Education Dissertations   Education, School of

January 1st, 2018

Smarter Balanced Math Assessment: An Examination of Its Construct Validity for Students with Known Reading and/or Writing Disabilities

Dana A. Bailey
Seattle Pacific University

Follow this and additional works at: https://digitalcommons.spu.edu/soe_etd

Part of the Special Education and Teaching Commons

Smarter Balanced Math Assessment: An Examination of Its Construct Validity

for Students with Known Reading and/or Writing Disabilities

by

Dana A. Bailey

Seattle Pacific University

Smarter Balanced Math Assessment: An Examination of Its Construct Validity

for Students with Known Reading and/or Writing Disabilities

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Education

Seattle Pacific University

2018

Approved by:

_____
(Dr. Eigenbrood, Chairperson, Dissertation Committee)

_____
(Dr. Preciado, Committee member)

_____
(Dr. Beers, Committee member)

Program Authorized to Offer Degree: _____SCHOOL OF EDUCATION_____

Date:_____APRIL 2018_____

_____
(Dr. Eigenbrood, Dean, School of Education)

Table of Contents

List of Tables

List of Appendices

Acknowledgments

I am grateful for the guidance and wisdom of Dr. Eigenbrood, for his assistance through this daunting project. I am equally grateful for the support of the committee members Dr. Preciado and Dr. Beers. Many guides and supporters coached me along the way, chiefly Dr. Steve Simpson, who unfailingly agreed whenever I needed help; technical help in writing or emotional support to keep going - I am forever in your debt. My classmates provided a constant flow of encouragement and support, through our classes and this final big step. Thank you Lisa, Leanna, Ronak, Kaley, Cari and Nalline for the reality checks, answers to late night questions and quality help throughout.

I would never have embarked on this journey without unfailing support and encouragement of my husband. Thank you for providing me the time and space to work hard and achieve this goal. Your efforts to carry more of the load has not gone unnoticed and your patience has been limitless. To my children and extended family - I know I have been more absent and preoccupied through this process than we all would have liked. Thank you in equal measures for your persistent support and the much needed source of distraction. And to my parents, my first and best teachers, your wisdom and care has buoyed me throughout this and all my endeavors.

I am truly blessed to be encouraged and inspired by you all.

Abstract

The evolution of standards based assessment, beginning with the No Child Left Behind Act (NCLB, 2001) and progressing into the Every Student Succeeds Act (ESSA, 2015), maintains high stakes testing as an expected component of the public school experience. For students with specific learning disabilities (SLD) in reading and/or writing, meeting standard on the currently mandated Smarter Balanced Math Assessment (Office of Superintendent of Public Instruction [OSPI], 2018; SBAC, 2018) presents unique challenges. On the Smarter Balanced Assessment (SBA), students are expected to solve math problems that require language skills, specifically skills which are deficient in students with learning disabilities in reading and/or writing. In particular, Claim 3 (SBAC, 2018) requires students to *explain your thinking*, and *construct and evaluate an argument*. These are examples of test questions that, for a student with learning disabilities, may be difficult to answer regardless of math ability. Although SBA authors attempted to mitigate specific learning disabilities with allowable testing accommodations, one might ask, if the student is unsuccessful on these math questions, is it due to lack of math ability or due to the impact of the specific learning disability in reading and/or writing? Examining elements of non-construct variance and construct validity is worthy of investigation for secondary students, for whom meeting standard on these required exams holds lasting consequences. This paper will explore the elements of non-construct variance and construct validity of achieved math SBA scores as it pertains to students with documented learning disabilities in reading and/or writing.

Chapter I: Introduction

In the current standards-based movement, mandated high stakes tests are prevalent for most students. Currently, Common Core State Standards (CCSS) and the Smarter Balanced Assessments (SBA) are mandated in 46 states (Smarter Balanced Assessment Consortium, [SBAC], 2018). For students with specific learning disabilities (SLD) in reading and/or writing, performance on the required math section of the SBA may be impacted by these known disabilities. Challenges might manifest from the linguistic complexity or language demand of the questions, or the requirement to write extended responses in answering the questions. For middle school students, the SBA math test requires students to construct an argument, evaluate a response, and disaggregate application problems. This is especially evident in Claim 3 of the math section, Communicating Reasoning (SBAC, 2018). This claim evaluates students' abilities to "clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others" (Smarter Balanced Content Specifications, 2015, p.5). In layman's terms, a student must have the language skills necessary to effectively communicate his or her math ability.

Validity of the SBA math score is examined in this study in two ways for students with reading and/or writing disabilities. One is construct validity: Do the math questions require language skills to answer, and if so, to what extent is a learning disability in reading and/or writing impacting the ability to achieve in mathematics? The second is the impact of allowable testing accommodations on resulting individual achievement scores. Do these legal and required accommodations remove barriers to achievement and thus enhance the validity of the resulting scores, or does the use of accommodations

decrease validity because protocols have been altered? Results of these high stakes tests have educational consequences for students when the results are used as evidence in high stakes decision-making (Cumming, 2008; Haertel & Lorie, 2004). In the standards-based accountability and data driven decision-making culture that exists today in public schools, solid answers to these questions are imperative to strengthen the validity and interpretability of the results of the SBA for students with disabilities. The anticipated date of the math SBA becoming a graduation requirement is 2019, and thus, validity and interpretability of scores is of interest to students with disabilities, their parents and district policy makers.

Theoretical Framework

Two important theoretical frameworks provide underpinnings for this examination. The constructivist theory of learning, primarily the works of Piaget and Inhelder, describe the processes that children undergo as they learn new concepts. The evolution of understanding about how some students learn differently from typically developing peers is briefly explained. The theories of assessment, in particular of the assessment of students with known disabilities, add support to the overall framework of how students learn. Together these provide the structure for the essence of this examination: How can educational professionals and policy makers be assured that the current iteration of standardized assessments is both fair and valid for students with specific learning disabilities?

Theories of learning: Piaget and Inhelder. Generally perceived as a cognitive constructivist, Piaget studied the relationship between the assimilation and accommodation of concepts. Assimilation occurs through lived experiences while accommodation is the development of new mental structures or schema. Through the

interplay of assimilation and accommodation, when newly acquired knowledge conflicts with that which was previously learned, an intellectual state of disequilibrium exists (Piaget, 1953). The act of learning is described as working and manipulating new information to fit or expand the prior schema, or structure, of all that had been previously learned. Constructivists such as Piaget (1953) supported this theory of learning that knowledge is framed or constructed through experiences and connections are made between new information and all that had been previously learned.

A component of Piaget's theory that set his work apart from other constructivists was his explanation that learning happens at distinct stages of life (Piaget, 1953). As a child moves through the phases of sensorimotor, preoperational, concrete operation, and formal operation, the child continues to create a schema that is ever-expanding in response to the new information gained by experiences and the environment. This framework, coined *global states of learning*, was presented early in Piaget's work and focused on the type of logic used at each developmental level. Later in his writings, he and others focused on the mechanism of learning – the process that enabled new constructions, new perspectives, and new learning to occur. Piaget drew on his previous work as a biologist to craft the idea that cognitive structures are under construction as people experience the world (Piaget, 1970).

Piaget posed three modes of equilibrium resolution in the process of learning. The first mode was the assimilation of new ideas into the previous structure. In this model, new ideas or information fit or blended into the previous structure without dissonance. In the second model, the learner was presented with two logical ideas in which the learner found contradicting information that did not immediately make sense.

In the practice of individually experiencing cognitive dissonance, and building schema to make sense of new information, the learner practices building and constructing ever-expanding cognitive structures. The third mode of equilibrium resolution was the integration of an entire knowledge structure, uniting two systems of thought in totality, creating an intricate and more complex schema of understanding. It is through building these intricate and complex schema of understanding that learners continually learn.

Piaget sought to describe structures common to all children. Inhelder and other Genevan theorists used his theoretical framework to examine learning differences in individual children (Gallagher & Reid, 1981; Gazda & Corsini, 1980). Further, Genevan theory supplied a developmental sequence not tied to age, but rather focused on the mechanism of change. Inhelder, Sinclair, and Bovet (1974) specifically investigated the transition from one stage of cognitive development to the next, and the constant or predictable order of the succession of stages. Through this investigation, Inhelder et al. (1974) noted that "even in cases where there appear to be deviations in development, basically the operatory system is always constructed in the same way" (p. 17). The different speeds of acquisition of concepts seemed to follow the stages of development, and many students with perceived learning disabilities, especially in reading, followed the same path as typically developing peers. Gallagher and Reid (1981) reinforced this theory, claiming "all children except the seriously emotionally disturbed follow normal development patterns" (p. 171). The learning theory of Piaget that children move through the stages of development in the same order, coupled with the work of Inhelder and others, that individual children with learning difficulties follow the same path but at a

slower rate, laid the foundation for educators and researchers to further explore how students with learning disabilities learn differently (Piaget, 1970).

Theories of assessment: Ysseldyke, Christenson, and Thurlow. Large scale assessments are constructed to serve different purposes. The current iteration of the federally mandated assessment serves two purposes: accountability for districts, and soon individual student accountability in the form of a graduation requirement. Assessments used for accountability reasons provide feedback to districts about overall achievement for individual schools and districts as a whole. In order to draw sound conclusions from the results, several elements must be met with care. In the assessment of students with disabilities in particular, quality assessments for accountability measures need to measure student achievement not solely on high stakes tests, but also on other factors such as high school graduation rates, college acceptance, and college graduation rates, all with the focus on individual achievements (Gunzenhauser & Hyde, 2007). Another important element is understanding the negative consequences of having accountability systems that do not include all students (Pemberton, Rademacher, Tyler-Wood, & Perez Cereijo, 2006; Thurlow, House, Scott, & Ysseldyke, 2000). These consequences of excluding students with disabilities from large scale assessments intended to analyze and reform school systems include increased retention practices, increased rates of referrals into special education, unfair or misleading comparisons between districts, and vague practices on the reporting of students not taking the test.

Ysseldyke, Christenson, and Thurlow (1987) described ten factors that were important for reporting on individual student achievement. Included was the degree to which student performance was evaluated appropriately and frequently. To achieve this

goal, assessments must be both frequent and congruent with that which is being taught. With these two critical elements in place, evaluation results serve to validly inform the teacher about student progress, and provide essential information for effective decision-making about subsequent instruction.

Furthermore, assessments should be valid for those with whom they are used (Thurlow et al., 2000). Not all tests can be used as valid assessments for students with disabilities, and if the test is to be used to demonstrate accountability measures for students with disabilities, these students should be present in the norming process (Overton, 1992; Thurlow, Quenemoen, & Lazarus, 2011). Thurlow et al. (2011) claimed that quality assessments for the CCSS must include quality assessments for all students, including students with disabilities who receive special education services. Educators, parents, policy-makers, and citizens need to know the extent to which all students, including students with disabilities, are profiting from their educational programs (Ysseldyke & Bielinski, 2002). Students with disabilities have a unique set of characteristics, and assessments used for students with disabilities should be valid, and transparent enough to know when a student is inaccurately measured because of poor basic skills but concurrently may have stronger high level thinking skills (Thurlow et al., 2011).

These theoretical frameworks underscore the importance of attending to the learning needs of individual students as related to standardized assessment practices. Students with disabilities must be included in the norming process and visible in the results. Learning theories support that children move through cognitive development in predictable stages, and initial intelligence testing attempted to measure aptitude by

quantifying ability relative to chronological age. Disparity between expected growth and actual growth became the underpinnings of the evaluative measures that quantified the existence of a disability and the need for specialized instruction. Finding these differences is the foundation behind the selection or battery of examinations that psychologists use to determine the presence of a specific learning disability. The field of special education evolved on the premise that while students with disabilities follow the Piagetian stages of development, these stages appear at a later chronological age (Weekly, 1979). Students with disabilities must be evaluated in the current paradigm using the correct grade assigned assessment for validity in measurement and accountability, but it is yet an unanswered question if the current assessment is fair and valid for students with certain types of disabilities.

High Stakes Testing for Students with Disabilities

Special education eligibility assessment is a highly regulated and inherently discriminatory process, one which is laden with high stakes testing implications (Bayles, 2009). Students with perceived disabilities undergo specific and detailed high stakes testing in the special education assessment process (Kauffman, Hallahan, & Pullen, 2011). Using results of these tests, school psychologists and the special education team determine if a student does or does not have a qualifying disability. If the outcome is affirmative, the student is entitled to special education services, which are additional educational services intended to support learning within the constructs of a documented specific learning disability. Once placed in special education, students are provided specially designed instruction and granted access to accommodations that support individualized learning and access to education. These accommodations include both

learning and assessment experiences, and are documented in the student's Individual

Education Plan (IEP), a legal document binding a school system to provide the additional

supports necessary for individual student success. It is important to note that students

with specific learning disabilities have significantly different academic needs from

students who have intellectual disabilities. The chief indicator of this difference is that

students with specific learning disabilities typically have average or above average

intelligence. Students with specific learning disabilities in one category, for example in

writing only, typically have commensurate skills in other academic areas similar to non-

disabled peers (Kauffman et al., 2011). Through the special education evaluation process,

it is most likely that all categories of concern were evaluated, and thus a student with a

specific learning disability in reading and/or writing would have documented evidence

that he or she did not meet the criteria for a specific learning disability in math. Of

importance to this study is the distinction and clarity around the specific focus on

students with learning disabilities in reading and/or writing.

Testing designed and used for entry into special education is not the only testing

required of students served in special education. All public school students in

Washington and 45 other states are required to take the SBA in grades three through eight

and grade eleven annually (OSPI, 2018; SBAC, 2018). As such, students served in

special education with specific learning disabilities are evaluated as meeting or not

meeting standard using the same achievement markers as non-special education students.

Allowable and approved testing accommodations are legally obligated and accessible for

students with disabilities through the IEP process. The Smarter Balanced Assessment

Consortium (SBAC) authors considered these accommodations in the validity and

reliability statement for the assessment (Smarter Balanced Technical Report [SBTR], 2016). In the development and pilot of the SBA, test authors claimed that one intent of the Smarter Balanced system was to remove construct irrelevant barriers that prevented students from demonstrating their best performance (SBTR, 2016). Students with disabilities were specifically included in this statement, and elements that applied to students with physical disabilities were elucidated. In the purpose statement of the same document, the SBA test authors claimed that an important outcome of the SBA was to provide valid, reliable, and fair information concerning students' achievement in literacy and mathematics that was equitable for all students and subgroups of students (SBTR, 2016, p. 51). Test authors admitted the need to understand the "characteristics and needs of students with disabilities and address ways to design assessments and provide accommodation to get around the barriers created by their disabilities" (SBTR, 2016, p. 10), but did not specifically describe the accommodations that were intended to accomplish this goal. This need for valid, reliable, and fair information, and the quality assessment practices that require that all students and sub-groups of students be visible in the norming process and the reporting practices, is an important element of high stakes, standardized testing practices of today's schools.

