears in a variety of positions" (2.2.G)	1st: "Determine the unknown whole number in an addition or subtraction equation relating three whole numbers" (1.OA) 3rd: "Determine the unknown whole number in a multiplication or division equation relating three whole numbers" (3.OA)
	Understanding equivalence	2nd: "Solve equations in which the unknown number appears in a variety of positions" (2.2.G) 3rd: "Determine whether two expressions are equal and use '=' to denote equality" (3.5.A)	1st: "Understand the meaning of the equal sign, and determine if equations involving addition and subtraction are true or false" (1.OA)
	Work with variables	4th: "Represent an unknown quantity in simple expressions, equations, and inequalities using letters, boxes, and other symbols" (4.4.A)	3rd: "Solve two-step word problems using the four operations. Represent these problems using equations with a letter standing for the unknown quantity" (3.OA)
Generalized Arithmetic (GA)	Simplifying	2nd: "Add and subtract two-digit numbers efficiently and accurately" (2.2.C)	2nd: "Fluently add and subtract within 20 using mental strategies" (2.OA)
	Generalization	3rd: "Determine whether two expressions are equal and use '='	1st : "Apply properties of operations as strategies to add and subtract"

		to denote equality" (3.5.A)[a]	(1.OA) 3rd: "Apply properties of operations as strategies to multiply and divide" (3.OA)
Functions (F)	Repeating Patterns	K: "Copy, extend, describe, and create simple repetitive patterns" (K.2.A).	4th: "Generate a number or shape pattern that follows a given rule. Identify apparent features of the pattern that were not explicit in the rule itself" (4.OA)
	Linear patterns	1st: "recognize, create, and extend number patterns" (1.2.I) 2nd: "Create and state rule for patterns that can be generated by addition and extend the pattern" (2.2.F)	3rd: "Identify arithmetic patterns and explain them using properties of operations" (3.OA) 4th: "Generate a number or shape pattern that follows a given rule. Identify apparent features of the pattern that were not explicit in the rule itself" (4.OA)
	Nonlinear patterns	5th: "describe and create a rule for numerical and geometric patterns and extend the patterns" (5.4.A)	4th: "Generate a number or shape pattern that follows a given rule. Identify apparent features of the pattern that were not explicit in the rule itself" (4.OA)

II. ALGEBRAIC THINKING DIMENSIONS AND LEARNING TARGETS

Dimension 1: Modeling

Solving Open Number Sentences: According to the Washington State Standards, 2nd grade students should be able to "solve equations in which the unknown number appears in a variety of positions" (2.2.G) (Washington State K-12 Mathematics Standards, 2008, p. 4). According to the National Common Core Standards, 1st grade students should be able to "determine the unknown whole number in an addition or subtraction equation relating three whole numbers" (Common Core Standards for Mathematics, 2010, p. 15). Further, 3rd grade students should be able to "determine the unknown whole number in a multiplication or division equation relating three whole numbers" (Common Core Standards for mathematics, 2010, p. 23).

Understanding Equivalence: According to the Washington State Standards, 2nd grade students should be able to "solve equations in which the unknown number appears in a variety of positions" (2.2.G) (Washington State K-12 Mathematics Standards, 2008, p. 4). According to the National Common Core Standards, 1st grade students should be able to "understand the meaning of the equal sign, and determine if equations involving addition and subtraction are true or false" (Common Core Standards for Mathematics, 2010, p. 15).

Work with Variables: According to the Washington State Standards, 4th grade students should be able to "represent an unknown quantity in simple expressions, equations, and inequalities using letters, boxes, and other symbols" (4.4.A) (Washington State K-12 Mathematics Standards, 2008, p. 8). According to the National Common Core Standards, 1st grade students should be able to solve word problems that call for addition of three whole numbers whose sum is less than or equal to 20, e.g., by using objects, drawings, and equations with a symbol for the unknown number to represent the problem" (Common Core Standards for Mathematics, 2010, p. 15).

Specifically work with variables is not mentioned until the 6th grade standards, however, when students should be able to "write, read, and evaluate expressions in which letters stand for numbers" (Common Core Standards for Mathematics, 2010, p. 43).

Dimension 2: Generalized Arithmetic

Efficient Numerical Calculation: According to the Washington State Standards, 2nd grade students should be able to "add and subtract two-digit numbers efficiently and accurately" (2.2.C) and "add and subtract two-digit numbers mentally" (2.2.D) (Washington State K-12 Mathematics Standards, 2008, p. 6). According to the National Common Core Standards, 2nd grade students should be able to "fluently add and subtract within 100 using strategies based on place value, properties of operations, and/or the relationship between addition and subtraction" (Common Core Standards for Mathematics, 2010, p. 19).

Generalizing: According to the Washington State Standards, 3rd grade students should be able to "determine whether two expressions are equal and use '=' to denote equality" (3.5.A) (Washington State K-12 Mathematics Standards, 2008, p. 6). According to the National Common Core Standards, 1st grade students should be able to "apply properties of operations as strategies to add and subtract (Commutative property of addition and Associative property of addition)" (Common Core Standards for Mathematics, 2010, p. 15). Further, 3rd grade students should be able to "apply properties of operations as strategies to add and divide (Commutative property of multiplication, Associative property of multiplication, and Distributive property)"(Common Core Standards for Mathematics, 2010, p. 23).

Dimension 3: Functions

Repeating Patterns: According to the Washington State Standards, kindergarten students should be able to "copy, extend, describe, and create simple repetitive patterns" (K.2.A). The National Common Core Standards does not appear to incorporate repeating patterns.

Linear Patterns: According to the Washington State Standards, 1st grade students should be able to "recognize, create, and extend number patterns" (1.2.I) and 2nd grade students be able to "create and state rule for patterns that can be generated by addition and extend the pattern" (2.2.F) (Washington State K-12 Mathematics Standards, 2008, p. 2-3). According to the National Common Core Standards, 3rd grade students should be able to "identify arithmetic patterns" (Common Core Standards for Mathematics, 2010, p. 23).

Nonlinear Patterns: According to the Washington State Standards, 5th grade students should be able to "describe and create a rule for numerical and geometric patterns and extend the patterns (5.4.A) (Washington State K-12 Mathematics Standards, 2008, p. 10). According to the National Common Core Standards, 4th grade students should be able to "generate a number or shape pattern that follows a given rule" (Common Core Standards for Mathematics, 2010, p. 29).

III. CONTENT ORGANIZATION

The student algebraic thinking assessment will consist of approximately 25-28 items, resulting in a total of approximately 27-31 points. Items are written at a reading level appropriate for an elementary school student audience; therefore the items will aim to be readable at a 2nd grade or lower readability level. Teachers may read the items to students who cannot read at a 2nd grade level.

Tests include the following item types:

1) **Answer Only items:** The student solves the problem presented and provides a numerical answer. This numerical answer must match the numerical answer in the answer key to be awarded 1 point. If students are asked to construct a diagram or picture, the diagram or picture must match exactly the diagram or picture in the answer key to be awarded 1 points.

2) **Short-Answer items:** The student will construct a short response. The student may be asked to show all of their work, write a sentence, or explain the answer given.

The student algebraic thinking test is designed to be administered in one sitting in approximately 30 minutes. The test is not a timed test so the teacher may choose to allow more than 30 minutes of time or discontinue testing at the conclusion of 30 minutes.

Each test will contain a variety of items so that all three dimensions of algebra are addressed. Each grade level, grades 1-5, will have its own specific test with items of increasing difficulty. Further, the dimensions will be emphasized in different proportions for the different grades based on when that skill is expected to be mastered in the curriculum. Work with variables, for example, is a 4th grade standard, therefore although it will be included it will not be emphasized on the assessment until grade 3. Teachers will have access to all 5 versions of the assessment, so if needed they may give students easier or more difficult versions as needed.

IV. ITEM / TEST SCORING

An item-specific rubric will be provided for each Short-Answer item for grading purposes. Short-Answer items will be scored with a 3-level scoring rubric (0-2).

Each operational test will include approximately 23-25 Answer Only items worth one point each for a total of 23-25 points plus 2-3 Short-Answer items worth two points each for a total of 4-6 points. This leads to a grand total of 27-31 points. Thus Answer Only items account for approximately 80% of the score and Short-Answer items account for approximately 20% of the score.

V. REPORTING SCHEMA AND ITEM DISTRIBUTION

Each algebraic thinking dimension will be worth approximately 3-17 points out of the total points. Modeling will account for the majority of the points, with about 55% of the total score,

with the remaining points divided up about equally between Generalized Arithmetic (22.5%) and Functions (22.5%).

1st / 2nd GRADE					
Code	Dimension Details	Number of Questions	Number of Questions per Dimension	Number of Points	Number of Points per Dimension
M	Solving Open Number Sentences	6	16	6	16-17
	Understanding Equivalence	6-7		7-8	
	Work with Variables	3		3	
GA	Efficient Numerical Calculations	1-2	2	1-2	3-4
	Generalizing	1		2	
F	Repeating Patterns	2	6	2	5-7
	Linear Patterns	2-3		2-3	
	Nonlinear Patterns	1-2		1-2	

3rd / 4th / 5th GRADE					
Code	Dimension Details	Number of Questions	Number of Questions per Dimension	Number of Points	Number of Points per Dimension
M	Solving Open Number Sentences	5-6	14-15	5-6	15-16
	Understanding Equivalence	5		6	
	Work with Variables	4		4	
GA	Efficient Numerical Calculations	2-3	4-5	2-3	6-7
	Generalizing	2		4	
F	Repeating Patterns	-	6	-	6
	Linear Patterns	3		3	
	Nonlinear Patterns	3		3	

VI. GENERAL CONSIDERATIONS

1) The amount of reading required will be kept to a minimum and the aim for the overall readability will be approximately a 2nd grade level. Teachers may read items to students as needed at their discretion.

2) If used, character names will be representative of both genders. These names will be short and easy to read.

3) No stem or answer to one item will provide a clue or answer to a different item.

4) If desired, students may use classroom manipulatives or tools, including rulers, base ten blocks, unifix cubes, algebra tiles, etc. They may not use calculators.

5) If necessary, extra scratch paper may be used, although it should be unnecessary.

6) Specific rules for Answer-Only and Short-Answer items are given in the Item Specifications.