A significant measure in the 1997 reauthorization of Individuals with Disabilities Education Act (IDEA, 1997), which was reaffirmed in 2004 (IDEA, 2004), required that students with disabilities be allowed and/or mandated the opportunity to access the same assessment opportunities as non-disabled peers. As previously stated, sixth, seventh and eighth grade students are required to take four at-grade-level Smarter Balanced Assessments. This requirement is held in tension with special education law that requires

students with disabilities receive specially designed instruction at their individual level of academic development. The net result is that students served in special education are experiencing tests that assess content not yet presented, or skills not yet learned. If a school system has a process or policy of moving students who are not proficient on the mandated test into remedial tracts, or using these scores for the entrance or denial of higher level courses, these students could be denied access based on a handicapping condition, which may be in direct violation of Section 504 of the Rehabilitation Act of 1973 (Rehabilitation Act, 1973).

A confluence of demands adds pressure to the already tenuous standardized assessment conditions. Federal regulations designed to ensure equitable access to rigorous standards and assessments for all students, challenge students, teachers, and school administrators to balance these needs simultaneously (Jamgochian & Ketterlin-Geller, 2015). Specific adherence to special education law requires students with disabilities be given the same opportunities as non-disabled peers, including participation in the mandated standardized testing. The SBA is assigned at grade level for students in special education, rather than at developmental, instructional, or academic level. Some students with learning disabilities in reading and/or writing are thus ill-prepared to show mastery on high stakes tests, specifically tests that cover information not yet learned (Jamgochian & Ketterlin-Geller, 2015). Testing accommodations used to support the individual needs of the students are intended to mitigate the impact of the disability of the construct being assessed. Absent from current research is the impact of a specific learning disability in reading and/or writing on individual student scores for the math Smarter Balanced Assessment.

Validity and Interpretability of Results of High Stakes Tests

Central to the problem of mandated assessments for students with disabilities is the objective interpretation of results. If school officials are to use the outcomes of annual mandated testing to withhold high school graduation, school officials must be certain that the stated result is both valid and reliable. Validity generally refers to how well a test measures what it claims to measure (Vogt, 2005). It is vital that an assessment be valid for the results to be accurately interpreted and applied. Reliability generally refers to a test being free from measurement or random error (Vogt, 2005). Reliability can be supported with confidence intervals, or the range of values with a known probability of including the true value (Vogt, 2005). In layman's terms, the confidence interval is a range of scores that are highly likely to include the true score.

Of interest to this study is construct validity, or how accurately items on the test measure the stated construct or topic of the individual question. Threats to construct validity include extraneous variables that compete with the independent variable; in this case, does a student's reading ability interfere with the test's accuracy in measuring that student's mathematic skills? Addressing threats to validity, especially examination of confounding variables, strengthens the interpretation of results. Test authors expect and account for a certain degree of random variance, variance that is evident and evenly spread over the data set (Field, 2013; Vogt, 2005). However, when specific variance is evident within a data set, impacting a specific group of students, and is not evenly or randomly distributed, this skews the data in a specific direction and makes interpretation less valid. (Field, 2013; Vogt, 2005). The question remains, does the math SBA math test assess only math skills or do language skills confound the resulting scores?

Statement of Problem

In the current and anticipated paradigm of state assessments, students served in special education face challenges to meet these requirements. The specific focus of this study is the impact of a specific learning disability in reading and/or writing on the student's ability to achieve on the required math assessment. Testing accommodations for individual students are allowable on all sections of the Smarter Balanced Assessment (SBAC, 2018). The underlying premise is that if testing accommodations are appropriately applied, and the student does not have a disability in mathematics, the reading and/or writing disability would be mitigated. As this is a relatively new assessment, empirical research regarding the SBA concerning students with SLD is limited.

Purpose of the Study

The primary purpose of this research is to examine the existence of a statistically significant difference between a group of middle school students with disabilities in reading and/or writing and students without these disabilities on SBA math scores. One concept to be investigated is construct validity, to further explore if the math questions on the SBA also inadvertently assess the ability to read, write, and communicate, and to what level provided accommodations mitigate learning disabilities in reading and/or writing. A second concept to be investigated is the existence of a compound influence for a student who has both reading and writing disabilities. A third concept to explore is to examine if the unique confidence interval, reported with the achieved score, can be used as a differential boost to improve interpretability for students with disabilities in reading and/or writing.

This study is intended to evaluate the presence of non-construct variance in the SBA math test for students with reading and/or writing disabilities. Research methods are designed to illuminate if systematic variance is present on the math SBA scores for certain groups of students.

Significance of the Study

Results of this study contribute to the substantive and practical significance in the area of validity and reliability in the assessment of students with specific learning disabilities in reading and/or writing. Substantively, the exploration of the variables in this study begin to build the body of evidence in the authentic assessment of students with specific learning disabilities. The exploration of the variables highlights the nuances of assessment where systematic variance can disadvantage a specific group of students. This study offers practical significance to school personnel required to interpret resulting scores in high stakes systems, especially regarding impending graduation requirements and other policy decisions. Deeper understanding of if, and how, reading and writing disabilities impact success on math assessments will aid classroom teachers in strategic teaching and focused intervention. If a statistically significant relationship exists between specific learning disabilities in reading and/or writing on SBA math scores, educators can advocate for equity in meeting local graduation requirements. The results of this study may confirm for special education teachers the important role that reading and writing play in meeting proficiency indicators in mandated math assessments, especially given the language requirements of Claim 3 on the SBA: Communicating Reasoning (SBAC, 2018). While research has been conducted on this same construct for English Language

Learners (ELL's), research is less robust in the application for students with specific

learning disabilities.

Terms and Definitions

1) Accommodations: Supports and services provided to help a student access the

   general education curriculum and validly demonstrate learning.

   Accommodations do not substantively alter the constructs being taught or

   assessed. In this study, accommodations are narrowly defined as supports

   deemed appropriate by test authors and selected by the Individual Education

   Plan team.

2) Confidence interval: Additional score reported by the Smarter Balanced

   Assessment Consortium to quantify the relative accuracy of the stated

   achievement score. This score is reported as a range of positive or negative

   points above and below the stated achievement score that represents a

   mathematical calculation range of which will likely include the true score.

3) Differential boost: When a testing accommodation produces a greater gain in

   scores for students with disabilities than for students without disabilities.

4) High stakes testing: A test in which the resulting scores are used to determine

   weighty consequences such as graduation, entrance or denial to advanced

   courses or programs.

5) Individual Education Plan (IEP): The legal document that describes the

   academic need and outlines an annual academic plan for a student with a

   disability.

6) Non-construct variance: Variance in statistical procedures that is not explained by the construct being measured.

7) Random error: Expected error within an assessment and cannot be eliminated. This error is assumed to have random distribution over the population sample.

8) Smarter Balanced Assessment (SBA): Current iteration of the federal mandated assessment required of all public schools in 46 states.

9) Specific Learning Disability (SLD): Difficulty with a specific learning task for an individual with typical intelligence resulting in a discrepancy between perceived ability and actual performance.

Research Questions

1) Is there a statistically significant difference in SBA math scores between students with reading disabilities and students who do not have disabilities?

2) Is there a statistically significant difference in SBA math scores between students with writing disabilities and students who do not have disabilities?

3) Is there a statistically significant difference in SBA math scores between students with reading and writing disabilities and students who do not have disabilities?

Hypotheses

Null hypothesis 1: There is no statistically significant difference in SBA math scores between students with specific learning disabilities in reading and students without disabilities.

Null hypothesis 2: There is no statistically significant difference in SBA math scores between students with specific learning disabilities in writing and students without disabilities.

Null hypothesis 3: There is no statistically significant difference in SBA math scores between students with specific learning disabilities in both reading and writing and students without disabilities.

Research Design

This study was designed to evaluate if there was a construct-based difference for students with specific learning disabilities in reading and/or writing on the math SBA. As such, an Analysis of Variance (ANOVA) with four groups was most appropriate. A convenience sample was be used; all middle school students from a large suburban Washington State school district were potential participants. The control group consisted of randomly selected students not served in special education, and the comparison group consisted of 111 students with specific learning disabilities in reading and/or writing. These comparison groups were further defined as students with reading, writing, and combined reading and writing disabilities as outlined by Washington State discrepancy scores (American Psychiatric Association [APA], 2014; OSPI, 2018). All students in the comparison group had a current and compliant Individual Education Plan. This ensured that the researcher was reasonably assured that students were appropriately placed in special education. All selected participants were in the assigned grade level math class, attempting to address a confounding variable of individual math skill and prior instruction; therefore, students in advanced and remedial general education math classes were excluded. SBA math scores for students in grades six through eight were be evaluated. The dependent variable is the continuous score on the SBA math assessment, and the independent variable is the presence or absence of a documented language disability.   Groups were defined as (1) control (not special education); (2) students with a

disability in reading; (3) students with a disability in writing; and (4) students with disabilities in reading and writing.

Limitations of study

The effects of test anxiety, test fatigue, and motivation are all elements that influence test success (Carter et al., 2005; Haladyna & Downing, 2004); however, these potential influencers were not investigated as part of this study. While all students participated in these assessments under the same general environment as dictated by the test directions for administration, students uniquely experience testing. The specificity of focus for students with only reading and/or writing disabilities resulted in relatively small groups. Generalizability of these results to other groups dissimilar to the studied group should be made with caution.

Summary and structure of dissertation

This chapter provided an overview of the introductory elements of this dissertation study, including the background, theoretical basis, problem statement, and purpose of study. Four subsequent chapters elaborate on the literature review, research methods, results, and discussion of results. Chapter Two examines in context the influences of high stakes testing, the language of mathematics, validity, and the concept of construct irrelevant variance. The examination of validity includes a discussion of testing accommodations, as this is a component of test structure, of analysis, and specifically of the SBA validity statement. Chapter Two also includes a review of the limited available research on the impact of language disabilities on math achievement, and includes parallel studies involving English Language Learners on the same topic. Chapter Three provides a description of the methods for this study, including the

statistical methods and data analysis. Chapter Four provides a summary of the results of the study, including descriptive and inferential statistics related to the research questions. Chapter Five presents a discussion of the statistical and practical significance of the findings and the implications for practitioners. Limitations and threats to internal and external validity are discussed. Suggestions for future studies that examine the additional impact of the use of testing accommodations, as well as implications of the unique student confidence interval, reported with student SBA math scores are offered.

Chapter 2: REVIEW OF THE LITERATURE

Introduction

This chapter provides a review of existing literature on the evolution of federally mandated tests, and the interpretation of resulting scores, and situates the assessment of students with disabilities within this trend. Arguments against high stakes testing are presented, including the question of evaluating high-level thinking with low-level assessment formats such as multiple-choice questions. Influences of linguistic complexity, language demand, and the language of mathematics are presented as relevant to this study of mathematics assessment. Use of testing accommodations as mitigating elements to the validity of the assessment results are discussed. Validity, specifically construct validity, is discussed with supporting empirical studies. Research for English Language Learners (ELL's) is more robust than research for students with disabilities on this topic and is presented to provide parallel comparisons.

Historical Elements of Special Education

As early as the 1800s, schools specially designed to educate or support individuals with disabilities were built (Kauffman, Hallahan, & Pullen, 2011; Hallahan & Mercer, 2001). In the late 1800's, researchers, educators, and medical professionals were beginning to examine, quantify, and name instances where students seemed to learn differently. Scientists and practitioners such as Adolf Kussamaul wrote on his notion of *word blindness*, which was followed by Rudolf Berlin, who added specificity to the definition and is noted to be the first to use the term *dyslexic* (Hallahan & Mercer, 2001). In 1905 Ophthalmologist W.E. Brunner introduced in the United States a report on childhood reading difficulties. Other educators such as Marion Monroe and Samuel Kirk

continued the work to define and provide educational insight for students with disabilities

(Kauffman et al., 2011). Monroe introduced the discrepancy concept as a way of

identifying students with disabilities (Hallahan & Mercer, 2001). Most authorities cite

Kirk as the originator of the term *learning disabilities* (Hallahan & Mercer, 2001).

Although the definition of learning disabilities evolved over time, his was the first

attempt at defining the elements that were uniquely different from other sources of

disability manifestations. Kirk also developed an assessment approach for pinpointing

specific learning disabilities in children, the Illinois Test of Psycholinguistic Abilities

(Hallahan & Mercer, 2001). Kirk's original definition follows:

> Children with special specific learning disabilities exhibit a disorder in one or more of
>
> the basic psychological processes involved in understanding or in using spoken and
>
> written language. These may be manifested in disorders of listening, thinking,
>
> talking, reading, writing, spelling or arithmetic. They include conditions which have
>
> been referred to as perception handicaps, brain injury, minimal brain dysfunction,
>
> dyslexia, developmental aphasia, etc. They do not include learning problems that are
>
> due primarily to visual, hearing or motor handicaps, to mental retardation, emotional
>
> disturbance or to environmental disadvantage. (U.S. Office of Education 1968, p. 34)

Bateman (1992) offered her own definition. This definition proved to be significant

because it reintroduced Monroe's earlier notion of measuring discrepancy between

observed achievement and expected potential as a way to identify students with learning

disabilities. Bateman's (1992) definition follows:

> Children who have learning disorders are those who manifest an educationally
>
> significant discrepancy between their estimated potential and actual level of

performance related to basic disorders in the learning process, which may or

may not be accompanied by demonstrable central nervous system dysfunction,

and which are not secondary to generalized mental retardation, educational or

cultural deprivation, severe emotional disturbance or sensory loss. (p. 34)

The work of these early scientists, medical professionals, and teachers helped frame

thinking around how students learn, how some students learn differently, and how

education and assessment might be inherently different for students with learning

disabilities.

Evolution of Standardized Tests as Accountability Measures

Tests that have been constructed and field tested to ensure a high degree of

reliability and validity are called standardized tests (Tileson, 2004). In an effort to raise

the quality of educational achievement for American students, the system of tests that

merely measured comparative, or normed, progress of students had to be changed to one

in which the content knowledge was specified first, and then student progress was

measured against how much of that knowledge had been acquired (National Council on

Education Standards and Testing, 1992). This recommendation began the system of

accountability and nationally mandated standardized tests. The goal of these new

standard exams was to promote high expectations rather than minimum competencies,

provide focus but not a prescribed curriculum, and be nationally administered,

mandatory, and dynamic. National standards were necessary to ensure educational

opportunity and equity for all young Americans, especially those not doing well in school

due to low expectations (National Council on Educational Standards and Testing, 1992).

Administrators and school personnel use these standardized tests to evaluate students and

the quality of programs, schools, and districts (Berlak et al., 1992). Tileson (2004)

suggested that using standardized tests allows educators the ability to compare the scores

of students, determine whether the students are making sufficient progress, and make

decisions about teaching and learning.

The progression of legislation with the intention of improving education for all

children in the United States began with the Elementary and Secondary Education Act

(ESEA) originally signed into law by President Lyndon B. Johnson on April 11, 1965

(ESEA, 1965). While the major focus of this legislation was to improve the education of

students living in poverty, it began the evolution of mandated standardized testing

intended to build rigor in education, and provided accountability measures for schools

and districts. Reauthorized by President Carter in 1978 and President Reagan in 1981,

these efforts strengthened the focus on basic education, allocated financial resources to

the schools in most need, and provided local control of educational resource allocation.

The reauthorization in 1994 by President Clinton included the Improving America's

Schools Act, which mandated the creation of core area standards and accountability

assessments for each. President G.W. Bush continued this education reform in the 2002

reauthorization of ESEA and renamed it No Child Left Behind (NCLB). The significant

measures of this bill included the mandate for annual testing in selected grades, specific

focus on underperforming groups, and protocols for reporting achievement scores to the

public. From this legislation, states were required to create standardized assessments. In

the context of the current study, the Washington Assessment of Student Learning

(WASL) commenced in 1997, became a graduation requirement in 2006, and was

replaced by the High School Proficiency Exams (HSPE)/Measurement of Student

Progress (MSP) in 2010. The HSPE/MSP became a graduation requirement in 2012 for

Washington students and was used until 2015. Subject-specific End of Course (EOC)

exams in math and science were initiated in 2012 and continued to be used as a

graduation requirement for students graduating in 2017 (The Washington State Board of

Education, 2017).

Additional legislation in the 1997 reauthorization of the Individuals with

Disabilities Education Act (IDEA) provided legislation that required school districts to

include students with disabilities in mandated state tests as a function of access (IDEA,

1997; IDEA, 2004). These accountability measures continued to evolve, become more

specific, and carry weightier consequences for both students and school districts.

Thurlow, Lazarus, Thompson, and Morse (2005) claimed out-of-level testing, while

allowable by state control, became scores not included in the aggregate scores of the

student population, effectively removing students with disabilities tested in this manner

from reporting practices. While these out-of-level students were being granted the access

to take the test, their scores were not being represented in the aggregate score reports.

The moderate rate of transition in and out of special education presented a different issue

as it revealed an inconsistency of how students were categorized (Ysseldyke & Bielinski,

2002) and this flux to group membership further confounded the attempt to reveal

progress for groups of students with disabilities.

The most recent reauthorization of the ESEA was signed by President Obama on

December 10, 2015 and renamed Every Student Succeeds Act (ESSA). In a continued

national effort to strengthen and measure the success of school programs, the Common

Core State Standards (CCSS, 2018) have been adopted by 46 states . In alignment with

the CCSS, the Smarter Balanced Assessment Consortium (SBAC) authored the Smarter

Balanced Assessment (SBA), which Washington State adopted verbatim in 2010 (SBAC,

2018). The SBA was piloted in Washington State in 2012, fully implemented in 2014,

and became a graduation requirement in 2017 for English Language Arts, and will

become a graduation requirement in 2019 for mathematics (OSPI, 2018; SBAC, 2018).

This current iteration of high stakes testing of Washington students is administered to all

public-school students in grades three through eight, and eleven, and is largely

administered via computer. Students are assessed on the test assigned to his or her grade,

regardless of the grade level of instruction that is provided in his or her classes (OSPI,

2018).

  Among the stated reasons for Washington schools to shift to the SBA as cited by

the Office of Public Instruction (OSPI, 2019) included that 1) this test measures the right

standards, 2) is less expensive, 3) provides quicker results and 4) is more accessible to

learners with disabilities (OSPI, 2018). This progression of standardized testing used as

accountability measures for schools and districts, paired with the 1997 IDEA

reauthorization that required that all students - including those with disabilities - have

access to and are required to take the same assessments as their non-disabled peers, has

resulted in the current assessment protocols for Washington students. Recent legislation

in *Endrew F. v. Douglas County, School District* (2017) provided further support by

adding that a special education program must not be *de minimus*, or lacking significance

or importance. This legislation effectively raises the bar for the requirements of the

program for special education. No longer will minimum requirements be acceptable;

instead, rigorous and robust educational programming will be required. As our current

state measure of school achievement is the SBA, combined with the looming graduation requirements, results of *Endrew F. vs. Douglas County School District* may have the potential to be significant to both the special education programs delivered and the measures of which special education students are evaluated.

Opponents to High stakes Testing

Progressivist theorists of education are in opposition to both the standards based movement and the high stakes testing model currently employed. Jerome Bruner (1960) highlighted the importance of building curriculum where pervading and powerful ideas are at the forefront. Kohn (2000) suggested that even quality assessments cannot cover the depth and breadth of information presented to middle school students, and assessments that pull to the surface pervading and powerful ideas are rarely well done in a multiple choice or short answer format. Bruner (1960) also cautioned that the perceived evolution of American high schools would include an element of competitiveness; one that would require special care not only for students who are quickly moving through material, but "more especially for the student - and he represents an important segment of our younger population - who is not the fast, early, and steady producer" (p. 80). This caution seems to speak directly to students with disabilities, and the care required to measure their skills and abilities even though they may not be meeting standards at the same pace as classmates.

Theorists who oppose high stakes testing claim concern of the validity for all students and the inherent disadvantage for specific subgroups of students. Discussing validity, Kohn (2004) argued that non-instructional factors explain most of the variance among test scores when schools or districts are compared. The variance found in these

tests are largely based in social or economic reasons (Kohn, 2000). Kohn (2004) asserted

that both students of color and those with disabilities are most disadvantaged in this

model of high stakes testing. If students with disabilities are required to participate in

high stakes testing, this disadvantage must be mitigated, frequently done so through the

use of testing accommodations. One example of appropriately mitigating the demands of

high stakes testing provided by Kettler (2015) is the accommodation of extended time. In

the example of extended time, students are able to complete the task without being

measured against a fluency marker. Kettler (2015) stated that this is "especially

appropriate for students with processing speed disorders" (p. 301). Addressing high

stakes testing specifically, Kohn (2004) claimed, "virtually all relevant experts and

organizations condemn the practice of basing important decisions, such as graduation or

promotion, on the results of a single test" (p. 55). A final validity concern raised is that

institutions cannot use the same test as both the lever and instrument to affect change in

the raising of standards, and as the tool used to measure achievement (Kohn, 2000). In

layperson terms, this equates to teaching to the test and results only in short term gains.

As previously stated, the SBA math or alternate high stakes test will be a graduation

requirement in the 2019 school year (OSPI, 2018) and the middle school SBA iterations

can be seen as practice or formative attempts to meet this standard.

     The narrowing of curriculum is an unintended, yet inimical result of high stakes

testing (Berliner, 2011). Currently in Washington State, reading, writing, math, and

science are subjects that are scrutinized at this high stakes level, while civics, the arts, and

technical education are not. Berliner (2011) analyzed the amount of time spent on non-

evaluated subjects and found substantial decrease in the time allocated to the non-

evaluated subjects of social studies, physical education, art and music. Given a finite amount of school time available, the minutes each day assigned to reading and math have grown at the expense of other subjects in elementary schools. Of grave concern is the connection between economic and other social factors to success on high stakes testing, and a widening gap of opportunity for fine arts instruction for students attending schools with a higher poverty rate than schools in affluent areas (Berliner, 2011). In following the spiral curriculum theory (Bruner, 1960), a narrow or limited curriculum in the elementary grades limits the opportunity to learn more deeply different topics through the middle and high school years. Berliner (2011) asserted that "the more narrow the curriculum is in youth, the less likely that the requisite background knowledge will be available in later grades and in the real world" (p. 299). The opposition to high stakes testing claim that the impact of the lessened focus on non-evaluated subjects, lack of validity, especially for marginalized groups, and the narrowing of curriculum, create an educational situation where best learning is not occurring in schools. Policy development, education measurement construction, and a deep understanding of validity are required to evaluate learning. Kohn (2000) and others debate if this is occurring in the current paradigm.

Impact of Language Disability on Math Assessment

Specific Learning Disabilities (SLD) can be diagnosed in a variety of categories. SLD in the category of language include disabilities in reading, writing, or communication, (APA, 2014) and for students with language disabilities this can manifest in struggles with fluency, comprehension, and the ability to clearly communicate thoughts in writing. For students with SLD in reading and/or writing, the

experience of taking a high stakes math test presents unique challenges that are perhaps different from the challenges experienced on other subject area assessments (Costa, Edwards, & Hooper, 2016; Uccelli, Barr, Menses, & Dobbs, 2015). A student with a disability in expressive language could struggle to perform well on a math question that requires him or her to construct an argument or explain his or her thinking. A student with a disability in writing may struggle to respond in writing comparable to non-disabled peers (Beach, Sanchez, Flynn & O'Connor, 2015; Mokhtari & Velten, 2015). Furthermore, reading and writing are closely related language skills (Shananhan, 2006) and nuances of programmatic or specifically designed instruction often flow between these related skills.

A student with a disability in reading may struggle to analyze or understand a question presented in linguistic format. Fluency is an important element in the ability to read, as reading without a certain level of fluency could cause a student to forget or become confused from the first part of the question to the end. Other components to the overall skill of reading, such as vocabulary, add to a student's ability to comprehend questions. Limited vocabulary, especially academic vocabulary, could inhibit a student's ability to successfully read and understand test questions. Fegans and Appelbaum (1986) claimed that narrative skill, a subset of overall language, is a strong predictor of academic success. The narrative skill is needed for deciphering word problems, essentially being able to discern between words that add context and words that illuminate specifics of a math problem. Each of these reading subset skills could be deficient for a student with a reading disability; deficits in several of these subset skills compounds the challenge of reading, comprehending, and being able to demonstrate learned skills in the area of

mathematics. Students who have learning deficits in both reading and writing face challenges in language based assessments, from reading and understanding the question to organizing and writing the answer.

Uccelli et al., (2015) expanded the construct of the language skills needed to achieve in school settings. These skills include the cluster of high-utility language skills that support text comprehension, and illuminated that these sophisticated language skills continue to develop through the upper elementary and middle school years. These skills include the student's ability to recognize and replicate an academic register; language characteristics that indicate a body of writing was written for academic purposes rather than casual use. Uccelli et al. (2015) further explored colloquial strengths in conversational fluency as uniquely different from academic language fluency and the impact this difference has on the students' ability to perform academic tasks. Specifically, Uccelli et al. (2015) distinguished this developmental shift in the degree of "lexical precision for students moving from upper elementary to middle school grades" (p. 341). This shift in lexical precision is imperative as students move to more linguistically complex material in the middle and high school grades, and may be lacking or lagging in students with specific learning disabilities in language. Lamb (2010) suggested that in assessing a student with a learning disability in reading on any subject area including math, the student may be unduly penalized twice, once for low reading performance and once for low mathematics performance resulting from the student's reading disabilities. Factoring in these components of how language interacts or impedes the ability of students with language barriers to achieve in academic tasks, it is questionable if the math SBA is assessing only math constructs, or also language skills.

Linguistic Complexity and Language Demand in Mathematics Assessments

Evaluating linguistic complexity is one method to quantify the impact of language demand on math questions. Although there is not a unified definition in the field of linguistics, Pallotti (2015) differentiated three discrete foci under the overall umbrella of linguistic complexity: structural complexity or the formal properties of texts and linguistic systems, cognitive complexity or the processing costs associated with linguistic structures, and developmental linguistic complexity. Of primary interest to this examination is developmental complexity, or the order in which linguistic structures emerge and are mastered in first and second language acquisition. Students with specific learning disabilities in reading and/or writing lag behind chronological peers in abilities to read, interpret, and communicate their ideas using language. Assessments that measure any construct through the vehicle of language are likely impacted by students' emerging or mastered developmental complexity. Students who have advanced developmental complexity show strengths through assessment processes, while students with lagging skills in developmental complexity do not.

Cawthon, Kaye, Lockhart, and Beretvas (2012) noted that a student with a reading disability might struggle to answer questions on a mathematics problem-solving task. Specifically they called into question the validity of interpreting these results because such results were unclear on if the performance represented the student's reading skill or mathematics problem-solving ability. Cawthon et al. (2012) further called into question the validity and reliability of test questions that were rich in academic language, or language that included vocabulary not frequently used in everyday speech. In mathematics, this is often called the language of math (Redish & Kuo, 2015). For

example, students with specific learning disabilities in reading may understand the English word's meaning, but if fluency is weak when reading a test item, this would effectively slow down the reading process and interrupt the cognitive connections necessary to successfully respond to the test item (Cawthon et al., 2012). For a student with weak fluency, he or she may not be able to remember the key components of the question throughout the entire question, thus limiting the ability to render an accurate response based on reading ability, regardless of content or construct of the question.

The Language of Mathematics

Mathematics and language are inextricably intertwined (Molina, 2012). Given the increased mandated federal testing, and in response to the National Council of Teachers of Mathematics'(NCTM, 2000) defined standards of what students should know and be able to do in mathematics, students are increasingly challenged on standardized assessments to read, create, use, and comprehend numerous mathematical representations as a way of demonstrating mathematical literacy (Matteson, 2006). Mathematicians have been among the first to recognize mathematics as a language (Wakefield, 2000). Using the simplistic definition of language offered by Harley (1995) that language is a system of symbols and rules that enable communication, along with the purpose statement offered by NCTM (2000), mathematics can be thought of as a language that must be meaningful if students are to communicate mathematically and apply mathematics productively. Berlak et al. (1992) noted that mathematics is a plural noun, not a single subject, as it encompasses several related domains. Building on this premise, that the language of mathematics holds the elements of its own language, we can explore learning the language of math as similar to learning a second language.