Appendix C: Item Specification for the Student Algebraic Thinking Skills Assessment
Grades 1 – 5
2012

The purpose of this test is to measure the level of algebraic thinking achievement that elementary school students (i.e., grades 1, 2, 3, 4, and 5) in Washington State have reached. Teachers should use the results of this test diagnostically. The test should be given at the beginning of each school year so the new teacher can plan effective opportunities for students to experience algebraic thinking on an as needed basis.

Assessment questions will be broken into three strands of algebraic thinking: 1) Modeling, which includes solving open number sentences, understanding equivalence, and work with variables; 2) Generalized Arithmetic, which includes efficient numerical calculations (i.e., simplify calculations using number relations and compensation strategies) and generalizing (i.e., utilizing mathematical properties like the commutative property, property of zero, etc.); and 3) Functions, which includes possessing the ability to recognize, describe, extend, and create patterns. The Washington State Learning Standards and National Common Core Standards document that these algebraic thinking standards are required at a variety of grade levels (i.e., grades 1 to 5). These are detailed below.

Code	Algebraic Thinking Dimension	Dimension Details	Washington State Standard Grade Level	Common Core Standard Grade Level
M	Modeling	Solving open number sentences	2nd	1st
		Understanding equivalence	2nd	1st
		Work with variables	4th	6th
GA	Generalized Arithmetic	Simplifying	2nd	2nd
		Generalizing	3rd	1st
F	Functions	Repeating Patterns	K	-
		Linear patterns	1st, 2nd	3rd
		Nonlinear patterns	5th	4th

The following generic scoring rules will be applied to all Short-Answer items. The items will be scored using a generic rubric for that specific dimension of algebraic thinking, with 2 points possible and a range of scores including 0, 1, and 2.

Scoring rules for items that assess the <u>Meaning of the Equal Sign:</u>
Student is asked to identify the equal sign symbol and then write 1-2 sentences about the meaning of the equal sign.

Score Awarded	Scoring Rules for the Meaning of the Equal Sign	Examples
2 points	Student's response indicates understanding of the equal sign symbol relationally (i.e., to mean the same as). Student likely uses the word "same".	• This symbol means the two numbers or things are the same • This symbol means is the same as • The same • The same as • That it is the same on both sides • It shows that something is even or balanced on both sides
1 point	Student's response indicates understanding of the equal sign operationally (i.e., to put the answer next). Student likely uses the word "answer" or "total".	• What the answer is • When you find the answer to a problem • When you're done adding it equals the total • When you add or subtract that is your answer • The total sum • What it all adds up to • When it is all done • What the total number is
0 points	Student's response is incoherent, incorrect, or unanswered. Student may simply repeat the word "equals".	• Blank • What will it equal • Equal • What the number is • To add • The symbol means equals • I don't know • Like 9+1=10

Scoring rules for items that assess <u>Generalizing a-a=0:</u>
Student is asked to write 1-2 sentences explaining why a-a=0 is always true or not always true.

Score Awarded	Scoring Rules for Generalizing a-a=0	Examples
2 points	Student's response is correct and shows effective reasoning. Student understands that subtracting the same number from itself results in zero and that this **<u>always</u>** works. Student likely uses the words "same number", "always", "any", "anything", "zero", etc. Student may provide an example but they provide additional reasoning.	• Well it's just like 7-7=0 and anything take away the same number is 0 • Because something – the same number is always 0 • Because if you take a number away from itself it will be 0 • Because any number minus itself is zero • A number take away the same number is always zero
1 point	Student's response is correct but the student either does not supply reasoning or the reasoning is undeveloped or incorrect. Student may provide only an example or not understand that this concept always works. Students may have partial understanding or supply part but not all of the answer.	• You're minusing the same number • The a's usually have to be the same number • Pretend a is 1, 1-1=0 • Because 10-10 does = 0, 0-0=0, 8-8=0 and even 1,000,000-1,000,000=0 • Because you subtract everything so there is nothing left • Because they are both the same number • Because it's taking away itself
0 points	Student's response is incoherent, incorrect, or unanswered. Student may simply repeat the problem, or not understand that the variables stand for numbers.	• Blank • You don't know what a is • How can a be a number? • You can count back from a number and end up with zero • Well I really can't explain • Not always true • I figured it out by counting • I don't know • I am not in algebra • Because I think a is 3

Scoring rules for items that assess <u>Generalizing a+0=a:</u>
Student is asked to write 1-2 sentences explaining why a+0=a is always true or not always true.

Score Awarded	Scoring Rules for Generalizing a+0=a	Examples
2 points	Student's response is correct and shows effective reasoning. Student understands that adding zero to a number results in that same number and that this **<u>always</u>** works. Student likely uses the words "same number", "always", "anything", "zero", etc. Student may provide an example but they provide additional reasoning.	• Because if you add anything to zero you will get the same number • Because anything plus zero = zero • Because if you add something with a zero you always get the number you add zero with • Because 0 is practically nothing so something plus 0 is that something • Anything + nothing always = that same figure that is in the first place • Because any number plus zero equals that number, for example 5+0=5
1 point	Student's response is correct but the student either does not supply reasoning or the reasoning is undeveloped or incorrect. Student may provide only an example or not understand that this concept always works. Students may have partial understanding or supply part but not all of the answer.	• Because zero plus zero equals zero • For example the a is 7, 7+0=7 so its always true • Because 1+0=1! • Because you don't add anything • So like if I got 6 dogs and I got no cats that = 6 dogs
0 points	Student's response is incoherent, incorrect, or unanswered. Student may simply repeat the problem, or not understand that the variables stand for numbers.	• Blank • It's hard! • I don't know • You have to have a secret number • Sometimes it could be true and sometimes it could be not always true • Because a+0=a, example: b+0=b • A is not a number • Because you don't know what a is

Scoring rules for items that assess <u>Generalizing a+b=b+a:</u>
Student is asked to write 1-2 sentences explaining why a+b=b+a is always true or not always true.

Score Awarded	Scoring Rules for Generalizing a+b=b+a	Examples
2 points	Student's response is correct and shows effective reasoning. Student understands that the order of numbers does not matter in addition sentences and that this **<u>always</u>** works. Student likely uses the words "same number", "always", "adding / addition", etc. Student may provide an example but they provide additional reasoning.	• Because a+b and b+a are just the same numbers in a different order • When the same numbers are added it doesn't matter what order they are in • In addition you can add anyway and get the same answer • Because in addition you can flip flop the numbers and they will always be the same • In multiplication and addition you will always get the same answer even if the numbers are turned around
1 point	Student's response is correct but the student either does not supply reasoning or the reasoning is undeveloped or incorrect. Student may provide only an example or not understand that this concept always works. Students may have partial understanding or supply part but not all of the answer.	• Because 7+3 is 10 and 3+7=10 • I added 9+10 and 10+9 and both equal 19 • Because they are the same numbers in different places • They are both the same just flipped around • Because a has to be the same number and so does b • Because they are the same but backwards
0 points	Student's response is incoherent, incorrect, or unanswered. Student may simply repeat the problem, or not understand that the variables stand for numbers.	• Blank • I don't know • All they did was switch them • Because on both sides there is an a and a b • Some are false and some are true • 1+2=2 it is 3 not 2 so it is false • Because a could be any number

Scoring rules for items that assess <u>Generalizing a-b=b-a:</u>
Student is asked to write 1-2 sentences explaining why a-b=b-a is always true or not always true.

Score Awarded	Scoring Rules for Generalizing a-b=b-a	Examples
2 points	Student's response is correct and shows effective reasoning. Student understands that the order of numbers does matter in subtraction sentences and that this is **<u>always</u>** the case. Student likely uses the words "same number", "always", "subtracting", "negative numbers" etc. Student may provide an example but they provide additional reasoning.	• When you subtract you can't do what you do in addition so 7-6=1 but you can't do 6-7 because it will be a negative answer • Because one is smaller than the other and you could get a negative number, a could = 5 and b could = 3, 5-3 is not equal to 3-5 • Since a is bigger than b, if b subtracted by a, the number will be negative but if and b are equal, the sum will always be 0
1 point	Student's response is correct but the student either does not supply reasoning or the reasoning is undeveloped or incorrect. Student may provide only an example or not understand that this concept always works. Students may have partial understanding or supply part but not all of the answer.	• One example is 9-8=1 and 8-9 you can't do because 8 is smaller than 9 • No because you would get a negative in one of them • If you subtract something it cannot go backward • It only works when a and b are the same number • If a=5 and b=3 5-3=2 but 3-5=-2 • Because you have to subtract the smaller number from the bigger • Because the small number can't take away the big number
0 points	Student's response is incoherent, incorrect, or unanswered. Student may simply repeat the problem, or not understand that the variables stand for numbers.	• Blank • I don't know • You never know what it will be plus I do not know what b and a is • You need to know the numbers first • What are the letters? • This is so hard • They are the same letters • They are not always equal • The letters can change

Scoring rules for items that assess <u>Generalizing ax0=a:</u>
Student is asked to write 1-2 sentences explaining why ax0=a is always true or not always true.

Score Awarded	Scoring Rules for Generalizing ax0=a	Examples
2 points	Student's response is correct and shows effective reasoning. Student understands that a number multiplied by zero equals zero and that this is **<u>always</u>** the case. Student likely uses the words "anything", "zero", "always", "multiplying / timesing", etc. Student may provide an example but they provide additional reasoning.	• Usually when you have a number times 0 the answer is 0 • Anything times 0 will equal 0 • Let's say a is 4. 4x0=4 is incorrect. When you multiply anything with zero it always ends up being zero as an answer • a+0 would equal a but ax0 would equal 0 because anything times zero equals zero • If you have something times 0 it equals 0 no matter what it equals 0
1 point	Student's response is correct but the student either does not supply reasoning or the reasoning is undeveloped or incorrect. Student may provide only an example or not understand that this concept always works. Students may have partial understanding or supply part but not all of the answer.	• If a was 2 then 2x0=0 it will not work • Because it is multiplying and it would be 0 • Because like 1x0=0 it is always 0 because it can't be 1 unless you do 1x1 then it would be true • 0x0=0 • Because if it was 5x0 would be 0 that's how it works 5x0=0 • Because ax0 doesn't equal a. It equals 0. So ax0=0 not a.
0 points	Student's response is incoherent, incorrect, or unanswered. Student may simply repeat the problem, or not understand that the variables stand for numbers.	• Blank • I don't know • Because if a is negative then that answers wrong • Because if you times by a zero it will still be the same number • Because it depends if you are trying to use the commutative property • a could equal 8 and with the zero, it could equal to 8 again

Scoring rules for items that assess <u>Generalizing ax1=a:</u>
Student is asked to write 1-2 sentences explaining why ax1=a is always true or not always true.