Wakefield (2000) claimed that the purpose of a second language is to communicate in situations where one's native language is ineffective. The purpose of mathematical language is communication with others; the terms of this language become useful only to the extent that their meanings can be shared (Berlak et al., 1992). Thus, it would suggest that mathematics is both a language created by human beings and used to communicate a specific set of technical information. Boero, Douek, and Ferrari (2008) furthered this discussion and added that the language of mathematics requires a mastery of one's natural language, both words and structures, to incorporate this natural language within the context of mathematical syntax. Molina (2012) expanded this discussion in her claim that students experiencing a difficulty in mathematics may not be grounded in reading English, but rather in understanding the language of math, and claimed that it was irrelevant if the student was fluent in English or not. In looking at the two math problems below, individuals not familiar with the language of math would fail to recognize that the two tasks represent the same problem. To solve the second problem, one would need to know the meaning of Σ, and the instructions laden in the language of math.

1) $N = 1 + 3 + 5 + 7$

2) $\sum_{n=1}^{4} 2n - 1$

Approaching the language of mathematics as a second language, Metsisto (2005) presented that mathematics teachers do not need to become reading specialists to help students read mathematical text, but they do need to recognize that students need help reading in mathematical contexts. An additional struggle in the mastery of the language of mathematics is that mathematics is a language of order (Adams, 2003), and students need fluency in the rules that govern mathematical syntax or order, as well as fluency in

the communicative properties of the symbols and words used to communicate

mathematical reasoning. The order in which operations are written or read is not

necessarily the order in which they are performed (Adams, 2003) and students must use

the order of operations clues to achieve the correct answer. The acronym PEMDAS

(parentheses, exponents, multiplication and division and addition and subtraction) is used

to help students remember the order in which operations are done. If a student were to

solve an equation left to right, as one would in English, he or she would achieve

inaccurate results, not from inaccurate calculations, but from failing to follow the order of

the language.

In DiGisi and Fleming's (2005) work, literacy specialists working with struggling

readers in the math classroom realized that students needed to be taught how to read the

questions and write the answers to demonstrate mathematical understanding. The

complexities of the English language can be in part due to the lexical ambiguities that can

cause misunderstanding and dysfluency in students. Several linguistic and vocabulary

issues in the language of math exacerbate the struggles of students who also struggle to

read and/or write in English. Words with multiple meanings, homophones, words

sounding nearly alike, and technical vocabulary are all opportunities for students with

specific learning disabilities in language to misunderstand the vocabulary and open the

learning process to error (DiGisi & Fleming, 2005).

Metsisto (2005) claimed that "research has shown that mathematics texts contain

more concepts per sentence and paragraph than any other type of text; written in a very

compact style, with little redundancy" (p. 2). This parlance, or vernacular of

mathematics can further be categorized into technical mathematical vocabulary,

procedural vocabulary, that which tells the student what to do, and descriptive vocabulary

that test writers use to provide context for the math problems (DiGisi & Fleming, 2005).

Technical mathematic vocabulary can further be categorized into words that have specific

mathematic meaning but are otherwise uncommon in everyday use, sub-technical

vocabulary that has more than one meaning depending on the content area, and symbolic

vocabulary that includes the special alphabet and non-alphabetic symbols used in

mathematics (Herner & Lee, 2005).

Polysemous terms. Herner and Lee (2005) claimed that mathematics includes

some of the most difficult and unfamiliar vocabulary for students, and without the proper

vocabulary, students face difficulties with the conceptual understanding. Polysemous

terms are terms that have multiple meanings, both within and between subjects. These

terms can add additional vocabulary impediments, and because these words are not

practiced with sufficient reinforcement, mastery is a challenge for some students. Clauses

that carry a specific mathematical meaning, but are also used in everyday speech, can

provide a different kind of misunderstanding. For example, *what's the difference* in

everyday language can mean *who cares*, but in mathematical language it is a procedural

direction to *subtract.* Polysemous terms within the field of mathematics include *degree*,

which can be used to measure temperature, in statistics as a degree of freedom, and in

geometry as a degree of an angle. Examples of mathematical vocabulary that has a

different meaning outside of the math classroom include product, scale, and factor. Table

1 illustrates examples of polysemous terms that have a different meaning in different

contexts.

Table 1

Polysemous Terms

Word	Mathematical context	Everyday context
Volume	Amount of space taken up by an object	Decibel level of sound
Product	Multiply	Item produced to sell
Ruler	Used to measure length	Person in authority
Cubed	Raised to the 3rd power	Cooking term
Face	Flat surface on a solid	Front of a head
Scale	Ratio of size	Used to weigh yourself
Range	Difference between two values	Stove

Building on the concept that mathematics has its own language and borrowing from the teaching techniques in learning a second language, students draw on everyday experiences and language to fill in or support that which is unclear. Given the specific examples of polysemous vocabulary, this technique can be faulty for students trying to use every day experiences to gain mathematical vocabulary context.

Homophones. Adding to this vocabulary struggle are homophones and similar spellings. In class, a student could hear *sum* and *some* and lack the literary sophistication to be able to discern if the teachers mean *to add* or a *collection of*. Homophones and words with similar spellings can be confusing to students with language disabilities. This dysfluency in language compounds the learning of math by slowing down the processing, requiring students to first confirm the accurate meaning of the term, and then address the mathematical question. Similar to polysemous terms, these terms introduce confusion to the learning process. Given the trend in CCSS to practice real life examples (CCSS, 2018) and use descriptive vocabulary to add context, students may struggle to separate terms that are being used in specific mathematical ways and terms that are added for

contextual understanding. Table 2 provides several examples of homophones that might cause confusion for students.

Table 2

Homophones

Mathematic word	Common word
Plane	Plain
One	Won
Whole	Hole
Weight	Wait
Two	Too
Eight	Ate

Assonance. Assonance refers to words that are near homophones, sounding very similar. As stated above, these nuances of language proficiency or deficiency impact students' abilities to fluently work through word problems and stay on pace with the class. The small difference between the words *of* and *off* create important mathematical meaning when used in word problems such as "the percent of something is quite distinct from the percent off something" (Metsisto, 2005, p. 14). Paired assonant words such as *quart* and *court*, *altitude* and *attitude*, *sphere* and *spear*, can introduce vocabulary based misunderstandings, which can lead to inaccurate responses grounded in a language error rather than mathematical reasoning error. These examples provide context for why vocabulary and reading fluency prove to be important elements in the learning of the language of mathematics, and why students with disabilities in reading struggle with deciphering and responding to math questions, apart from math reasoning or calculations presented in the problem.

Matteson (2006) claimed the complex nature of mathematical concepts is further complicated by the compactness of presentation, such as understanding the meaning of a formula. Mathematicians describe a deceptively simple expression as *eloquent*. The eloquent compact presentation of mathematical expressions offers yet another area for students with disabilities in language to struggle (Matteson, 2006). Her example explained that students were often presented a question using one representation categories (verbal, symbolic, graphical, or numerical) and asked to respond in a different representational category. This translation between representational categories, the use of technical and confusing vocabulary, and the compact nature of the language of math combine to create a situation where a student with disabilities in reading and/or writing might face challenges in math achievement that ultimately are not based in math ability. These elements are apparent in the math SBA questions (see Appendices A, B, and C for sample questions for grades six through eight that require students to translate between representational categories, decipher technical vocabulary, and discern between vocabulary that provides context to the question and vocabulary that gives mathematical direction). The development of mathematical literacy involves learning the terms, signs, symbols, and rules for use of language and simultaneously, learning to read and write messages in that language in order to communicate with others (Berlak et al., 1992). These are all skills that students must be able to do to show mathematical proficiency on the math SBA.

Language-based issues in mathematics are problematic for all students (Molina, 2012). Unfortunately for students with reading disabilities, concepts in mathematics that are actually quite simple and can be presented in a way that is far more complex than

necessary to convey meaning (Matteson, 2006). Teachers can support mathematical language learning by using cognates, or words that derive from the same Latin origin. DiGisi and Fleming (2005) recommended that teachers instruct students to recognize types of vocabulary used in math questions, provide students with strategies for reading the questions and identifying what they need to do, and most importantly, provide students with ample opportunities to practice explaining their thinking or showing their work. Teachers can strengthen clarity by avoiding being inexact in their own language, and in disallowing the perpetuation of inexact use of mathematical language. As an example, in a rudimentary lesson on geometry, a teacher could ask the class to select the *bigger half,* this is erroneous and misleading, as by definition, half describes an equally divided whole. An example of perpetuating inexact language offered by Matteson (2006) that some secondary students used the term *graph* to represent the grid itself. In precise mathematical language, the graph as a noun is a diagram showing the relation between variable quantities, each measured along one of a pair of axes at right angles. As a verb, to graph specifically means to plot. Explicit language by the teacher, and accountability to the students to use explicitly correct language, strengthens mastery of the language of mathematics.

Borrowing from strategies used in teaching students a second language, teachers of mathematics must constantly and consistently model both languages (Wakefield 2000). Molina (2012) supported this approach, suggesting that it "may be obvious that language is as critical in mathematics as in any other discipline, the role of language in mathematics entails far more than vocabulary or definitions, encompassing a broad landscape of language based issues" (p. 2). Given these elements and approaching the

learning of the language of mathematics as similar to a second language, it is plausible that students with reading and/or writing disabilities would struggle to learn the language of mathematics to a greater degree than peers without disabilities.

Accommodations in High Stakes Testing

Inextricably linked to high stakes test taking is the use of testing accommodations for students with disabilities. The use of accommodations on standardized tests is not only allowable, it is required. Testing accommodations fall into four general categories: alterations to the presentation/response format, timing/scheduling, setting, and assistive technology (Elliott, Kratochwill, & McKevitt, 2001; Fuchs, Fuchs, & Capizzi, 2005; Jamgochian & Ketterlin-Geller, 2015). Called *supports* on the SBA, universal supports allowed under the Directions for Administration (DFA) for all students (SBAC, 2018), might include the use of an electronic highlighter, computer assisted spell check, English dictionary or English glossary, and the ability for students to manipulate the print. For students with disabilities, designated supports are allowed under the DFA and are provided by the testing administrator as defined in the student's Individual Education Plan (IEP). These designed supports are allowable on all sections of the SBA and include items such as text-to-speech, color contrast, and the ability to request that tools be turned off for a student who might find them distracting (SBAC, 2018). The use of testing accommodations within standardized testing protocols is a direct result of the 1997 reauthorization of the IDEA, which requires that all students with disabilities be offered access to the assessment opportunities offered to non-disabled peers (IDEA, 1997; Thurlow et al., 2005), and allow students to earn valid, but not necessarily optimal scores (Fuchs et al., 2005). Ketterlin-Geller, Yovanoff, and Tindal (2007) clarified the

importance of the use of testing accommodations for students with disabilities in their claim that "accommodations supports students with disabilities by removing construct irrelevant variance caused by physical, cognitive or sensory barriers to accessing the material" (p. 331).

Testing accommodations are considered effective if they do not change the construct of the test, and provide needed differential boost or benefit to students with disabilities when compared to students without disabilities (Kettler, 2015). Elliott et al. (2001) researched the frequency and impact of common accommodations, and specifically focused on math and science high stakes assessments. Researchers reported that extended time, an example of a timing accommodation, was one of the most frequently employed tools, and did not appear to affect the construct being measured. Study results showed that a majority (75.5%) of students with disabilities showed a medium or large positive effect on achievement scores, while approximately half (55%) of the students without disabilities showed a medium or large positive effect on achievement scores when granted the accommodation of extra time. This positive improvement for all students appears to have been considered on the development of the SBA, as all sections of the SBA allow for extended time for all students. Another frequently employed tool is the read aloud or text-to-speech accommodation. This is an example of a presentation or format accommodation. Elliott et al. (2001) reported that in the cases studied, human readers or audiotaped readers provided the strongest effect on scores across grade levels and subject areas for both students with and without disabilities. Dolan, Hall, Banerjee, Chun and Strangman (2005) also studied the text-to-speech accommodation and found that while it had a negligible impact on short reading

passages (2% gain), this accommodation showed a significant impact for students with disabilities on longer reading passages (21.7% gain). Cho, Lee, and Kingston (2012) further studied the effectiveness of testing accommodations for math assessments and found similar results. Their study also suggested that students with disabilities who have high academic ability may be better able to make effective use of accommodations. Accommodations are intended to reduce the impact of personal characteristics that may limit a student's ability to show what he or she knows and can do, and when appropriately implemented, will help teachers and school officials unravel the interaction of group membership (e.g., students with and without disabilities) and the differential benefit of accommodations (Ketterlin-Geller et al., 2007).

Researchers disagree on the validity of the resulting achievement scores when testing accommodations are used on standardized tests. Cook, Eignor, Sawaki, Steinberg, and Cline (2010) investigated the use of testing accommodations in standardized assessments, specifically if the use of testing accommodations evaluated a construct differently when the same construct was assessed under standard conditions. Cook et al. (2010) sought to determine if accommodated assessments ultimately lead to more valid interpretations of student achievement reported on standardized assessments. Results of this study showed that reading and writing constructs were highly correlated, and that the goodness of fit indicators could not be consistently interpreted. Cook, et al. (2010) ultimately found in their study that the assessment held some degree of validity when students with disabilities were accommodated. Crawford and Ketterlin-Geller (2013) argued that when accommodations were appropriately used on a standardized test, the resulting score had higher validity because said accommodation removed barriers that

were connected to the student's disability. This echoes the conclusions in Dolan et al.'s (2005) research that concluded, "any testing solutions that reduce construct irrelevancy will improve the validity of decisions made upon test scores" (p. 25). This research seems in opposition to the findings in Cho et al. (2012) who found no consistent interaction between a student's accommodation status and academic ability. As the research is inconsistent, it leaves practitioners and policy-makers without clear answers as to how best to support students with disabilities in high stakes testing situations.

This question of the use of testing accommodations is an important component to the validity and interpretability of resulting assessment scores. Two unanswered questions remain: if accommodations alter or do not alter the construct being measured and if so, how should educational professionals interpret the resulting assessment scores? To comply with the IDEA for students in special education, accommodations must be offered. In the procedures for the SBA, universal accommodations are available to any student, general or special education, and designated supports are appropriate for students served in special education. It is yet undetermined if these accommodations provide a statistically significant boost to the scores of non-disabled students. The answer to this question is important because it will add to the overall understanding of how to measure constructs with and without accommodations on high stakes mandated assessments.

High Stakes Consequences and Options for Secondary Students

In 2019, the successful passage of the mandated math SBA has high school graduation implications, subject to changes in state law (OSPI, 2018; SBAC, 2018). This practice of withholding graduation if the student was unsuccessful in meeting standard on high stakes testing has been affirmed through litigation (*Brookhart v. Illinois State Board*

of Education, 1983; *Rene v. Reed*, 2001) as appropriate if certain conditions are met (Amrein-Beardsley, 2009; Yell, Katsiyannis, Collins, & Losinski, 2012). These conditions include1 1) the requirement that all students, including those with disabilities, must be required to pass the same exit exam to graduate, 2) all students must be given adequate notice that the test will be required to graduate, 3) the test must be fair, thus assessing that which is taught, and 4) that students with disabilities must be permitted to utilize reasonable accommodations. Given these elements, the SBA meets the definition offered by Haladyna (2006) that high stakes testing is a test where the resulting scores are used to determine consequences such as entrance or denial to advanced classes, graduation or promotion, and other significant opportunities for students. Students with specific learning disabilities may intend to further their education after high school and the expectation of meeting standard on these high stakes tests may be an unfair roadblock to an otherwise capable student.

Some secondary students with diagnosed learning disabilities find it difficult to meet standard on high stakes exams (Thurlow, Albus, & Lazarus, 2015). On the SBA, any student who does not meet standard on the high school exam will be allowed additional attempts (OSPI, 2018); however, students with multiple failed attempts may be obligated to pursue alternates to meeting the graduation requirements of being at standard on high stakes exams. Examples of allowable alternatives to meeting standard include a collection of evidence, taking an off-grade level test, having the cut score lowered, or utilizing a grade point average comparison to prove that the student has mastered the material (OSPI, 2018). Students requiring an alternative to this graduation requirement may aspire to pursue a higher level of education and the lack of a traditional high school

diploma could be a hindrance to admittance to higher education institutions. Prior to

diploma issuance, students with multiple failed attempts may also be denied access to

higher-level coursework. Students may be placed in remedial or support classes,

effectively denying the student opportunities such as advanced classes or additional

electives, courses that are available for non-disabled peers. However, under the typical

six-period day constriction, these classes would not be made available to students served

in special education if the student faced a competing interest of the aforementioned

support class. These advanced classes or additional electives could be seen as the

educational best fit for students with reading and/or writing disabilities. Again this

conflict speaks to access – Are students with disabilities afforded access to the same

educational opportunities as afforded typically developing peers?

Validity and Construct Validity

Validity refers to the credibility of experimental results, the degree to which the

results can be applied to the general population of what is being studied (Kallet, 2004)

and the degree to which each interpretation or use of a test score is supported by the

accumulated evidence (American Educational Research Association [AERA], 2014).

One component to overall validity, internal validity, refers to the credibility of a study

and is determined by the degree to which conclusions drawn from an experiment

correctly describe what transpired during the study, and that no other variables were

actually the cause of the results (Trochim, 2006; Vogt, 2005). An example of a threat to

internal validity is failure to identify alternate reasons for the results of the construct

being studied. If a study examined differences related to gender on achievement and did

not include examination of aptitude or giftedness, the interpretation of the results could

be faulty. External validity, another component to overall validity, refers to whether and to what degree the results from a particular student can be generalized to a larger population (Vogt, 2005). Threats to external validity are well documented (Gall, Gall, & Borg, 2006) and include the categories of history, maturation, testing, instrumentation, regression, selection, and experimental mortality. An example of a threat to external validity is attempting to make a generalization to a population that is dissimilar to the participants that were initially studied. Careful attention to threats to validity increase the strength of the results of an examination.

The Smarter Balanced Technical Report (2016) stated that the SBA "adheres to the Standards as prescribed by AERA" (p. 22) and includes the essential validity elements of careful test construction, adequate score reliability, appropriate test administration and scoring, accurate score scaling, equating and standard setting, attention to fairness, equitable participation, and access. SBA authors claimed that bias is minimized through universal design and accessibility resources (Smarter Balanced Technical Report, 2016), and include the availability and assignment of test accommodations that are available for all students, including students with disabilities. The directions for administration assure that these assessments are administered in a standardized manner sufficient to yield data that support valid inferences (Smarter Balanced Technical Report, 2016). Content validity addressed in the same report lies in the premise that the knowledge, skills, and abilities measured by the Smarter Balanced assessments are consistent with the ones specified in the Common Core State Standards (CCSS, 2018). The Smarter Balanced test authors concluded that validity is an ongoing process, including shared responsibility between the test authors and the test

administrators. Caution is needed especially in the interpretation of standardized test results used in manners not intended by test authors (Smarter Balanced Technical Report, 2016).

Messick (1989) described construct irrelevant variance, or non-construct variance, as one way to explain systematic error. Systematic error is error that does not occur by chance, but rather by an inaccuracy inherent in the measurement tool (Trochim, 2006). Addressing systematic errors in standardized testing is critical for the validity (Drost, 2011) and resulting interpretability of the student scores. Systematic errors can be difficult to detect and difficult to analyze statistically, because all the data varies in the same direction. Elements of construct-irrelevant variance or non-construct variance found in high stakes tests call into question the validity of an assessment designed for typically developing students when taken by students with a specific learning disability. Given the high stakes nature of these accountability measures, and the implications for students on resulting scores, construct validity is of prime importance.

Haladyna and Downing (2004) described construct-irrelevant variance as error that arises from systematic error. Systematic error can be compared with random error, defined as the difference between any observed and true score for each student. However, random error is uncorrelated with true and observed scores. Systematic error is not random, but is group or person specific (Haladyna & Downing, 2004). In the case of high stakes testing, this is a mathematical computation for the variance or difference between the measurement of the stated construct and the achieved score. This variance is a result of a variable unrelated to the construct being measured. Systematic errors are a main concern of validity because systematic errors do not cancel out (Drost, 2011).

Instead, they contribute to the mean score of those being studied, causing the mean value

to be either inflated or deflated. If evidence is found that systematic error exists in a test,

validity and interpretability are compromised.

Construct validity refers to how interpretable the results are, or to the degree that

inferences can be made from a study (Trochim, 2006). In other words, construct validity

can be seen as a labeling issue, does the test claim to measure what is actually measured?

If a test is weak in construct validity, interpretations must be made with caution. In

following the previously offered definition of high stakes testing, the interpretation of the

resulting scores has significant and lasting results. Haladyna (2002) furthered this

discussion of interpretation of test results and made the claim that standardized testing for

groups of disadvantaged students – e.g., students from poverty, with limited English

proficiency, and with disabilities, must be done with precise focus on purposeful

disaggregated reporting. In the current SBA model, student scores are reported as a

continuous point total, with a predetermined cut score; point totals below this cut score

are deemed not at standard and point totals above are deemed proficient or above

standard. This reporting system is identical for all students.

The basis of construct validity and construct validation was described by

Cronbach and Meehl (1955). Building on the premise that construct validity and

validation are not mechanical computations, and are never thoroughly complete, others

have sought to describe methods that continue on this goal. Shepard (1993) added to

Messick's (1989) definition of construct validity by asserting that the first requirement is

a clear definition of the purpose of the test, and after the purpose is agreed upon, then the

appropriate validity measures can be applied. In the cases of high stakes tests such as the

Scholastic Aptitude Test (SAT) and American College Test (ACT), Shepard (1993) outlined how these important tests for college entrance have failed to follow this premise, and this created the opportunity where a student might be erroneously accepted or denied entrance to college based solely on a measure that was not validated for that purpose. Haladyna (2002) described several ways to evaluate validity, suggesting that analyzing test structure validity is a time-consuming task and cautioned that validity analysis is typically directed at specific interpretations and uses, not all interpretations. It is feasible that individuals in school systems using the scores from standardized tests such as the SBA will lack deep knowledge in test item validity, and may interpret the resulting scores in ways not intended. One method suggested by Haladyna (2002) that is specifically targeted to test item construction is to evaluate test item responses and to seek patterns of differences between groups and over time. Given that the SBA is a relatively new test, evaluating differences over time between groups has not yet been studied.

Haertel and Lorie (2004) continued this discussion and added the element for standards based test validation of the *cut score*. The placing or selecting of the cut score for any standards based measurement must be closely aligned to the defined performance standard, or the agreed upon conception of the minimum acceptable level of proficiency on the measured construct. With an agreed upon minimum proficiency marker, the cut score can be established so that passing or meeting standard generally indicates that a student is accurately described by the performance standard, and the performance standard is accurately described by the student (Haertel & Lorie, 2004). OSPI reports that the cut score for meeting standard is in flux; a new cut score will be named as this test moves to tenth grade. OSPI also reports that two different cut scores will denote

students *college ready* as opposed to *acceptable for high school graduation* (OSPI, 2018). Using the computer adapted testing procedures, once a student answers questions, either correctly or incorrectly, subsequent items are presented based on the accuracy of the student's responses and difficulty of the questions (Shapiro, Dennis, & Fu, 2015). Reporting procedures for the SBA include an individual confidence interval, a mathematically computed range for which the true score of the student's achievement on a particular test will lie. A narrow confidence interval would indicate that the reported score is close to the predicted score if the student were to take the test again on a different day (Veldkamp, 2016). Said a different way, a narrow confidence interval would lead educational leaders to make inferences on the reported score that it is more reliable. A wide confidence interval would lead educational leaders to the conclusion that perhaps the student was engaging in a great deal of guess work, and thus, the true score, if the test were taken on another day, might be much higher or lower than the reported score. Understanding how these confidence intervals reveal additional information about students' abilities is imperative when approaching the high stakes consequences for students.

Empirical Studies

      Construct validity in Key Math Revised assessment. Rhodes, Branum-Martin, Morris, Romski, and Sevcik (2015) presented research addressing this question of construct validity in high stakes testing. This research began with the premise that mathematics ability alone does not predict mathematics test performance; linguistic demands may also predict achievement. This research was focused on the Key Math Revised (Connolly, 2007) assessment, which is one test often used for special education

evaluation, or tests used to identify and qualify students for special education. This research sought to quantify if language, mathematics, or a combination of skills in language and mathematics was a statistically significant predictor of math achievement.

This analysis sought to add to the literature characterizing the construct validity of the Key Math Revised (KM-R) for students with less severe language disabilities. A sample of 264 participants from an urban Atlanta school district was selected, and analysis was conducted on demographics, including chronological age, Peabody Picture Vocabulary (Dunn & Dunn, 2007) language age, students' intelligence quotient, current grade, and mother's and father's years of education. Of note, the sample represented 64% male students. This over representation of male students in special education is well documented, even at elementary grades (Coutinho & Oswald, 2005; Piechura-Couture, Heins, & Tichenor, 2013).

This same study used Confirmatory Factor Analysis (CFA) to address the central research question; neither language alone nor mathematics skills alone were found to have a good fit to the model. However, when allowed to covary and to evaluate language and mathematics skills together, a good fit was found to the model. In detailed inspection of results, items toward the beginning of the test protocol were more highly laden with language skills, and successful completion of these items was more highly predicted by language skills. Items more toward the end of the subtests tended to be more highly predicted by mathematics skills. Several items in the KM-R (Connolly, 2007) were entirely predicted by language, while only one item was predicted entirely by mathematics.

Rhodes et al. (2015) concluded that language ability appeared to have a threshold effect; that is, participants with language skills high enough to succeed on the items at the beginning of the test could access the items predicted largely by mathematics abilities toward the end of the test. They concluded that only students with high language skills can access the questions on the KM-R (Connolly, 2007) math test that have high predictability for mathematics skills, thus excluding language-deficient students from exhibiting certain mathematical skills that they might indeed possess. This research adds to the growing body of research on how to evaluate math skills in questions that are laden with language tasks.

This same issue of students not having access to the questions that may be more predictive of math ability is true on the computer-adapted section of the SBA math test. The program is designed to follow a test blueprint and monitors correct and incorrect student answers. The program provides an individual student with different questions dependent on correct and incorrect answers, until conditions are met to ascertain a valid score (SBAC, 2018). Specific to the math test, students are also asked questions in and out of the grade span, in the effort to precisely evaluate the students' knowledge. If the student does not correctly answer the questions in the first two thirds of the assessment, the remaining questions are either not asked, or are asked from a lower grade level set of questions. This is elemental to the question of validity for the SBA math assessment, if the students fail at initial or easier questions that are more heavily laden on language constructs, are they given the opportunity to see and answer questions that are less heavily laden on language constructs and more predictive of higher mathematics skills?

Parallel Research for English Language Learners. Research on how limited language affects assessments is robust when discussing students learning English as a second language. This body of research is of interest to the current study because although the dynamic nature of language acquisition is different for students learning English rather than students with disabilities in language, the effects on assessment seem to be similar. Shaftel, Belton-Kocher, Glasnapp, and Poggio (2006) provided a link between the study of students with disabilities and students learning English while being assessed in mathematics.

Cormier, McGrew, and Ysseldyke (2014) evaluated the linguistic demand of the Woodcock-Johnson Test of Cognitive Abilities for students who were learning English and found that three items significantly rated as high in linguistic demand: verbal comprehension, general information, and concept formation for middle school aged students. This seems to follow a parallel discussion to students with disabilities in language at least in the concept of verbal comprehension. Cormier et al. (2014) concluded that the significant results obtained were also for some "native English speakers, such as children and adolescents who have speech and language difficulties, given that their scores may also be attenuated due to this testing variable" (p. 620). This study provided insight into the analogous relationship between the study of language difficulties for students learning English as a second language and students struggling to master the elements of language as a function of disability.

Abedi and Lord (2001) studied the construct of assessing mathematics achievement for students who struggle communicating in English. In this study, researchers found that the discrepancy between performance on verbal and numeric

format problems strongly suggested that factors other than mathematical skill contributed to student success in the ability to solve word problems. They concluded that dysfluent academic vocabulary impacted students' ability to perform as well on mathematics questions as the same construct evaluated solely in numeric fashion. This research also reported the largest differences in math performance to be found between students in different math classes; students from remedial math classes, while controlling for English Language Learning status, posed the highest improvement score between standard questions and questions on the same math construct with simplified language.