Score Awarded	Scoring Rules for Generalizing ax1=a	Examples
2 points	Student's response is correct and shows effective reasoning. Student understands that a number multiplied by one equals the original number and that this is **<u>always</u>** the case. Student likely uses the words "same number", "always", "multiplying / timesing", etc. Student may provide an example but they provide additional reasoning.	• If you multiply a number by 1 it will always be that number • Because something times 1 is always the same number • Everything you times by 1 is what you timesed the 1 by • Anything x 1 is always the first number • Because anything multiplied by one does not ever change • Anything times one is always one
1 point	Student's response is correct but the student either does not supply reasoning or the reasoning is undeveloped or incorrect. Student may provide only an example or not understand that this concept always works. Students may have partial understanding or supply part but not all of the answer.	• If you do 40 times 1 it's still the same • Yes because 5x1=5 you just multiply by one • Because 1x1=1 and 10x1=10 • Because it means a 1 time • Because you always have 1 group of a so it is a • Because 1 doesn't change the number
0 points	Student's response is incoherent, incorrect, or unanswered. Student may simply repeat the problem, or not understand that the variables stand for numbers.	• Blank • I don't know • Anything times 1 = 1 • How they find the answer rany way • Because ax1=a • Because you never change the number

ALGEBRAIC THINKING DIMENSIONS AND LEARNING TARGETS

Dimension 1: Modeling

Solving Open Number Sentences: According to the Washington State Standards, 2nd grade students should be able to "solve equations in which the unknown number appears in a variety of positions" (2.2.G) (Washington State K-12 Mathematics Standards, 2008, p. 4). According to the National Common Core Standards, 1st grade students should be able to "determine the unknown whole number in an addition or subtraction equation relating three whole numbers" (Common Core Standards for Mathematics, 2010, p. 15). Further, 3rd grade students should be able to "determine the unknown whole number in a multiplication or division equation relating three whole numbers" (Common Core Standards for mathematics, 2010, p. 23).

Item Format:

- Open-Answer items will be used to test this learning target

Stimulus Attributes:

- Items will include directions to complete the answer, such as "Solve the problem and write the answer in the box", "Solve the problem", or simply "Solve"
- Items may also include showing a number sentence and asking students to circle "true" or "false"
    - True or False statements might include 3+4=7; 8=5+3; 7=6+7; etc.
- Item stems / directions will be as short as possible (i.e., less than 10 words) and stated in command format
- Whole numbers will be presented in one of the various non-traditional formats of the equation a+b=c
    - Examples of these different formats include: 5+__=9, __-6=12, or 36=__x6
- Items for 1st & 2nd Grade Students
    - Items for 1st and 2nd grade students will include addition and subtraction operations only
    - One or two digit numbers are allowed for 1st and 2nd grade students
- Items for 3rd, 4th, & 5th Grade Students
    - One, two, or three digit numbers are allowed for 3rd, 4th, and 5th grade students
    - Items for 3rd grade students will include addition, subtraction, and multiplication operations only
    - Items for 4th and 5th grade students may include addition, subtraction, multiplication, and/or division operations
- A box will be used to denote where the student should write the answer

Mathematical Vocabulary and Terms:

- Terms that may be used: *number, solution, problem, solve*

- Terms that may not be used: *equivalence*

Item Characteristics:

- Items may ask students to solve problems in the format a+__=c, __-b=c, ax__=c, ___=a-b, and __/b=c,
- Items may ask students to solve problems with numbers in either single digit or double digit numbers
- Items may ask students to circle "true" or false"

Example Items:

1. Solve.

8 + __ = 15

2. Solve the problem and write the answer in the box.

__ - 3 = 12

3. Circle True or False.

8=5+13 True False

Understanding Equivalence: According to the Washington State Standards, 2nd grade students should be able to "solve equations in which the unknown number appears in a variety of positions" (2.2.G) (Washington State K-12 Mathematics Standards, 2008, p. 4). According to the National Common Core Standards, 1st grade students should be able to "understand the meaning of the equal sign, and determine if equations involving addition and subtraction are true or false" (Common Core Standards for Mathematics, 2010, p. 15).

Item Format:

- Open-Answer items will be used to test this learning target

Stimulus Attributes:

- Items will include directions to complete the answer, such as "Solve the problem and write the answer in the box", "Solve the problem", or simply "Solve"
- Items may also include showing a number sentence and asking students to circle "true" or "false"
  - True or False statements might include 5+3=5-3; etc.
- Item stems / directions will be as short as possible (i.e., less than 10 words) and stated in command format
- Whole numbers will be presented in one of the four versions of the equation a+b=c+d
- Items for 1st & 2nd Grade Students
  - Items for 1st and 2nd grade students will include addition and subtraction operations only
  - One or two digit numbers are allowed for 1st and 2nd grade students
- Items for 3rd, 4th, & 5th Grade Students
  - One, two, or three digit numbers are allowed for 3rd, 4th, and 5th grade students

- Items for 3rd grade students will include addition, subtraction, and multiplication operations only
- Items for 4th and 5th grade students may include addition, subtraction, multiplication, and/or division operations

- A box will be used to denote where the student should write the answer

Mathematical Vocabulary and Terms:

- Terms that may be used: *number, number sentence, solution, problem, solve, equation, variable*
- Terms that may not be used: *equivalence*

Item Characteristics:

- Items may ask students to solve problems in the following formats:
  a+__=c+d __+b=c+d a+b=__+d a+b=c+__

Example Items:

1. Solve the problem and write the answer in the box.

8 + __ = 3 + 9

2. Solve.

6 + 7 = __ + 4

3. Circle True or False.

5 + 3 = 5 – 3 True False

Work with Variables: According to the Washington State Standards, 4th grade students should be able to "represent an unknown quantity in simple expressions, equations, and inequalities using letters, boxes, and other symbols" (4.4.A) (Washington State K-12 Mathematics Standards, 2008, p. 8). According to the National Common Core Standards, 1st grade students should be able to solve word problems that call for addition of three whole numbers whose sum is less than or equal to 20, e.g., by using objects, drawings, and equations with a symbol for the unknown number to represent the problem" (Common Core Standards for Mathematics, 2010, p. 15). Specifically work with variables is not mentioned until the 6th grade standards, however, when students should be able to "write, read, and evaluate expressions in which letters stand for numbers" (Common Core Standards for Mathematics, 2010, p. 43).

Item Format:

- Open-Answer items will be used to test this learning target

Stimulus Attributes:

- Items will include directions to complete the answer, such as "What is c? Write the answer."
- Item stems / directions will be as short as possible short (i.e., less than 10 words) and stated in command format
- Whole numbers will be presented
- Items for 1st & 2nd Grade Students
  - Items for 1st and 2nd grade students will include addition and subtraction operations only
  - One or two digit numbers are allowed for 1st and 2nd grade students

- Items for 3rd, 4th, & 5th Grade Students
    - One, two, or three digit numbers are allowed for 3rd, 4th, and 5th grade students
    - Items for 3rd grade students will include addition, subtraction, and multiplication operations only
    - Items for 4th and 5th grade students may include addition, subtraction, multiplication, and/or division operations
- Variables will be noted using one of the following letters: a, b, c, e, g, x, or y
- A box will be used to denote where the student should write the answer

Mathematical Vocabulary and Terms:

- Terms that may be used: *number, solution, problem, variable, solve, value of, equation*
- Terms that may not be used: *equivalence*

Item Characteristics:

- Items may utilize the same variable once or twice only in an equation and ask the student to solve for that letter
    - Items may include up to three numbers and three unknown variables
        - This variable may be used once or repeated
        - Students must understand that if a variable is repeated it represents the same number each time
        - If two different variables are included, the number represented by one of the variables will be provided (i.e., If x=3, what is x+y=10?)
        - Items for 4th and 5th grade students may utilize two different variables and include two different equations so students may solve for the two different variables (i.e., if x+x+y=12 and y-x=3 what is x and y?)
    - Multiple work with variables items in the same assessment will not use the same letter to denote the variable (i.e., #4 will use 'c' and #9 will use 'x' to denote the variable)

Example Items:

1. 10+x=27, what is x?
2. If x=3, and x+y=12, what is y?
3. If x+x+y=12 and y-x=3 what is x and what is y?
4. What is c? Write the answer in the box.

c + c + 4 = 16

5. What is x? Write the answer in the box.

10 – x – x = 6

Dimension 2: Generalized Arithmetic

Efficient Numerical Calculation: According to the Washington State Standards, 2nd grade students should be able to "add and subtract two-digit numbers efficiently and accurately" (2.2.C) and "add and subtract two-digit numbers mentally" (2.2.D) (Washington State K-12 Mathematics Standards, 2008, p. 6). According to the National Common Core Standards, 2nd grade students should be able to "fluently add and subtract within 100 using strategies based on place value, properties of operations, and/or the relationship between addition and subtraction" (Common Core Standards for Mathematics, 2010, p. 19).