Thomas, Van Garderen, Scheuermann, and Lee (2015) expanded this discussion of evaluating mathematic achievement by clarifying the concept of the language of mathematics as having both expressive and receptive aspects. The receptive skills of speaking and listening, are claimed to be developmental skills, whereas the reading and writing skills are claimed to be learned skills. This research focused on the language demands in the discipline of mathematics, specifically in technical vocabulary and question construction. Specifically, the vocabulary demands are in words that are used differently than in everyday speech, such as *thousand* and *thousandth*, in words that are rarely used in everyday speech such as *coefficient*, or in words that hold an entirely different meaning in a math context such as *prime*. For students with limited language proficiency, using context or construct clues is a tool that is less helpful when approaching a math question. Extraneous or distractor information is more difficult to extract and discard, and information in tables is not necessarily read from left to right. Sentence structure clues also may be less helpful. For example, the main idea may not be at the beginning of the sentence, and the cue words such as *first, next*, and *finally*, may or

may not provide relevant information. Providing math assessments in simplified language provides one opportunity for students struggling with the language of the questions to show ability in mathematical skills (Haag, Heppt, Roppelt, & Stanat, 2015). Again, the principal question of whether assessments are measuring math constructs or being confounded with language issues persists.

 Conclusion

Students who are served in special education for specific learning disabilities in reading, writing, or both reading and writing, are required by law to receive the opportunity to access all components of a public-school setting, including the federally mandated high stakes testing. The current high stakes test for middle school students in 46 states is the SBA. This exam is aligned to the Common Core State Standards (CCSS, 2018). Although parents can exempt their student from this test, in 2019 the successful passage of the SBA math test or alternative will be a graduation requirement (SBAC, 2018). This meets the description of a high stakes test as defined by Haladyna (2006) as having substantial consequences that are highly impactful in large-scale events such as graduation and issuance of a standard diploma.

Middle school SBA exams can be viewed as practice opportunities to this high stakes high school testing experience. Middle school students with specific learning disabilities in language may be secondarily impacted by these disabilities in their ability to achieve at predicted levels on the SBA math test. This question, whether the construct of math ability is accurately measured in this high stakes test, absent of the confounding variables of reading and writing abilities, is yet to be determined, especially in the SBA Claim 3 of Communicating Reasoning which includes *explain your thinking* and

construct an argument (SBAC, 2018). Confounding the issue further is the selection and application of allowable accommodations. The validity of the resulting achievement score must be ascertained before these scores are used to offer or withhold access to higher-level classes, determine promotion, or other weighty consequences for students with disabilities.

While there is much research in the validity of high stakes testing of students with disabilities in the area of disability (Hock, Brasseur-Hock, Hock, & Duvel, 2017; Parkin, 2016; Reed, Cummings, Shaper, & Bincarosa, 2014), research is less robust in the study of non-construct variance and construct validity on a high stakes math test for students with language disabilities. Parallel research is robust in the area of language deficits in English Language Learners on the impact of math achievement and offers similar models to consider. This research aims to add to the growing body of evidence in high stakes testing for students with disabilities and validity questions on the interpretation and use of these results.

Chapter 3: Methods

Introduction

The purpose of this study was to examine if the math Smarter Balanced Assessment is a valid instrument for assessing students with specific learning disabilities in reading and/or writing. Specifically, this study was an effort to contribute to the gap in the literature regarding the high stakes nature of the current federally mandated assessments. Additionally, this investigation attempted to contribute to the establishment of a pathway for future research in valid assessment of students with disabilities, including the use of accommodations and the evaluation of confidence intervals that are reported with individual student SBA scores. This chapter describes the methods and procedures that were used to measure the impact of a specific learning disability in reading and/or writing on the SBA math score. Information on the participants, sampling procedures, and group design are first detailed. Next, data collection and study limitations are described. Finally, ex-post facto research design with a one-way Analysis of Variance (ANOVA) follows, including independent and dependent variables, and analysis of effect size using eta squared (η^2).

Research Questions

The following are the research questions examined:

1. Is there a statistically significant difference on SBA math scores between students with reading disabilities and students who do not have disabilities?

2. Is there a statistically significant difference on SBA math scores between students with writing disabilities and students who do not have disabilities?

3. Is there a statistically significant difference on SBA math scores between students with both reading and writing disabilities and students who do not have disabilities?

Hypotheses of the Study

-   Null hypothesis 1: There is no statistically significant difference on SBA math scores between students with specific learning disabilities in reading and students without disabilities.

-   Null hypothesis 2: There is no statistically significant difference on SBA math scores between students with specific learning disabilities in writing and students without disabilities.

-   Null hypothesis 3: There is no statistically significant difference on SBA math scores between students with specific learning disabilities in both reading and writing and students without disabilities.

Participants and Sampling

The population from which the sample was selected was comprised of middle school students from a large suburban school district in Washington State. Five large comprehensive middle schools ranging in size from 900 to 1,100 students comprise this district. This district reported a total student enrollment in grades K-12 of 20,040 in 2016, with 7.9% receiving free/reduced lunch and 8.9% receiving special education services. District-wide attendance rate was high, with a minimal 0.3% unexcused absent rate for the 2015-16 school year. The ethnicity report as of October 2015 for the entire district is reported in Table 3. The remaining 0.05% percentage were comprised of Native

Hawaiian, Pacific Islander, American Indian and Alaskan Native students; these groups were so small that they cannot be reported in this table without risking privacy violations.

Table 3

Ethnic Distribution of District

Ethnic Group	District Averages
Asian	25.5%
Black	1.7%
Hispanic/Latino	7.8%
Two or more races	6.8%
White	57.3%

Table 4 shows the SBA math district average scores were substantially higher than the Washington State SBA math average scores in 2016.

Table 4

District vs. State Average Math Scores

Grade	District SBA Math Average Scores	State SBA Math Average Scores
6th grade	75.1%	48.0%
7th grade	76.6%	49.8%
8th grade	78.5%	47.8%

Participants for the comparison group were all students in grades six through eight in the 2015-2016 school year who had a specific learning disabilities in reading and/or writing and a current Individual Education Plan (IEP). The comparison group was comprised of 65.1% boys and 34.9% girls. Participants in the control group were randomly selected from a convenience sample of all district students in grades six through eight in the 2015-2016 school year who did not have a disability, and were

assigned to grade level math class. The decision to exclude students in advanced math classes addressed a potential confounding variable of individual math talent and/or additional advanced math instruction. In the 2015-16 school year, 47% of seventh and eighth grade students selected an advanced math class. Participants in the control group were randomly selected in a stratified sample to match the comparison group gender ratio, and were comprised of 76.47% males and 23.53% females. Participants ranged in age from 10 years 8 months to 13 years 10 months. The sample in the study included 64 sixth grade students, 71 seventh grade students, and 74 eighth grade students.

Students were assigned to the three comparison groups according to disability; Group 1 (control) was comprised of students without a disability, Group 2 was comprised of students with only a reading disability, Group 3 was comprised of students with only a writing disability, and Group 4 was comprised of students with comorbid reading and writing disabilities. Students with disabilities in math, communication, or behavior were excluded from this study. This decision was made to eliminate a potential confounding variable of additional disabilities or learning struggles in mathematics. This district does not offer remedial math classes outside of special education classes. In the comparison groups, no students were concurrently receiving support as English Language Learners. While this was unplanned, it also addressed a third confounding variable of English Language acquisition. All students in the studied group claimed English as the language spoken at home and did not receive additional supports for language acquisition while at school. To assure fidelity to the conditions of the comparison and control group, a 10% random sampling of each group was verified by inspection of individual student

cumulative files. No students meeting the conditions of the comparison group were excluded from the study.

Middle school participation in the SBA was high for both this district and Washington State with a 98% participation rate for both the district and state in 2016 (OSPI, 2018). As allowed by the Smarter Balanced Assessment Consortium and the federal government, students could be opted out of this assessment by their parents. During the 2015-16 school year, less than 3% of middle school students in the district opted out of the SBA as requested by their parents. This high participation rate suggests that the sampling pool was representative of the district.

Instrumentation and Data Collection

SBA scores were extracted using the Query function of Skyward, the data management system for this district. For analysis, variables of age, race, language, disability, English Language Learner (ELL), gender, grade, linear score on the SBA mathematics, ordinal score on the SBA mathematics, confidence interval for the SBA math score, and the current math course were available. Social Economic Status (SES) was not available for this study. The district superintendent provided consent for access to this data, and students were reported with identifying numbers rather than student names to assure confidentiality of data. As this was an ex-post facto study, an institutional review board was not deemed necessary.

All students participated in the computer assessment within a four-week period in late spring of 2016. Students who took the test in paper/pencil form as determined appropriate by the individual student's IEP, were scored using the same criteria as students who were assessed via a computer. Accommodations were utilized, both

universal and designated supports, as allowed for students on the SBA. Universal supports are those which are available to all students based on student preference, selection, and recommendations by a known adult. These include but are not limited to: digital calculator, breaks, spell check, and highlighter functions on the computer based assessment. Designated supports are provided by an informed adult who has knowledge of the unique needs of the student such as a teacher or parent. These designated supports include but are not limited to: contrast of background and text, text-to-speech, and the ability to turn off universal tools that might provide a distraction to the student. Analysis of the impact of these accommodations on achievement scores is beyond the scope of this study.

SBA scores are reported in two ways: as a linear score, between 2,473 and 2,652, and as ordinal data, with scores of 1 and 2 considered not-proficient and scores of 3 and 4 considered proficient. The cut score is determined by the test authors and is the score of which scores above are deemed proficient. Table 5 illustrates specific minimum and maximum possible scores along with the cut score for grades six, seven, and eight (SBA, 2018).

Table 5

Smarter Balanced Math Score Information

Grade	Minimum poss.	Cut score	Maximum poss.
6th grade	2,473	2,552	2,609
7th grade	2,484	2,567	2,634
8th grade	2,504	2,586	2,652

Under examination for this study are the math scores on the SBA for students with reading and/or writing disabilities. Administration protocols of this assessment were

scrutinized by testing officials and building test coordinators were required to report all testing anomalies. For the 2015-2016 administration of the SBA, the district assessment coordinator report claimed no testing anomalies were reported for any student included in this study.

This test is untimed and each student may continue testing until he or she independently determines that he or she is finished. Students may complete different sections of the test on different days. The Smarter Balanced Assessment Consortium published data about reliability and validity of this test in the Smarter Balanced Technical Report (2016) and the overall reliability and validity statement is presented in Appendix D. Test authors reported that overall estimated reliability coefficients were high and in the acceptable range for a large scale, high stakes test (SBTR, 2016). Although the reliability coefficients are lower for individual claims, test authors reported that this is expected, as the number of items in the individual claim sections is smaller.

Design and Procedures

The research design of this study was ex-post facto and it examined the relationship of reading and/or writing disabilities on SBA math scores. All students in the study were tested within a four-week period in late spring 2016, and precisely followed the directions for administration for the Smarter Balanced Assessment. Prior to administration of the test, all students were provided practice sessions to review test format, and were assigned individual computers with the supports of headphones, universal, or designated accommodations. Testing locations included classrooms, computer labs, or library spaces. If a specific student was warranted to have accommodations as dictated by his or her IEP, these accommodations were pre-loaded

into the system to be available to the student during all sections of the test (both English

Language Arts and Math). Four sections of the Smarter Balanced Assessment are given

in this assessment window: English Language Arts (ELA) performance task and ELA

Computer Adapted Test (CAT) and math performance task and math CAT. After the

completion of all four tests, the testing window was closed. The Smarter Balanced

Assessment Consortium scored the tests and reported scores back to the district within

four months of test closure. These administration procedures exactly followed the

directives in the DFA of the SBA (SBAC, 2018).

Variables Studied

- Independent Variable: disability in reading/writing or no disability in reading/writing.

- Dependent Variable: math score on the SBA

Following convention for educational research, an alpha level of $p < .05$ was used to

reject the null hypothesis (Gall et al., 2006). The overall linear mathematics score was

used for analysis. A univariate ANOVA is used to analyze the relationship between one

or more factors on a dependent variable (Field, 2013). An ANOVA F ratio is generated

and if significant, reveals differences between groups being studied, but does not

specifically indicate where group differences lie (Gall et al., 2006). With a significant

univariate effect in the ANOVA, several procedures may be used to determine where the

significant differences lie. Field (2013) suggested performing a Gabriel post hoc analysis

when group membership is small. The Gabriel post hoc analysis also adjusts the

significance levels in multiple or post hoc comparisons to reduce the chance of Type 1

error (Vogt, 2005).

The ANOVA with four groups produced three significant tests which corresponded to the key effects that are examined in this study: (a) main effects for reading disability; (b) main effects for writing disability; (c) main effects for both reading and writing disability. Field (2013) suggested that calculating the effect size using the square root of the eta squared (η^2) statistic provides an analysis of the effect size or practical application.

Analysis

The researcher used SPSS version 24 general linear model to examine the descriptive and inferential statistics in the analysis of the research questions. Descriptive statistics included means, standard deviations, skewness, and kurtosis for the dependent variable, and were used to ensure parametric procedures would be appropriate (Field, 2013). However, Field (2013) stated that ANOVA with repeated measures only requires approximately normal data because it is robust to violation of normality.

In order to determine the nature of the relationship between the independent variable and dependent variable, inferential statistics were computed. A one-way analysis of variance (ANOVA) was appropriate due to sample size, single dependent variable, and single independent variable (Gall et al., 2006) and was used to examine the main effects of the independent variable groups.

Valid results from an ANOVA require several statistical assumptions be met (Field, 2013). One assumption is the homogeneity of variances, evaluated with the Levene's test for homogeneity of variances (Field, 2013). Significant results on the Levene's test indicate that the mean variances are similar, and thus an ANOVA would not be an appropriate analysis procedure (Field, 2013). Another assumption is that the

dependent variable is measured at the continuous level and is normally distributed. There should be no significant outliers in any group, and the groups should consist of at least two distinct categorical groups.

Effect Size

The effect size can be calculated in several ways, and represents a standard measure of practical significance (Vogt, 2005). Eta squared measures the degree of association between the effect and the dependent variable and represents practical significance of the study (Vogt, 2005). If the value of the measure of association is squared it can be interpreted as the proportion of the variance in the dependent variable that is attributed to each effect. Eta squared and partial eta squared measure are estimates of the degree of association for the sample, and measures the strength of the phenomena. Cohen (1988) suggested guidelines for interpretation as a small ($\eta^2 = .01$), medium ($\eta^2 = .09$), and large ($\eta^2 = .25$) effect size.

Conclusion

Chapter Three described the research questions and related hypotheses for this study, as well as the population, sample, and statistical methods used. Variables evaluated were the presence or absence of a reading and/or writing disability, and math score on the Smarter Balanced Assessment, reported both as a linear score and as an ordinal score. Descriptive information about the sample population, grade, ethnicity, and age was evaluated. With a significant F ratio on the ANOVA, post hoc tests were calculated.

The results of each statistical analysis are presented in Chapter Four. The assumptions and statistical procedures used in this study are presented. Inferential

statistics are presented and summarized in terms of significance for each research

question.

Chapter 4: Results

Introduction

The results of this study will be presented in order according to the research questions presented in Chapter One. To reiterate, this study examined the impact of a specific learning disability in reading and/or writing on achieved scores on the middle school SBA math assessment for students with disabilities. The statistical methods were summarized in Chapter Three. Prior to reporting results from the inferential tests, descriptive statistics and frequencies of the variables are described. Next, the findings from the univariate statistical procedures of the ANOVA and effect size analysis are reported. Finally, a summary of the most salient findings and whether the specific hypotheses were rejected or accepted is detailed.

Population and Sample

As previously stated, the population from which the sample was selected is a large suburban school district with five large comprehensive middle schools of approximately 1,000 students each. Participants were selected for the control group by randomly selecting students with no documented disability from the entire middle school population. The comparison group was selected by choosing all students with reading and/or writing disabilities from the entire district. The number of participants from each individual school is shown in Table 6, illustrating the distribution of participants. These numbers include both the control sample and the comparison sample.

Table 6

Distribution of Students for Participating Middle Schools

	Number of participants
Middle school #1	34
Middle school #2	37
Middle school #3	45
Middle school #4	49
Middle school #5	44

Descriptive data on all 209 participants is presented in Table 7. Participants ranged in age from 10 years, 8 months to 13 years, 10 months. As anticipated, there was an uneven distribution of males and females in the comparison groups, with 76.7% of the studied group being male and 23.3% being female. Although gender was not specifically examined in this study, mirroring the gender ratio was deemed important, so a stratified control sample was selected, resulting in 76.5% males and 23.5% females comprising the control group.

Table 7

Gender, Grade and Special Education Status Statistics for Participants of Entire Study

	Number of males	Number of females
Grade 6	46	18
Grade 7	53	18
Grade 8	61	13
In Special Education	56	17
Not in Special Education	104	32

Ethnic distribution of all participants is reported in Table 8. While ethnicity is not specifically examined in this study, these percentages provide a frame of reference for the reader about the population for which this study was conducted.

Table 8

Ethnic Distribution of All Participants

Ethnic group	*n*	Participant averages
Asian	39	12.8%
Black	5	3.0%
Hispanic / Latino	22	9.1%
Two or more races	7	4.2%
White	136	70.9%

Descriptive Statistics

Prior to computing inferential statistics, descriptive data were generated and analyzed. To avoid potential sources of bias, Field (2013) suggested checking for assumptions of normality. Assumptions for normality in statistical tests ensure that the statistical test is likely to result in appropriate and interpretable results. The data were first scanned for missing scores and outliers. No missing scores were found and no cases were excluded. The absence of outliers, or scores that are very different from the group, suggested that this assumption was met. The assumption that the data be normally distributed was evaluated first with visual inspection of the histograms (see Appendix E), followed by evaluation of skewness and kurtosis indices. Field (2013) stated that at a significance or *p* value of < .05, the absolute value of 1.96 or less falls into the range of normally distributed data (p. 184). The data were initially checked for normal distribution on the entire sample of 209 students. Means, skewness, and kurtosis indices for the entire data set are reported in Table 9. For the variable of Smarter Balance math score, skewness and kurtosis were well within acceptable limits (Skewness = .17, Kurtosis = - .18), meeting the assumption for normally distributed data. For the variable of confidence interval, the skewness score was within acceptable limits and the kurtosis

score was above recommended limits (Skewness = 1.67, Kurtosis = 3.9). Inspection of

the frequency charts indicated that the frequency of the confidence interval of 22 was

more prevalent than all other scores. It is unclear why middle school students would more

frequently achieve this score. For the entire data set, the assumption for normally

distributed data for the variable of math scores was met.

Table 9

Descriptive Statistics for Studied Variables for All Participants

	n	Mean	*SD*	Skewness	Kurtosis
SBA math score	209	2610.67	101.8	.17	-.18
Confidence interval	209	46	4.78	1.67	3.9

Next, variables were examined separately to check for the assumption of

normality for the students with specific learning disabilities and students without

disabilities. Results of this analysis showed that the variable of Smarter Balanced math

score in split groups remained within the skewness and kurtosis limits (Skewness = - .34,

Kurtosis = - .79). As the variable of confidence interval for the control group remained

unchanged and was within acceptable limits for normally distributed data, both skewness

and kurtosis for the studied group was higher (Skewness = 1.37, Kurtosis = 1.7).

Descriptive statistics for the comparison group are displayed in Table 10. However, both

variables remained within acceptable limits. Therefore, the assumption of normally

distributed data was met specifically for the comparison group of students with

disabilities.

Table 10

Descriptive Statistics for Studied Variables for Control and Comparison Groups

	n	Mean	Standard Deviation	Skewness	Kurtosis
Control group					
SBA math score	136	2645.44	95.19	.22	- .79
Confidence interval	136	22.87	3.46	.78	.11
Comparison group					
SBA math score	73	2545.82	79.99	- .33	.22
Confidence interval	73	25.89	6.12	1.37	1.7

Finally, Smarter Balanced Assessment math scores were examined for normality in the sample subgroups as directed by the three research questions. The control group is presented again for comparison purposes. For all studied groups – i.e., reading disability, writing disability, and both reading and writing disabilities – skewness and kurtosis scores were within acceptable limits on the Smarter Balanced math score variable and are presented in Table 11. For the variable of confidence interval, skewness and kurtosis scores for the studied groups of writing disability and both reading and writing disability were within acceptable limits (see Table 11). For the reading disability subgroup on the variable of confidence interval, both skewness and kurtosis exceeded recommended limits (Skewness = 2.05, Kurtosis = 5.92). The small member size of the reading disabilities only group is one reason why these scores might be higher. As ANOVA is generally robust to minor violations of normally distributed data (Field, 2013), the assumption of normally distributed data was met.

Table 11

Descriptive Statistics for Disability Subgroups

	n	Mean	Standard Deviation	Skewness	Kurtosis
Control group					
SBA math score	136	2645.48	98.19	.22	-.78
Confidence interval	136	22.87	3.46	.77	.1
Reading disability					
SBA math score	14	2558.92	68.96	-.616	1.68
Confidence interval	14	25.92	6.79	2.05	5.92
Writing disability					
SBA math score	22	2557.31	80.56	1.17	.164
Confidence interval	22	25	5.26	1.17	.16
Reading and writing disability					
SBA math score	37	2534.02	83.68	-.542	.039
Confidence interval	37	26.4	6.43	1.19	.917

The final assumption evaluated was homogeneity of variances. Researchers evaluate homogeneity of variances to ascertain if the variance across groups is equal. This assumption for an ANOVA is often tested with a Levene's test. A non-significant result on the Levene's test indicates that the variances between groups are similar, and thus it is appropriate to perform additional inferential statistical procedures. The Levene's result for this study was $F(3, 205) = 2.07$, $p = .105$. This non-significant result indicates that the variances between groups were similar and thus the assumption was met. Said a different way, the variances for the studied groups were different enough to warrant inferential statistics in the form of an ANOVA. Field (2013) stated "if the Levene's test is non-significant ($p > .05$) then the variances are roughly equal and the assumption is tenable" (p. 193).

Given the assumptions of normally distributed data overall and within groups, absence of outliers, and homogeneity of variances, further inferential analysis was appropriate.

Inferential Statistics

The main focus of this research was to evaluate the impact of a reading and/or writing disability on SBA math scores for middle school students. The overall model of evaluating all of the students with disabilities as compared to the control group of students without disabilities resulted in an omnibus $F(3,205) = 19.68$, $p < .001$. This omnibus score determined that an overall significant difference was found and is presented in Table 12.

Table 12

ANOVA Results

	Sum of squares	*df*	Mean sum of squares	*F*	Sig.
SBA Math Score					
Between groups	48221.61	3	160740.53	19.68	.000
Within groups	1673646.61	205	8164.13		
Total	2155868.22				

Further investigation was needed to determine where, or between which groups, the significant differences were found. Field (2013) suggested computing the Gabriel post hoc analysis when group membership is small. The post hoc comparison results are displayed in Table 13. These results indicated significant findings between the control group and all three comparison groups.

Table 13

Gabriel Post hoc Comparison Results

Group	Learning Disability	Mean Difference	Standard Error	Sig.
Control	Reading	105.97	24.07	.000
	Writing	107.58	19.91	.000
	Reading & writing	130.87	16.33	.000
Reading	Control	-86.54	25.36	.001
	Writing	1.61	30.89	1.0
	Reading & writing	24.90	28.35	.935
Writing	Control	-88.15	20.76	.000
	Reading	-1.61	30.89	1.0
	Reading & writing	23.29	24.32	.912
R & W	Control	-111.45	16.75	.000
	Reading	-24.9	28.35	.935
	Writing	-23.29	24.32589	.912

Note. The mean difference is significant at the 0.05 level.

After interpreting the between subjects results, practical significance or effect sizes were analyzed. Field (2013) suggested the eta squared (η^2) calculation is appropriate to determine effect sizes when using an ANOVA. The effect size for this study is $\eta^2 = .224$. Interpreting this result, 22.4% of the variance in the model is explained by group membership, or level of disability. Statistical significance testing assesses the reliability of the association between the independent and dependent variables (Tabachnick & Fidell, 2013), in this study, the examination of the impact of a reading and/or writing disability on SBA math achievement. Cohen (1988) suggested guidelines for interpretation as a small ($\eta^2 = .01$), medium ($\eta^2 = .09$) and large ($\eta^2 = .25$) effect size. Given these statistics, results in this study reflect a medium to large effect or practical significance.

Research Question 1

Research Question One evaluated the impact of a specific learning disability in reading on math SBA scores. As previously reported, the ANOVA showed a statistically

significant difference with a medium to large effect size ($F(3, 205) = 19.69$, $p < .001$, $\eta^2 =$ .224). The Gabriel comparison showed a statistically significant difference between the control group and the group of students with a reading disability, with the mean difference calculated at $M = 105.97$, $p < .001$. There were no statistically significant differences between the reading group and either the writing group or the reading and writing group. With these results, the researcher rejected the null hypothesis for Question One.

Research Question 2

Research Question Two evaluated the impact of a specific learning disability in writing on math SBA scores. As previously reported, the ANOVA showed a statistically significant difference with a medium to large effect size ($F(3, 205) = 19.69$, $p < .001$, $\eta^2 =$ .224). The Gabriel comparison showed a statistically significant difference between the control group and the group of students with a writing disability, with the mean difference calculated at $M = 107.58$, $p < .001$. There were no statistically significant differences between the writing group and either the reading group or the reading and writing group. With these results, the researcher rejected the null hypothesis for Question Two.

Research Question 3

Research Question Three evaluated the impact of a specific learning disability in reading and writing on math SBA scores. As previously reported, the ANOVA showed a statistically significant difference with a medium to large effect size ($F(3, 205) = 19.69$, $p < .001$, $\eta^2 = .224$). The Gabriel comparison showed a statistically significant difference between the control group and the group of students with both reading and writing

disabilities, with the mean difference calculated at $M = 130.87$, $p < .001$. There were no statistically significant differences between the reading and writing group and either the reading group or the writing group. With these results, the researcher rejected the null hypothesis for Question Three.

Summary of Findings

An ANOVA was computed to test the impact of a reading and/or writing disability on student math scores on the middle school SBA. Significant findings for all three research questions were found. The Gabriel post hoc adjustments to control for Type 1 error were all significant. Inspection of the means plot (see Appendix E) showed that the impact of a reading or a writing disability impacts the outcome of math SBA scores in a similar manner, while the impact of both a reading and a writing disability impact the math SBA scores to a higher degree. The results of the effect size calculation showed a medium or large effect, indicating that these results have non-trivial implications for students, teachers, and educational policy-makers.

While not a foci of this study, the confidence interval information adds clarity to the practical importance of these findings. Reported for individual students, the confidence interval indicates the range in which the true score will lie. For the students in this study, when confidence intervals were added to an individual student's score, 10 more students, or 14%, would have achieved the proficient score (see Appendix F). In conclusion, these results support a rejection of the null hypothesis for all three research questions.

Conclusion

This chapter presented the findings of the research. Prior to performing the statistical procedures, data were analyzed to check for violations of parametric assumptions. Descriptive statistics were computed for all variables in aggregated and disaggregated groups, and reported. Parametric statistical data related to the research question were reported. The results showed in the main effect of the studied group, at an alpha level of $p < .05$, there was a statistically significant difference between the studied and control groups on the SBA math score. Consequently, the researcher rejected all three null hypotheses. Effect size and practical implications were calculated and reported, finding a medium to large effect size.

The following chapter provides a summary of the purpose of this study, the methodology employed, limitations, and the practical significance of the results. Finally, suggestions for future research are offered.

Chapter 5: Summary and Discussion

Introduction

The purpose of this study was to investigate the impact of reading and writing disabilities on achieved scores on the math Smarter Balanced Assessment. A second purpose was to explore if the use of testing accommodations as specified by the SBA test authors mitigated the effects of a reading and/or writing disability on this particular high stakes math assessment. A third purpose was to analyze individual students' confidence intervals to ascertain if the addition of the calculated confidence interval provided a differential boost enough to score the students in the proficient range. This chapter provides a discussion of the results found in this study, including results from inferential statistical analyses to investigate the stated problem. Finally, a description of research and educational implications, study limitations, and suggestions for future research are presented.