Item Format:

- Short-Answer items will be used to test this learning target

Stimulus Attributes:

- Items will include directions to complete the answer, such as "Circle True or False" or "Solve"
- Item stems / directions will be as short as possible (i.e., less than 10 words) and stated in command format
- Whole numbers will be presented
- A box will be used to denote where the student should write the answer

Mathematical Vocabulary and Terms:

- Terms that may be used: *number, solution, problem, simplify, solve, equation, variable*
- Terms that may not be used: *equivalence, function*

Item Characteristics:

- Items may utilize addition, subtraction, or multiplication strategies, and may include multiple operation strategies (i.e., both addition and subtraction) in the same problem
- Items may include 3 numbers on one side of the equal side and request the student to solve for the answer on the right side of the equal sign
- Items may be simple to solve mentally if simplification strategies are used
- Items may ask students to look at an equation and the method a fictional student used to solve the equation, and then ask students to explain or show why this strategy will work
- Items may ask students to solve problems with numbers ranging from 0 to 999
- Items for 1st & 2nd Grade Students
    - Items for 1st and 2nd grade students will include addition and subtraction operations only
    - One or two digit numbers are allowed for 1st and 2nd grade students
- Items for 3rd, 4th, & 5th Grade Students
    - One, two, or three digit numbers are allowed for 3rd, 4th, and 5th grade students
    - Items for 3rd, 4th, and 5th grade students may include addition, subtraction, and multiplication operations

- Numbers utilized should require calculation if efficient numerical calculation strategies are not used (i.e., all numbers used should not be single digits and the answer should not be solved in one's head without simplification strategies)

Example Items:

1. Solve the problem and write the answer in the blank. Please show all of your work.

35 + 57 – 56 = __

2. Look at the equation.

34+46+12=___

Tom solved the equation like this: 30+40+10+12

Explain or show why Tom's strategy will work

3. Circle True or False.

15+7-6=15+1 True False

4. Circle True or False.

(9x57)+57=10x57

5. Solve.

___ + 54 = 37 + 55

Generalizing: According to the Washington State Standards, 3rd grade students should be able to "determine whether two expressions are equal and use '=' to denote equality" (3.5.A) (Washington State K-12 Mathematics Standards, 2008, p. 6). According to the National Common Core Standards, 1st grade students should be able to "apply properties of operations as strategies to add and subtract (Commutative property of addition and Associative property of addition)" (Common Core Standards for Mathematics, 2010, p. 15). Further, 3rd grade students should be able to "apply properties of operations as strategies to add and divide (Commutative property of multiplication, Associative property of multiplication, and Distributive property)"(Common Core Standards for Mathematics, 2010, p. 23).

Item Format:

- Short-Answer items will be used to test this learning target

Stimulus Attributes:

- Item stems / directions will be as short as possible (i.e., less than 10 words) and stated in command format
- Items will ask students to circle 'Always True' or 'Not Always True'
- Items will ask students to explain why (i.e., 'How do you know?') and will include lines to explain their thinking
- A number sentence including variables and/or numbers will be presented at the top
- Three blank lines will be used to denote where the student should write the answer

Mathematical Vocabulary and Terms:

- Terms that may be used: *number, solution, problem, solve*

- Terms that may not be used: *equivalence*

Item Characteristics:

- Items may include variables denoted by letters
- Items may display a variety of important mathematical properties, including the commutative property, the associative property, the identity property, and the property of zero
- Items for 1st & 2nd Grade Students
  - Items for 1st and 2nd grade students will include addition and subtraction operations only
- Items for 3rd, 4th, & 5th Grade Students
  - Items for 3rd, 4th, and 5th grade students may include addition, subtraction, and multiplication operations
- Items may ask students to circle 'Always True' or 'Not Always True' and explain how they know their answer is correct

Example Items:

1. Look at the problem. Circle whether this is 'Always True' or 'Not Always True'. Please explain how you know.

   a + 0 = a

   Circle: Always True Not Always True

   Please explain how you know.

2. Circle whether this is 'Always True' or 'Not Always True'.

   a – b = b – a

   Circle: Always True Not Always True

   How do you know?

Dimension 3: Functions

Repeating Patterns: According to the Washington State Standards, kindergarten students should be able to "copy, extend, describe, and create simple repetitive patterns" (K.2.A). The National Common Core Standards does not appear to incorporate repeating patterns.

Item Format:

- Open-Answer items will be used to test this learning target

Stimulus Attributes:

- Items will include directions to complete the answer, such as "Draw the figure that would come next in the pattern" or "What shape would come 9th in the pattern?"
- Item stems / directions will be as short as possible and stated in command format
- Items will present either a figural repeating patterns
- A box will be used to denote where the student should draw the answer

- Students may be asked to "Show all of your work" and be provided with space to do so

Mathematical Vocabulary and Terms:

- Terms that may be used: *number, solution, problem, solve, rule, pattern, figure, draw, equation, variable, simplify*
- Terms that may not be used: *equivalence*

Item Characteristics:

- Items may ask students to write the next shape or draw the next figure in the pattern
- Figural repeating patterns may ask the student to write the name of the shape that will be in the 'next' figure (i.e., a near generalization)
- Figural repeating patterns may ask the student to write the name of the shape in the 20^{th} figure (i.e., a far generalization)
- Items may ask students to provide the rule of the repeating pattern

Example Items:

1. Look at the pattern made by shapes. Draw the figure that would come 9^{th} in the pattern.

Linear Patterns: According to the Washington State Standards, 1^{st} grade students should be able to "recognize, create, and extend number patterns" (1.2.I) and 2^{nd} grade students be able to "create and state rule for patterns that can be generated by addition and extend the pattern" (2.2.F) (Washington State K-12 Mathematics Standards, 2008, p. 2-3). According to the National Common Core Standards, 3^{rd} grade students should be able to "identify arithmetic patterns" (Common Core Standards for Mathematics, 2010, p. 23).

Item Format:

- Open-Answer items will be used to test this learning target

Stimulus Attributes:

- Items will include directions to complete the answer, such as "Draw the figure that would come next in the pattern", "Write the rule for this pattern", or "Write the missing number"
- Item stems / directions will be as short as possible (i.e., less than 10 words) and stated in command format
- Items will present either a numerical linear pattern or a figural linear pattern
- Whole numbers will be used if numbers are presented
- One or two digit numbers are allowed for all students
- A box will be used to denote where the student should write the answer
- Students may be asked to "Show all of your work" and be provided with space to do so

Mathematical Vocabulary and Terms:

- Terms that may be used: *number, solution, problem, solve, rule, pattern, figure, draw, equation, variable, simplify*
- Terms that may not be used: *equivalence*

Item Characteristics:

- Items may ask students to write the next number or draw the next figure in the pattern
- Items may ask students to write a specific number or draw a specific figure in the pattern
- Figural linear patterns may ask the student to write the number of blocks or dots that will be in the 'next' figure (i.e., a near generalization)
- Figural linear patterns may ask the student to write the number of blocks or dots in the 10^{th} or 20^{th} figure (i.e., a far generalization)
- Items may ask students to provide the rule of the numerical pattern

Example Items:

2. Write the rule for this pattern in the blank.

41, 45, 49, 53, 57, 61

3. Write the missing number.

72, 69, 66, 63, 60, ___

4. Draw the figure that would come next in the pattern. Write the number of dots it would have in the blank.

Nonlinear Patterns: According to the Washington State Standards, 5^{th} grade students should be able to "describe and create a rule for numerical and geometric patterns and extend the patterns (5.4.A) (Washington State K-12 Mathematics Standards, 2008, p. 10). According to the National Common Core Standards, 4^{th} grade students should be able to "generate a number or shape pattern that follows a given rule" (Common Core Standards for Mathematics, 2010, p. 29).

Item Format:

- Open-Answer items will be used to test this learning target

Stimulus Attributes:

- Items will include directions to complete the answer, such as "Draw the figure that would come next in the pattern", "Write the rule for this pattern", or "Write the missing number"

- Item stems / directions will be as short as possible (i.e., less than 10 words) and stated in command format
- Items will present either a numerical or figural nonlinear pattern
- Whole numbers will be used if numbers are presented
- One or two digit numbers are allowed for all students
- A box will be used to denote where the student should write the answer
- Students may be asked to "Show all of your work" and be provided with space to do so

Mathematical Vocabulary and Terms:

- Terms that may be used: *number, solution, problem, solve, rule, pattern, figure, equation, variable, simplify*
- Terms that may not be used: *equivalence, exponential*

Item Characteristics:

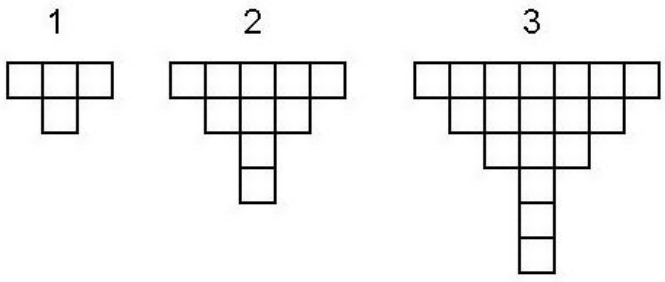
- Items may ask students to write the next number or draw the next figure in the pattern
- Items may ask students to write a specific number or draw a specific figure in the pattern
- Figural nonlinear patterns may ask the student to write the number of blocks or dots that will be in the 'next' figure (i.e., a near generalization)
- Figural nonlinear patterns may ask the student to write the number of blocks or dots in the 10^{th} or 20^{th} figure (i.e., a far generalization)
- Items may ask students to provide the rule of the numerical pattern

Example Items:

5. Write the missing number.

2, 3, 5, 8, 12, 17, ___

6. Draw the figure that would come next in the pattern. Write the number of dots it would have in the blank.



Appendix D: Algebraic Thinking Assessment Administration Directions

Timing: Time is not important in this assessment; it is not a timed assessment, and therefore it is recommended that you allow enough time for all students to complete the assessment. Given that this may not be possible, however, it is recommended you allow at least 30 minutes for each administration of the assessment tool.

Manipulatives: Students may use any manipulatives they desire. Please make basic manipulatives (i.e., cubes, tiles, etc.) available to students if they wish to use them.

Directions to read to students:

"This test is designed to find out what you know; it is not for a grade. So try your best, and please show all of your work. When you are done you may read a book from your desk. Look at the first page. Write your teacher's name in the blank at the top. Now, solve each of the problems on this test. Solve [write 'solve' on the board] *means to find the answer and write it. Explain* [write 'explain' on the board] *means to show or tell me about your answer and how you know what you know. Figure* [write 'figure' on the board] *means the shape made out of blocks or dots. Write your answers in the boxes. When you get to the bottom of a page, go on to the next page. Remember, try your best and show all of your work. If you're not sure, it's okay to go on to the next problem."*

When students are finished:

"Make sure you complete the bottom section on the last page. It says 'I am a: boy or girl'. Please mark boy if you are a boy and girl if you are a girl. Then write your age next to 'What is my age?' Do you speak English at home with your family? If yes, mark yes, if you speak a different language at home, mark no. Finally, it says 'I am good at math'. If you think you are good at math mark yes, if you think you are kind of good at math mark kind of and if you don't think you're good at math mark no. Now pass your papers to the front."

Appendix E: Math Survey for Elementary School Teachers

The purpose of the survey below is to better understand how Washington elementary school teachers teach math and algebraic thinking skills. **Algebra** has been defined by the National Council of Teachers of Mathematics (NCTM) as teaching students to understand patterns, relations, and functions; represent and analyze mathematical situations; use mathematical models to represent and quantify; and analyze change. Information you provide will be kept confidential and reported anonymously. Your help is greatly appreciated!

1. What is your name? _____
2. What school do you teach at? _____
3. What grade level do you currently teach?_____
4. At the end of this school year, how many years of K-5 teaching experience will you have? _____
5. How many students are in your current class?_____
6. What level of education have you obtained? Please check all that apply.

    ☐ Bachelors Degree

    ☐ Masters Degree

    ☐ Professional Certificate

    ☐ Administrative Certificate

    ☐ Ph.D.

    ☐ Other: _____

7. Which (if any) math curriculum do you use?

_____

_____

8. Think about the current school year when answering the question. Please check the appropriate box.

	Never	Rarely (e.g., twice a month)	Sometimes (e.g., once a week)	Often (e.g., 2-3 times a week)	Daily
How often do you teach math lessons that include algebraic thinking topics?	☐	☐	☐	☐	☐

9. Think about your current students when answering the question. Please check the appropriate box.

	Very Poor	Poor	Fair	Good	Very Good
Please rate the math skills of your current students.	☐	☐	☐	☐	☐
Please rate the algebraic thinking skills of your current students.	☐	☐	☐	☐	☐

10. Please rate the extent to which you agree with each of the following. Please check the appropriate box.

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
I feel well prepared to teach math.	☐	☐	☐	☐	☐
I feel well prepared to teach algebraic thinking skills.	☐	☐	☐	☐	☐
My students grasp algebraic thinking skills easily.	☐	☐	☐	☐	☐

11. What professional development activities have you participated in?

	Less than 3 months ago	3-12 months ago	1-5 years ago	More than 5 years ago
I have received professional development in teaching students algebraic thinking skills.	☐	☐	☐	☐
I have received professional development in teaching math (not algebraic thinking skills).	☐	☐	☐	☐

12. Is there anything else you would like me to know about teaching math?

_____

_____

_____

Thank you so much for your time!
The information you have provided is greatly appreciated.

Appendix F: Current Assessments of Algebraic Thinking

Authors	Ages	Skills Assessed	Purpose	Assessment Details	Psychometric Quality
Secondary					
Sylvan & Let's Go Learn Algebra Readiness Check-Up http://letsgolearn.com/fit4algebra/	'Pre-Algebra' students (i.e., likely Grades 6-8)	• Comparing and converting • Coordinate graphing • Decimal operations • Fraction operations • Geometry • Integer operations • Ratios and proportions	To determine if the student is "ready" for algebra	21 items Multiple-choice	• No information provided
Russell, O'Dwyer, & Miranda, 2009 Diagnostic Algebra Assessment System (DAAS)	Grades 6-12 (students in 'Algebra I')	• Concept of variable • Equality • Graphing	To diagnose algebraic misconceptions & inform teachers	10-12 items Multiple-choice	• Pilot studies • Validity studies • Misconception studies • Three-year trial investigating full and partial effects
McGraw-Hill Acuity Algebra	Grades 6-12 (students in 'Algebra I')	• Number sense, computation, and estimation • Variables and expressions • Inequalities and equations • Graphic inequalities and linear equations • Functions • Quadratic and exponential functions • Radical expressions and equations • Rational expressions and equations • Geometry • Statistics and probability	To assess algebra readiness, proficiency, and on-going progress	Multiple-choice & constructed response 1 readiness exam 13 formative assessments 1 proficiency exam	• No information provided • Developed using CTT & IRT
Kuchemann, 1981	Ages 13-15 (students in 'Algebra I')	• Variables	To investigate student understanding of variables	Approximately 26 items	• Convergent validity • Intervention study

				Open answer	
Foegen, Olson, & Impecoven-Lind, 2008 Project AAIMS (Algebra Assessment and Instruction: Meeting Standards)	Grades 6-12 (students in 'Algebra I')	• Solving simple equations • Applying the distributive property • Working with integers • Combining like terms • Working with variables • Manipulating expressions involving integers • Exponents • Order of operations • Basic graphing • Solving simple equations • Solving problems involving patterns and functions	To use as a progress-monitoring tools	50-60 items Open answer 5 minute time limit 2 versions: Basic Skills, Algebra Foundations	• Reliability analyses • Criterion validity correlations Parallel form reliability • Growth in construct over time • Teacher use studies
Asquith, Stephens, Knuth, & Alibali, 2007	Grades 6-8	• Work with variables • Equivalence	To use in comparison with teacher interview data	4 items Constructed response 3 versions	• Inter-rater reliability
Alibali, Knuth, Hattikudur, McNeil, & Stephens, 2007	Grades 6-8	• Equivalence	To measure student equivalence knowledge longitudinally	3 items Constructed response 3 versions	• Inter-rater reliability
Elementary					
Rittle-Johnson & Alibali, 1999	Grades 4-5	• Equivalence	To measure gains from an instructional intervention	Approximately 25 items Open answer Multiple-choice	• Inter-rater reliability
Hattikudur & Alibali, 2010	Grades 3-4	• Equivalence	To measure gains from an equality instructional intervention	6-8 items Open answer Constructed response	• Inter-rater reliability
McNeil & Alibali, 2005a	Ages 7-11	• Equivalence	To investigate the effects of arithmetic knowledge on	7 items Open answer	• No information provided

			equivalence knowledge	Constructed response Multiple-choice	
McNeil & Alibali, 2005b	Grades 3-5, 7	• Equivalence	To investigate the effects of the equal sign used in different contexts	Approximately 7 items Constructed response Multiple-choice 3 versions	• Inter-rater reliability
McNeil, 2007	Grades 1-4	• Equivalence	To investigate the association between age and equivalence knowledge	12 items Open answer	• No information provided
Fuchs, Compton, Fuchs, Hollenbeck, Craddock, & Hamlett, 2008 Dynamic Assessment (DA)	Grade 3	• Open number sentences with variables • Multiplication equations with variables • More than one missing variables	To measure gains from a mathematics problem solving instructional intervention	30-45 minute interview	• Convergent validity
Rittle-Johnson, Matthews, Taylor, & McEldoon, 2011	Grades 2-6	• Equivalence	To assess equivalence knowledge	37 items Open answer True / False Constructed response 2 versions	• Face validity • Factor analyses • Convergent validity • Internal reliability • Test-retest reliability • Inter-rater reliability • Alternate-forms reliability
Jacobs, Franke, Carpenter, Levi, & Battey, 2007	Grades 1-5	• Relational thinking • Simplifying • Work with variables • Generalization	To measure gains from a professional development program	8-25 items Open answer True / False 3 versions: 1st grade, 2nd/3rd grade, 4th/5th grade	• No information provided

References

Alibali, M. W., Knuth, E. J., Hattikudur, S., McNeil, N. M., & Stephens, A. C. (2007). A longitudinal examination of middle school students' understanding of the equal sign and equivalent equations. *Mathematical Thinking and Learning, 9*(3), 221-247.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, Calif.: Brooks/Cole Pub. Co.

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

Asquith, P., Stephens, A. C., Knuth, E. J., & Alibali, M. W. (2007). Middle school mathematics teachers' knowledge of students' understanding of core algebraic concepts: Equal sign and variable. *Mathematical Thinking and Learning, 9*(3), 249-272.

Baker, S., Gersten, R., & Lee, D. S. (2002). A synthesis of empirical research on teaching mathematics to low-achieving students. *Elementary School Journal, 10*, 51-73.

Ball, D. L. (1990). The mathematical understandings that prospective teachers bring to teacher education. *Elementary School Journal, 90*, 449-466.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389-407.

Baroody, A. J., & Gannon, K. (1984). The development of the commutativity principle and economical addition strategies. *Cognition and Instruction, 1*, 321-339.

Baroody, A. J., & Ginsburg, H. P. (1983). The effects of instruction on children's understanding of the equals sign. *The Elementary School Journal, 84*(2), 198-212.

Basterra, M. R., Trumbull, E., & Solano-Flores, G. (Eds.), (2011). *Cultural validity in assessment: A guide for educators*. New York: Routledge.

Becker, J. R., & Rivera, F. (2005). Generalization strategies of beginning high school algebra students. In H. L. Chick & J. L. Vincent (Eds.), *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education, Vol. 4*, (pp. 121-128). Melbourne: PME.

Behr, M., Erlwanger, S., & Nichols, E. (1976). How children view equality sentences. (PMDC Technical Report No. 3). Tallahassee: Florida State University. (ERIC Document Reporduction Service No. ED 144 802).

Bezuszka, S. J., & Kenney, M. J. (2008). The three R's: Recursive thinking, recursion, and recursive formulas. In C. E. Greenes & R. Rubenstein (Eds.), *Algebra and algebraic thinking in school mathematics (seventieth yearbook)* (pp. 81 - 97). Reston, VA: NCTM.

Billings, E.M., Tiedt, T., & Slater, L. H. (2008). Algebraic thinking and pictorial growth patterns. *Teaching Children Mathematics 14*, 302-308.

Birenbaum, M., Tatsuoka, K. K., & Gutvirtz, Y. (1992). Effects of response format on diagnostic assessment of scholastic achievement. *Applied Psychological Measurement, 16*(4), 353-363.

Birenbaum, M., Tatsuoka, C., & Xin, T. (2005). Large-scale diagnostic assessment: Comparison of eighth graders' mathematics performance in the United States, Singapore and Israel. *Assessment in Education, 12*, 67-81.

Black, P. (1993). Formative and summative assessment by teachers. *Studies in Science Education*, *21*, 49-97.

Black, P.. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-71.

Blanton, M. L., & Kaput, J. J. (2003). Developing elementary teachers' algebra eyes and ears. *Teaching Children Mathematics*, *10*(2), 70-77.

Blanton, M. L., & Kaput, J. J. (2004). Elementary grades students' capacity for functional thinking. *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education, 2*, 135-142.

Blanton, M. L., & Kaput, J. J. (2005). Characterizing a classroom practice that promotes algebraic reasoning. *Journal for Research in Mathematics Education, 36*(5), 412-446.

Booth, L. R. (1984). *Algebra: Children's strategies and errors. A report of the strategies and errors in secondary mathematics project*. Windsor, Berks: NFER-Nelson.

Brennan, R. L. (1992). *Elements of generalizabilty theory* (*rev. ed.).* Iowa City, IA: American College Testing.

Byrnes, J. P., & Wasik, B. A. (2009). Factors predictive of mathematics achievement in kindergarten, first and third grade: An opportunity-propensity analysis. *Contemporary Educational Psychology*, *34*(2), 167−183.

Cai, J., & Moyer, J. C. (2008). Developing algebraic thinking: Some insights from international comparative studies. In C. E. Greenes & R. Rubenstein (Eds.), *Algebra and algebraic thinking in school mathematics* (pp. 169-182). National Council of Teachers of Mathematics 2008 Yearbook. Reston, VA: NCTM.

Campbell, J. R., Hombo, C. M., & Mazzeo, J. (2000). *NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Capraro, M. M., Ding, M., Matteson, S., & Capraro, R. M. (2007). Representational implications for understanding equivalence. *Research in Brief, 107*(3) 86-90.

Carpenter, T.P., Franke, M.L., & Levi, L. (2003). *Thinking mathematically: Integrating arithmetic and algebra in the elementary school*. Portsmouth, NH: Heinemann.

Carpenter, T.P., & Levi, L. (2000). Developing conceptions of algebraic reasoning in the primary grades. *Research Report #002*. Madison, WI: National Center for Improving Student Learning and Achievement in Mathematics and Science.

Carpenter, T. P., Levi, L., Berman, P. W., & Pligge, M. (2005). Developing algebraic reasoning in the elementary school. In T. A. Romberg, T. P. Carpenter, & F. Dremock (Eds.), *Understanding mathematics and science matters* (pp. 81-98). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Carpenter, T.P., Levi, L., & Farnsworth, V. (2000). Building a foundation for learning algebra in the elementary grades. *In Brief: Vol. 1, No. 2.* Madison, WI: National Center for Improving Student Learning and Achievement in Mathematics and Science.

Carraher, D.W. & Schliemann, A. D. (2007). Early algebra and algebraic reasoning. In F. Lester (Ed.) *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 669-705). Charlotte, NC: Information Age Publishing.

Carraher, D. W., Schliemann, A. D., Brizuela, B. M., & Earnest, D. (2006). Arithmetic and algebra in early mathematics education. *Journal for Research in Mathematics Education, 37*(2) 87-115.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333-353.

Chazan, D. (1996). "Algebra for all students?" *Journal of Mathematical Behavior, 15*, 455-477.

Chazan, D. (2000). *Beyond formulas in mathematics and teaching: Dynamics of the high school algebra classroom*. New York: Teacher's College Press.

Clason, D. L., & Dormody, T. J. (1994) Analyzing data measured by individual Likert-type items. *Journal of Agricultural Education*, 35(4), 31- 35.

The Common Core Standards Writing Team. (2011). *Progressions for the Common Core State Standards in Mathematics*.

Clement, J. (1982). Algebra word problems solutions: Thought processes underlying a common misconception. *Journal for Research in Mathematics Education, 13*(1), 16-30.

Confrey, J. (1990). A review of the research on student conceptions in mathematics, science, and programming. *Review of Research in Education, 16*, 3-56.

Crocker, L. (2006). Introduction to measurement theory. In J.L. Green, G. Camilli, & P.B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 371-384). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measures*. New York: Wiley.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

DiLeo, J. (2007). *A study of a specific language arts and mathematics software program: Is there a correlation between usage levels and achievement?* Unpublished doctoral dissertation, Indiana University of Pennsylvania, Indiana, PA.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009*). Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken, NJ: Wiley.

Driscoll, M. (1999). *Fostering algebraic thinking: A guide for teachers, grades 6-10.* Educational Development Center, Inc.: Portsmouth, NH.

Ely, R. (2010). Nonstandard student conceptions about infinitesimals. *Journal for Research in Mathematics Education, 41*(2), 117-146.

Embretson, S., & Yang, X. (2006). Item Response Theory. In J.L. Green, G. Camilli, & P.B. Elmore (Eds.), *Handbook of Complementary Methods in Education Research* (pp. 385-409). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Englard, L. (2010). Raise the bar on problem solving. *Teaching Children Mathematics, 17*(3), 156-163.

Eric, C. C. M. (2009). Mathematical modelling as problem solving for children in the Singapore mathematics classrooms. *Journal of Science and Mathematics Education in Southeast Asia, 32*(1), 36-61.

Ericsson, K. A. & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Falkner, K. P., Levi, L., & Carpenter, T. P. (1999). Children's understanding of equality: A foundation for algebra. *Teaching Children Mathematics, 6*(4), 232-236.

Fast, L., Lewis, J., Bryant, M., Bocian, K., Cardullo, R., Rettig, M., & Hammond, K. (2010). Does math self-efficacy mediate the effect of the perceived classroom environment on standardized test math performance? *Journal of Educational Psychology*, 102, 729-740.

Ferrucci, B. J., Kaur, B., Carter, J. A., & Yeap, B. H. (2008). Using a model approach to enhance algebraic thinking in the elementary school mathematics classroom. In C. E. Greenes & R. Rubenstein. (Eds.), *Algebra and algebraic thinking in school mathematics: Seventieth yearbook* (pp. 195-210). Reston, VA: National Council of Teachers of Mathematics.

Filloy, E. & Rojano, T. (1984). From an arithmetical thought to an algebraic thought. In J. Moser (ed.), *Proceedings of PME-NA VI,* (pp. 51-56). Madison, Wisconsin.

Fisher, K. J., Borchert, K., & Bassok, M. (2011). Following the standard form: Effects of equation format on algebraic modeling. *Memory & Cognition, 39*(3), 502-515.

Foegen, A., & Olson, J. R. (2007). Effects of teachers' access to student data on algebra progress. *Project AAIMS: Technical Report #15.*

Foegen, A., Olson, J. R., & Impecoven-Lind, L. (2008). Developing progress monitoring measures for secondary mathematics: An illustration in algebra. *Assessment for Effective Intervention, 33*, 240-249.

Fuchs, L. S., Compton, D. L., Fuchs, D., Hollenbeck, K. N., Craddock, C. F., & Hamlett, C. L. (2008). Dynamic assessment of algebraic learning in predicting third graders' development of mathematical problem solving. *Journal of Educational Psychology*, *100*(4), 829-850.

Fujii, T., & Stephens, M. (2008). Using number sentences to introduce the idea of variable. In C. Greenes & R. Rubenstein (Eds.) *Algebra and algebraic thinking in school mathematics: Seventieth Yearbook*, pp. 127-140. Reston, VA: National Council of Teachers of Mathematics.

Garcy A. (2009). The longitudinal link between student health and math achievement scores. *Journal of Education for Students Placed at Risk, 14*(4), 283-310.

Geary, D. C. (2006). Development of mathematical understanding. In D. Kuhl & R. S. Siegler (Vol. Eds.), Cognition, perception, and language, Vol 2 (pp. 777-810) in W. Damon (Gen. Ed.), *Handbook of Child Psychology* (6th Ed.). New York: John Wiley & Sons.

Geary, D. C. (in press). Cognitive predictors of achievement growth in mathematics: A 5-year longitudinal study. *Developmental Psychology.*

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement, 6,* 263-275.

Ginsburg, H. P. (1997). *Entering the child's mind: The clinical interview in psychological research and practice*. Cambridge University Press: Cambridge, UK.

Goddard, R. D., Hoy, W. K., & Woolfolk, A. (2000). Collective teacher efficacy: Its meaning, measure, and effect on student achievement. *American Education Research Journal, 37*(2), 479-507.

Gottfried, M. A. (2009). Excused versus unexcused: How student absences in elementary school affect academic achievement. *Educational Evaluation and Policy Analysis, 31,* 392-419.

Groen, G. J., & Poll, M. (1973). Subtraction and the solution of open sentence problems. *Journal of Experimental Psychology, 16,* 292-302.

Grouws, D. A., & Good, T. L. (1976). Factors associated with third- and fourth-grade children's performance in solving multiplication and division sentences. *Journal for Research in Mathematics Education, 7*(3), 155-171.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11,* 385-398.

Gustin, W. C., & Corazza, L. (1994). Mathematical and verbal reasoning as predictors of science achievement. *Roeper Review, 16*(3), 160–163.

Hambleton, R. K., Swaminathan, H., Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Han, W-J., & Bridglall, B.L. (2009). Assessing school supports for ELL students using ECLS-K. *Early Childhood Research Quarterly, 24*, 445-462.

Hart, K. M. (1980). *Secondary school children's understanding of mathematics: a report of the mathematics component of the concepts in secondary Mathematics and Science Programme*. London : Chelsea College of Science and Technology

Hattikudur, S., & Alibali, M. W. (2010). Learning about the equal sign: Does comparing with inequality symbols help? *Journal of Experimental Child Psychology, 107*, 15-30.

Herscovics, N. (1989). Cognitive obstacles encountered in the learning of algebra. In S. Wagner, & C. Kieran (Eds.), *Research issues in the learning and teaching of algebra* (pp. 60-92). Reston, VA: National Council of Teachers of Mathematics.

Herscovics, N., & Linchevski, L. (1994). A cognitive gap between arithmetic and algebra. *Educational Studies in Mathematics, 27*(1), 59-78.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*, 371-406.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2010). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal, 105*(1), 11-30.

Hill, H. C., & Shih, J. C. (2009). Examining the quality of statistical mathematics education research. *Journal for Research in Mathematics Education, 40*(3), 241-250.

Huff, K. & Goodman, D. P., (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). New York, NY: Cambridge.

Jacobs, V. R., Franke, M. L., Carpenter, T. P., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. *Journal for Research in Mathematics Education, 38*(3), 258-288.

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

Kajander, A., & Lovric, M. (2009). Mathematics textbooks and their potential role in supporting misconceptions. *International Journal of Mathematical Education in Science and Technology, 40*(2), 173-181.

Kalton, G. (1983). *Introduction to Survey Sampling.* Beverly Hills: Sage Publications.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. T. (2006). Validation. In R. L. Brennan, (Ed.), *Educational measurement* (pp. 17-64). Westport, CT: American Council on Education / Praeger.

Kaput, J. J. (2000). Teaching and learning a new algebra with understanding. National Center for Improving Student Learning and Achievement in Mathematics and Science, Dartmouth, MA.

Kaput, J. J. (2000). Transforming algebra from an engine of inequity to an engine of mathematical power by "algebrafying" the K-12 curriculum. National Center for Improving Student Learning and Achievement in Mathematics and Science, Dartmouth, MA:

Kaput, J. J. (2008). What is algebra? What is algebraic reasoning? In J. J. Kaput, D. W. Carraher, & M. L. Blanton (Eds.), *Algebra in the early grades* (pp. 5-17). New York, NY: Taylor & Francis Group.

Kaput, J. J., & Blanton, M. L. (2000). *Algebraic reasoning in the context of elementary mathematics: Making it implementable on a massive scale*. National Center for Improving Student Learning and Achievement in Mathematics and Science, Dartmouth, MA.

Kaput, J. J., & Blanton, M. L. (2001). Student achievement in algebraic thinking: A comparison of 3rd graders' performance on a state 4th grade assessment. In *Proceedings of the Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, (pp. 99-107).

Kaput, J. J., & Blanton, M. L. (2005). A teacher-centered approach to algebrafying elementary mathematics. In T. A. Romberg, T. P. Carpenter, & F. Dremock (Eds.), *Understanding mathematics and science matters* (pp. 99-125). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kaput, J. J., Blanton, M. L., & Moreno, L. (2008). Algebra from a symbol-ization point of view. In J. J. Kaput, D. W. Carraher, & M. L. Blanton (Eds.), *Algebra in the early grades* (pp. 19-55). New York, NY: Taylor & Francis Group.

Kaput, J., & Clement, J. (1979). Letter to the editor. *Journal of Mathematical Behavior, 2*, 208.

Kazemi, E. (2002). Exploring test performance in mathematics: The question s children's answers raise. *Journal of Mathematical Behavior, 21*, 203-224.

Kieran, C. (1981). Concepts associated with the equality symbol. *Educational Studies in Mathematics, 12*(3), 317-326.

Kieran, C. (1989). The early learning of algebra: A structural perspective. In S. Wagner, & C. Kieran (Eds.), *Research issues in the learning and teaching of algebra* (pp. 33-56). Reston, VA: National Council of Teachers of Mathematics.

Kieran C. (1992). The learning and teaching of school algebra. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning*. New York: Macmillan Publishing Company.

Kieran, C. (2004). The equation/inequality connection in constructing meaning for inequality situations. In M. J. Hoines & A. B. Fuglestad (Eds.), *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 1, pp. 143-147). Bergen, Norway.

Kieran, C. (2007). Learning and teaching of algebra at the middle school through college levels: Building meaning for symbols and their manipulation. In F. K. Lester, (Ed.), *Second*

handbook of research on mathematics teaching and learning (pp. 707–762). Reston, VA: National Council of Teachers of Mathematics.

Kilpatrick, J., & Izsák, A. (2008). A history of algebra in the school curriculum. In C. E. Greenes & R. Rubenstein (Eds.), *Algebra and algebraic thinking in school mathematics: Seventieth yearbook* (pp. 3–18). Reston, VA: The National Council of Teachers of Mathematics.

Kline, R. B. (2011). *Principles and practice of structural equation modeling (3rd ed.).* NY: The Guilford Press.

Knuth, E. J., Alibali, M. W., McNeil, N. M., Weinberg, A., & Stephens, A. S. (2005). Middle school students' understanding of core algebraic concepts: Equality & variable. *International Reviews on Mathematical Education, 37*, 68-76.

Knuth, E., Stephens, A. C., McNeil, N. M., & Alibali, M. W. (2006). Does understanding the equal sign matter? Evidence from solving equations. *Journal for Research in Mathematics Education, 36*(4), 397-312.

Kuchemann, D. (1978). Children's understanding of numerical variables. *Mathematics in School, 7*(4), 23-26.

Kuchemann, D. (1981). Algebra, in Hart, K. (Ed.) *Children's understanding of mathematics 11-16*. London: John Murray.

Kukla-Acevedo, S. (2009). Do teacher characteristics matter? New results on the effects of teacher preparation on student achievement. *Economics of Education Review, 28*(1), 49-57.

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.

Lannin, J. (2003). Developing algebraic reasoning through generalization. *Mathematics Teaching in the Middle School, 8*(7), 342-348.

Lannin, J., Barker, D., & Townsend, B. (2006). Algebraic generalization strategies: Factors influencing student strategy selection. *Mathematics Education Research Journal, 18*(3), 3-28.

Leighton, J. P., & Gierl, M. J. (Eds.) (2007). *Cognitive diagnostic assessment for education: Theory and practices*. Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*(3), 205-237.

Li, X., Ding, M., Capraro, M. M., & Capraro, R. M. (2008). Sources of differences in children's understandings of mathematical equality: Comparative analysis of teacher guides and student texts in China and the United States. *Cognition and Instruction, 26*(2), 195-217.

Li, X., & Li, Y. (2008). Research on students' misconceptions to improve teaching and learning in school mathematics and science. *School Science and Mathematics, 108*(1), 4-7.

Linacre, J. M. (2005). WINSTEPS: Multiple-choice, rating scale, and partial credit Rasch analysis [computer software]. Chicago: MESA.

Lindvall, C. M., & Ibarra, C. G. (1980). Incorrect procedures used by primary grade pupils in solving open addition and subtraction sentences. *Journal for Research in Mathematics Education, 11*(1), 50-62.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-448.

Lochhead, J. (1980). Faculty interpretations of simple algebraic statements: The professor's side of the equation. *Journal of Mathematical Behavior, 3*(1), 28-37.

Ma, L. (1999). *Knowing and learning elementary mathematics*. Mahway, NJ: Lawrence Erlbaum.

MacGregor, M. & Stacey, K. (1993). Cognitive models underlying students' formulation of simple linear equations. *Journal for research in mathematics education, 24*(3), 217-232.

Macgregor, M. & Stacey, K. (1997). Students' understanding of algebraic notation: 11-15. *Educational Studies in Mathematics, 33*, 1-19.

Maerten-Rivera, J., Myers, N., Lee, O. and Penfield, R. (2010), Student and school predictors of high-stakes assessment in science. *Science Education, 94*, 937–962.

Mann, R. L. (2004). Balancing act: The truth behind the equals sign. *Teaching Children Mathematics*. 65-70.

Martinez, M.V. & Brizuela, B.M. (2006). A third grader's way of thinking about linear function tables. *Journal of Mathematical Behavior, 25*(4), 285-298.

Mason, J. (2008). Making use of children's powers to produce algebraic thinking. In J. J. Kaput, D. W. Carraher & M. L. Blanton (Eds.), *Algebra in the early grades* (pp. 57-94). New York: Erlbaum.

Matthews, P., & Rittle-Johnson, B. (2008). In pursuit of knowledge: Comparing self-explanations, concepts, and procedures as pedagogical tools. *Journal of Experimental Child Psychology, 104*, 1-21.

McNeil, N. M. (2007). U-Shaped development in math: 7-year-olds outperform 9-year-olds on equivalence problems. *Developmental Psychology, 43*(3), 687-695.

McNeil, N. M. (2008). Limitations to teaching children 2+2=4: Typical arithmetic problems can hinder learning of mathematical equivalence. *Child Development, 79*(5), 1524-1537.

McNeil, N. M., & Alibali, M. W. (2000). Learning mathematics from procedural instruction: Externally imposed goals influence what is learned. *Journal of Educational Psychology, 92*(4), 734-744.

McNeil, N. M., & Alibali, M. W. (2004). You'll see what you mean: Students encode equations based on their knowledge of arithmetic. *Cognitive Science, 28*, 451-466.

McNeil, N. M., & Alibali, M. W. (2005a). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development, 76*(4), 883-899.

McNeil, N. M., & Alibali, M. W. (2005b). Knowledge change as a function of mathematics experience: All contexts are not created equal. *Journal of Cognition and Development, 6*(2), 285-306.

McNeil, N. M., Fyfe, E. R., Petersen, L. A., Dunwiddie, A. E., & Brletic-Shipley, H. (2011). Benefits of practicing 4=2+2: Nontraditional problem formats facilitate children's understanding of mathematical equivalence. *Child Development, 82*(5), 1620-1633.

McNeil, N. M., Grandau, L., Knuth, E. J., Alibali, M. W., Stephens, A. C., Hattikudur, S., & Krill, D. E. (2006). Middle-school students' understanding of the equal sign: The books they can't read can't help. *Cognition and Instruction, 24*, 367-385.

McNeil, N. M., Weinberg, A., Hattikudur, S., Stephens, A. C., Asquith, P., Knuth, E. J., & Alibali, M. W. (2010). A is for apple: Mnemonic symbols hinder the interpretation of aglebraic expressions. *Journal of Educational Psychology, 102*(3), 625-634.

Messick, D. M. (1988). On the limitations of cross-cultural research in social psychology. In M. Bond (Ed.), *The cross-cultural challenge to social psychology* (pp. 41-47). Newbury Park, CA: Sage.

Messick, S. (1989a). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp13-103). New York: Macmillan.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Mix, K. S. (1999). Preschoolers' recognition of numerical equivalence: Sequential sets. *Journal of Experimental Child Psychology, 74,* 309-332.

Moses, R. and Cobb, C. (2001) *Radical equations: Civil rights from Mississippi to the Algebra Project*. Beacon Press, Boston, MA

Moss, J., Beatty, R., Shillolo, G., & Barkin, S. (2008). What is your theory? What is your rule? Fourth graders build their understanding of patterns and functions on a collaborative database. In C. Greenes (Ed.), *Algebra and algebraic thinking in school mathematics: The National Council of Teachers of Mathematics 70th Yearbook (2008)* (pp. 155–168). Reston, VA: NCTM.

Mullis, I.V.S., Martin, M.O., & Foy, P. (2008). *TIMSS 2007 international report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muthen, L.K. & Muthen, B.O. (1998-2007). *Mplus User's Guide. Fifth Edition*. Los Angeles, CA: Muthen & Muthen.

Nathan, M. J., & Koedinger, K. R. (2000). An investigation of teachers' beliefs of students' algebra development. *Cognition and Instruction, 18*(2), 209-237.

National Council of Teachers of Mathematics. (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (1997). A framework for constructing a vision of algebra. *Algebra working group document*. Reston, VA: Author.

National Council of Teachers of Mathematics. (1998). *The nature and role of algebra in the K-14 curriculum*. Washington: DC: National Academy Press.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Mathematics Advisory Panel. (2008). *Foundations for success: Final report of the National Mathematics Advisory Panel*. Washington, D. C.: U. S. Department of Education. Retrieved from http://www.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf.

Ng, S. F., & Lee, K. (2009). The model method: Singapore children's tool for representing and solving algebraic word problems. *Journal for Research in Mathematics Education, 40*(3), 282-313.

Ng, S. S. N., & Rao, N. (2010). Chinese number words, culture, and mathematics learning. *Review of Educational Research, 80*(2), 180-206.

Nichols, P. (1994). A framework for developing cognitively diagnostic assessments. *Educational Research, 64*(4), 575-603.

Nichols, P. D., Meyers, J. L., & Burling, K. S. (2009). A framework for evaluating and planning assessments intended to improve student achievement. *Educational Measurement: Issues and Practice, 28*(3), 14-23.

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice, 28*(1), 3-9

Nicolson, R. L. (1990). Design and evaluation of the SUMIT intelligent teaching assistant for arithmetic. *Interactive Learning Environments, 1(*4), 265-287.

Nunes, T., Schliemann, A.D., & Carraher, D.W. (1993). *Mathematics in the Streets and in Schools.* Cambridge, U.K: Cambridge University Press.

OECD. (2010). *PISA 2009 results: What students know and can do: Student performance in reading, mathematics and science (Volume I),* OECD Publishing.

Okpala, C. O., Smith, F., Jones, E., & Ellis, R. (2000). A clear link between school and teacher characteristics, student demographics, and student achievement. *Education, 120,* 487-495.

Pajares, F. & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology, 24*, 124–139.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Perry, M. (1991). Learning and transfer: Instructional conditions and conceptual change. *Cognitive Development, 6*, 449-468.

Perry, M., Church, R. B., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development, 3*, 359-400.

Piaget, J. (1952). *The Origins of Intelligence in Children.* New York: International Universities Press. (Originally published 1936.)

Powell, S. R., & Fuchs, L. S. (2010). Contribution of equal-sign instruction beyond word-problem tutoring for third-grade students with mathematics difficulty. *Journal of Educational Psychology, 102*(2), 381-394.

RAND Mathematics Study Panel. (2003). *Mathematical proficiency for all students: Toward a strategic research and development program in mathematics education*. Santa Monica, CA: RAND Corporation MR-1643.0-OERI

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* 2nd edition. Newbury Park, CA: Sage.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development, 77*(1), 1-15.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology, 91*(1), 175-189.

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., & McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A construct-modeling approach. *Journal of Educational Psychology, 103*(1), 85-104.

Roberts, M. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice, 29*(3), 25-38.

Rosnick, P. (1981). Some misconceptions concerning the concept of variable. *The Mathematics Teacher, 74*(6), 418-420.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). Diagnostic measurement: Theory, methods, and applications. New York: The Guilford Press.

Russell, M., O'Dwyer, L.M., & Miranda, H. (2009). Diagnosing students' misconceptions in algebra: Results from an experimental pilot study. *Behavior Research Methods, 41*(2), 414-424.

Saenz-Ludlow, A., & Walgamuth, C. (1998). Third graders' interpretations of equality and the equal symbol. *Educational Studies in Mathematics, 35*(2), 153-187.

Scheuneman, J. D., & Slaughter, C. (1991). *Issues of test bias, item bias, and group differences and what to do while waiting for the answers* [ERIC]. Educational Testing Service. ERIC Number ED400294.

Schifter, D. (1999). Reasoning about Operations: Early Algebraic Thinking, Grades K through 6. In L. Stiff and F. Curio, (Eds.) *Mathematical Reasoning, K-12: 1999 NCTM Yearbook.* (pp. 62-81). Reston, VA: National Council of Teachers of Mathematics.

Schifter, D., Russell, S. J., & Bastable, V. (2009). Early algebra to reach the range of learners. *Teaching Children Mathematics, 16*(4), 230-237.

Schoenfeld, A., & Arcavi, A. (1988). On the meaning of variable. *The Mathematics Teacher, 81*, 420-427.

Seo, K.-H., & Ginsburg, H. P. (2003). "You've got to carefully read the math sentence…": Classroom context and children's interpretations of the equals sign. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills* (pp. 161–187). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Sfard, A., & Linchevski, L. (1994). The gains and the pitfalls of reification: The case of algebra. *Educational Studies in Mathematics, 26*(2/3), 191-228.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Shepard, L. A. (1982). Definitions of bias. In R.A. Berk (Ed.), *Handbook of Methods for detecting test bias.* Baltimore, MD: The Johns Hopkins University Press.

Sherman, J., & Bisanz, & J. (2009). Equivalence in symbolic and nonsymbolic contexts: Benefits of solving problems with manipulatives. *Journal of Educational Psychology, 101*(1), 88-100.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8,* 481-520.

Sleeman, D., Kelly, A. E., Mortinak, R., Ward, R. D., & Moore, J. L. (1989). Studies of diagnosis and remediation with high school algebra students. *Cognitive Science, 13,* 551-568.

Smith III, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2), 115-163.

Smith, J., & Thompson, P. W. (2008). Quantitative reasoning and the development of algebraic reasoning. In J. J. Kaput, D. W. Carraher & M. L. Blanton (Eds.), *Algebra in the early grades* (pp. 95-132). New York: Erlbaum.

Stacey, K. (1989). Finding and using patterns in linear generalizing problems. *Educational Studies in Mathematics, 20,* 147-164.

Statistical Package for the Social Sciences (2006). SPSS Version 15.0. SPSS Inc, Chicago: IL.

Stephens, A. (2005). Developing students' understandings of variable. *Mathematics Teaching in the Middle School, 11*(2), 96-100.

Stephens, A. C. (2006). Equivalence and relational thinking: Pre-service elementary teachers' awareness of opportunities and misconceptions. *Journal of Mathematics Teacher Education, 9,* 249-278.

Stephens, A. C. (2008). What "counts" as algebra in the eyes of preservice elementary teachers? *Mathematical Behavior, 27,* 33-47.

Tanish, 2. (2011). Functional thinking ways in relation to linear function tables of elementary school students. *The Journal of Mathematical Behavior, 30*(3), 206-223.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345-354.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Fredrickson, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum.

Taylor, C. S., & Lee, Y. (2011). Ethnic DIF and DBF in reading tests with mixed item formats. *Educational Assessment, 16,* 1-34.

Taylor, C. S., & Lee, Y. (in press). Gender DIF in tests with mixed item formats. *Applied Measurement in Education.*

Trends in International Mathematics and Science Study (TIMSS). (2007). *2007 Results*. Retrieved from http://nces.ed.gov/timss/results07.asp

U.S. Department of Education. (2007). *Digest of Education Statistics: 2007*. National Center for Education Statistics. Retrieved from http://nces.ed.gov/programs/digest/d07/.

Usiskin, Z. (1988). Conceptions of school algebra and uses of variables. In A. F. Coxford (Ed.s), *The Ideas of Algebra K-12* (pp. 8-19). National Council of Teachers of Mathematics: Reston, VA.

Usiskin, Z. (1995). Why is algebra important to learn? *American Educator*, 30-37.

van der Linden , W. J. & Hambleton, R. K. (1996). *Handbook of Modern Item Response Theory*. New York: Springer.

Van de Walle, J. A., Karp, K. S., & Bay-Williams, J. M. (2010). *Elementary and middle school mathematics: Teaching developmentally* (4th Ed.). New York: Pearson.

Wagner, S., & Kieran, C. (1989). An agenda for research on the learning and teaching of algebra. In S. Wagner & C. Kieran (Eds.), *Research issues in the learning and teaching of algebra* (pp. 220-237). Reston, VA: National Council of Teachers of Mathematics.

Waite, R. D. (2000). A study of the effects of *Everyday Mathematics*® on student achievement of third-, fourth-, and fifth-grade students in a large north Texas urban school district. *Dissertation Abstracts International, 61*(10), 3933A.

Walker, C. M., Zhang, B., & Surber, J. (2008). Using a Multidimensional Differential Item Functioning framework to determine if reading ability affects student performance in mathematics. *Applied Measurement in Education. 21,* 162-181.

Wang, J. (2005). Relationship between mathematics and science achievement at the 8th grade. *International Online Journal of Science and Mathematics Education, 5,* 1-17.

Warren, E., & Cooper, T. (2005). Introducing functional thinking in year2: A casey study of early algebra teaching. *Contemporary Issues in Early Childhood, 6*(2), 150-162.

Warren, E., & Cooper, T. J. (2008). Generalising the pattern rule for visual growth patterns: actions that support 8 year olds' thinking. *Educational Studies in Mathematics*, *67*(2), 171-185.

Warren, E. A., Cooper, T. J., & Lamb, J. T. (2006). Investigating functional thinking in the elementary classroom: Foundations of early algebraic reasoning. *Journal of Mathematical Behavior, 25,* 208-223.

Washington State K-12 Mathematics Learning Standards. (2008). *K-8 Algebra Strand*. Office of Superintendent of Public Instruction.

Weaver, J. F. (1973). The symmetric property of the equality relation and young children's ability to solve open addition and subtraction sentences. *Journal for Research in Mathematics Education, 4*(1), 45-56.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. *Handbook of Statistics, 26,* 4-44.

Wiliam, D., & Black, P. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment. *British Educational Research Journal, 22*(5), 537-548.

Yang, D., Reys, R. E., & Wu, L. (2010). Comparing the development of fractions in the fifth- and sixth-graders' textbooks of Singapore, Taiwan, and the USA. *School Science and Mathematics, 110*(3), 118-127.

Yee, L. P., & Hoe, L. N. (2009). *Teaching Primary School Mathematics: A Resource Book*. McGraw Hill Education: Singapore.