Discussion

Research questions. The evaluation of the overall model used to compare the mean scores of all students with disabilities to the mean scores of the control group of non-disabled students resulted in statistically significant findings, with a medium to large effect size (F(3, 205) = 19.69, p < .001, η^2 = .224). The Gabriel post hoc comparison (*Mean Difference* = 105.97, *Standard Error* = 24.07, p < .001) showed a statistically significant difference between the control group and the group of students with only a reading disability. This was the smallest group (n = 14, M = 2558.92, SD = 68.96). Several reasons can be offered for this small group membership. Research has shown that reading and writing are highly correlated and share a reciprocal relationship in

literacy development (Costa et al., 2015), thus when parsing to single disability groups, these numbers are likely to be relatively small. This small group membership is also evident in the group for students with only a disability in writing ($n = 22$, $M = 2534.02$, $SD = 83.68$). For the writing disability only group, the Gabriel post hoc comparison showed similar statistically significant differences as compared to the control group (*Mean Difference* = 107.58, *Standard Error* = 19.91, $p < .001$). Statistically significant results support that an evident disability in reading or writing does impact the ability to perform at the same level as peers without disabilities on this math assessment.

The largest difference was found between the control group and the group of students who had both reading and writing disabilities ($n = 37$, $M = 2558.92$, $SD = 68.96$). In this group, the Gabriel post hoc comparison (*Mean Difference* = 130.87, *Standard Error* = 16.33, $p < .001$) showed a statistically significant difference. The larger mean difference indicates that students with both reading and writing disabilities are more impacted than students who have only either reading or writing disabilities on this math assessment. This double impact is supported in the literature when evaluating other comorbid disabilities (Costa et al., 2015; Shananhan, 2006; Swanson, Jerman, & Zheng, 2009). With this data, the researcher rejected the null hypotheses for all three research questions.

Concerns arose in the selection of the control sample. From the population from which this sample was drawn, students may self-select into an advanced math path, effectively narrowing the control sample population by 47% for this study. Given that nearly half of the students in this district elect for an advanced math path, researchers can interpret that the mean ability level of the grade level assigned class is lower than the

mean level of the advanced students, due in part to advanced instruction and possibly to natural math talent or interest.

The number of students with math disabilities presented an unanticipated problem during data collection. The sample of students with disabilities initially included all students with disabilities in grades six through eight. Once students with behavior and communication disabilities were extracted, 374 students remained in the sample set. Of these students, 301 students had a math disability, effectively the largest disability category in the group of students with disabilities. Although not the focus of this study, it was noted that 48% of the students with math disabilities, 146 students, also had both reading and writing disabilities. This ultimately resulted in smaller than anticipated groups for the study. This research illuminated differences when disability groups were parsed out more specifically.

Generalizations made to other populations, districts dissimilar to the studied district, and students of different ages must be made with caution. It is understood that there may exist other factors that influence the overall math achieved score such as maturity, test anxiety and fatigue, and effort. These are all beyond the scope of this study. Although the fundamental validity question for the high stakes nature of this test for high school students is an underlying premise of this examination, these results can only be generalized to middle school students.

The parallel research for the assessment of ELL students provided additional insight. As noted in the literature review, language difficulties are more dynamic for ELL students than for students with specific language disabilities. The research on validity in assessment for ELL students is more robust than for students with disabilities,

and supports the suggestion that language does impact the ability to perform on math tests for ELL students (Cormier et al., 2014; Haag et al., 2015; Shaftel et al., 2006). Although research examining this same construct is less available for students with disabilities, continued exploration similar to Rhodes et al.'s (2015) study on the KeyMath – Revised (Connolly, 2007) and other large scale assessments such as the Kansas General Mathematic Assessment (Shaftel et al., 2006) is necessary to more fully understand how language deficits impact the ability to perform mathematics tasks. As shown in Rhodes et al. (2015) and Veldkamp's (2016) research, initial studies reveal that the beginning or early questions on large scale assessments present simpler mathematics concepts and more language laden question forms. Following computer adapted testing procedures, subsequent items are adjusted based on the accuracy of the student's response and the difficulty of the question (Shapiro et al., 2015). Because Rhodes et al.'s (2015) research indicated that initial, or usually easier, math problems are more laden with language, and the later, or supposedly more difficult math problems are less laden with language, this can be offered as one reason why language disabilities impact achieved math scores on computer adapted assessments. The Rhodes et al. (2015) study, in conjunction with the abundant research that explores the impact of language acquisition and mastery for ELL students (Boero et al., 2008; Shaftel et al., 2006), supports the premise language plays a role in the ability to answer math questions.

Testing accommodations. Accommodations are not only allowed on the SBA, they are mandatory. Following the suggestions of universal design, some accommodations are available for all students, such as extended time, allowances for breaks, and use of an embedded calculator. Accommodations selected by IEP teams for

students with disabilities are intended to mitigate the impact a disability has on the student's ability to perform to his or her capacity.

Significant research has been conducted on the selection process (Cawthon et al., 2012; Cook et al. 2010) and effectiveness (Elliott et al., 2001) of specific accommodations for high stakes tests. Researchers do not yet agree on either of these topics. The selection and use of accommodations is a crucial component of the validity and interpretability of resulting scores. The selection model or framework offered by Fuchs et al. (2005) provided one example of a structured system for the purposeful selection of accommodations. Evaluating the differential boost to students with and without disabilities under accommodated and not accommodated assessments will provide imperative insight to the validity measures of high stakes tests (Elliott et al., 2001). Anticipated in 2019, meeting standard on the SBA math test will become a graduation requirement. In order to validly assess the math skills of students with reading and/or writing disabilities, this question of validity must be answered. If providing accommodations does mitigate the impact of a specific language disability, then the resulting score will be valid for interpretation. If, as Cho et al. (2012) claimed, the use of accommodations diminishes the validity of the resulting scores because the accommodations provide a differential boost, and thus no longer fit the norm referencing standard, then policy-makers at the state and district level must be ready to respond to the needs of students with learning disabilities who may intend to pursue post-secondary education. Withholding a diploma, and the life implications therein, for failing to meet standard on this assessment may be statistically unsupported under the current paradigm.

Impact of confidence intervals. Confidence interval data offer insight to the range where the true achieved score might lie, and is a more accurate representation of the true score than the reported numerical score. The Smarter Balanced Assessment Consortium provides this explanation and refers to the confidence interval as an error band:

> Smarter Balanced tests provide the most precise scores possible within a reasonable time limit, but no test can be 100 percent accurate. The error band indicates the range of scores that a student would likely achieve if they were to take the test multiple times. (Smarter Balanced Reporting System User Guide, 2016, p. 120)

This often overlooked data point may prove to be informative in the future as policy-makers become more aware of its impact. In this study, when the confidence interval was added to the student's score, 10 more students, or 14% would have reached the cut score to be determined proficient. Additionally, with the confidence interval added to their score, several more students were within a minimal point number needed to reach the cut score (see Appendix F). Further research in the use or application of confidence intervals to increase interpretability of high stakes tests is needed.

Research Significance

Findings from this study have theoretical significance for understanding how students with specific learning disabilities are assessed in high stakes testing. It is important to recognize that these scores are, and will continue to be, a snapshot in time, rather than a representation of cumulative skill development. The current research added insights to the assessments for students with certain types of disabilities, and echoed the

results of previous studies (Cormier et al., 2014). Students with reading and/or writing disabilities learn differently, and assessing via high stakes tests authored for accountability purposes may not adequately reveal what students know and the level that this learning has been mastered. Cormier et al. (2014) articulated this as an investigation between the question of *difference* versus *disorder* (p. 610) and specifically investigated if differences in large scale assessments measure linguistic demand the same across student groups. Results of the current study revealed that it is plausible that students with reading and/or writing disabilities may have abilities in mathematics that are impacted by elements of the language of mathematics, the linguistic demand of question formats, and the language laden questions that are often found in the easier questions on math assessments. These findings echo previous research on the assessment of students with disabilities (Rhodes et al., 2015). This study also broadened understanding about the importance of parsing out individual disability categories rather than studying students with disabilities as a single group, and indicates the impact may be greater for students who have more than one specific learning disability.

Educational Significance

In order to use high stakes assessments for high stakes consequences, such as graduation, the assessment tool must be as valid and free from bias and construct irrelevant variance as possible. As the Smarter Balanced Assessment is a relatively new test, and considering that high school participation rates have not yet reached a level that would invite statistical analysis, it is difficult to extrapolate how high school students with reading and/or writing disabilities will perform on the mathematics assessment. In the 2015-2016 school year, less than 5% of high school students in the studied district

participated in the math SBA and thus, initial scores were repressed. When sufficient participation is reached to perform statistical analyses on results of the high school Smarter Balanced Assessments, replicating this study with high school participants will be helpful in continued efforts to validly evaluate students with reading and/or writing disabilities.

Prior to empirical research being completed, policy-makers and school personnel should use caution when assessing graduation consequences using this tool. Findings suggest that allowing the confidence interval calculated for each student to be one method of attempting to provide fairness in the absence of justified validity for students with disabilities.

Limitations

As with any study, several limitations are found within the present study. The small sample size of the individual groups, especially the group of reading only disability must be considered. In order to increase sample size, expanding the population from which the sample was drawn to include more districts concurrently, might open a confounding variable of differences in policies and procedures in the special education evaluation process, resulting in questionable group membership characteristics.

The Smarter Balanced Assessment is a relatively new assessment, only used nationwide since 2015 (SBAC, 2018). As with any new large scale assessment, adjustments are made as evidence is gained through experience (Smarter Balanced Technical Report, 2016). Students and teachers will become more familiar with this test, both in format and content. As students and teachers become more familiar with the

content and districts align curriculum to the Common Core State Standards, we can expect influences of inexperience or poorly aligned curriculum to be abated.

Although the schools in this district generally represent typical schools in the surrounding area, results may be limited to schools with similar demographics. It is important to note that this test is a moment-in-time. Elements of student anxiety, fatigue, effort, and sustained focus are all elements that influence test taking and achieved results. These were all beyond the scope of this study.

Future Research

The impact of testing accommodations is a robustly researched topic (Dolan, et al., 2005; Elliott et al., 2001; Kettler, 2015). Researchers do not yet agree if the use of testing accommodations improves or decreases the interpretability of resulting scores. Additionally, research has revealed the presence of a differential boost, or improved scores for students without disabilities while using accommodations (Dolan et al., 2005; Fuchs et al., 2005; Ketterlin-Geller et al., 2007). While the limits of allowable designated supports are clear within the DFA for the SBA, the opposite pendulum of opportunity is yet unknown. Additional research on the Smarter Balanced Assessment results would be probative to examine if providing similar accommodations to students without disabilities, but who struggle with academic tasks, would show higher achieved scores. The process for selecting testing accommodations is not tightly regulated. This allows for IEP teams to select packages of accommodations rather than specific targeted accommodations or accommodations that support a student's emotional rather than specific educational needs. Examining teacher selection processes for the assignment of testing accommodations may suggest best practices not currently widely known or

implemented. Further exploration on the use of the confidence interval score as an accommodation might provide an option for policy-makers in the absence of empirical research on systematic variance for students with reading and/or writing disabilities.

Repeating this study with larger samples or in subsequent years might offer insights not currently available. As students and school personnel become more familiar with this test, how items are measured, and best practices on the assignment of accommodations, greater insight may suggest other avenues to research.

Subsequent studies are needed to continue to explore the unique experiences of high stakes testing for students with reading and/or writing disabilities. The research design for these potential studies should include systems to detect unique variance for students with specific learning disabilities categorically rather than studying students with disabilities as one group.

Conclusion

The present study resulted in significant results for the three stated research questions; students with reading and/or writing disabilities do face an impact when participating in the math Smarter Balanced Assessment. These findings have important implications for research and the assessment of students with disabilities. Understanding how the language of math and linguistic demand impact the reading difficulty of a math problem helps classroom teachers seek ways to lessen the language demand while maintaining construct validity and rigor. The use of accommodations on standardized testing is an important element in the valid assessment of students with disabilities. As high stakes tests for accountability purposes are paramount in today's educational landscape, careful attention and further exploration on this topic is warranted.

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education. 14*(3), 219-234. http://dx.doi.org/10.1207/S15324818AME1403_2

Adams, T. (2003). Reading mathematics: More than words can say. *The Reading Teacher, 56*(8), 786-795.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

American Psychiatric Association. (2014). *Diagnostic and statistical manual of mental disorders: DSM-V*. Washington, DC: Author.

Amrein-Beardsley, A. (2009). The unintended, pernicious consequences of "staying the course" on the United States' No Child Left Behind Policy. *International Journal of Education Policy and Leadership, 4*(6), 1-13.

Bateman, B. (1992). Learning disabilities: The changing landscape. *Journal of Learning Disabilities, 25*(1), 29-36.

Bayles, M. (2009). *Perceptions of educators and parents of the California High School Exit Examinations (CAHSEE*) (Unpublished doctoral dissertation). California Lutheran University, Thousand Oaks, CA.

Beach, K., Sanchez, V., Flynn, L., & O'Connor, R. (2015). Teaching academic vocabulary to adolescents with learning disabilities. *Teaching Exceptional Children, 48*(1), 36-44. http://dx.doi.org/10.1177/00400591559483

Berlak, H., Newmann, F., Adams, E., Archbald, D., Burgess, T., Raven, J., & Romberg, T. (1992). *Toward a new science of education testing and assessment*. Albany, NY: State University of New York Press.

Berliner, D. (2011). Rational responses to high-stakes testing, the case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education, 41*(3), 287-302. http://dx.doi.org/10.1080/0305764X.2011.607151

Boero, P., Douek, N., & Ferrari, P. (2008). Developing mastery of natural language: Approaches to some theoretical aspects of mathematics. In L. D. English (Ed.), *Handbook of international research in mathematics education* (2nd ed., pp. 262-295). New York, NY: Routledge.

Brookhart v. Illinois State Board of Education, 697 F.2d 179 (7th Cir. 1983).

Bruner, J. (1960). *The process of education*. Cambridge, MA: Harvard University Press.

Carter, E., Wehby, J., Hughes, C., Johnson, S., Plank, D., Barton-Arwood, S., & Lunsford, L. (2005). Preparing adolescents with high-incidence disabilities for high stakes testing with strategy instruction. *Preventing School Failure, 49*(2), 55-62.

Cawthon, S., Kayne, A., Lockhart, L., & Beretvas, S. (2012). Effects of linguistic complexity and accommodations on estimates of ability for students with learning disabilities. *Journal of School Psychology, 50*(3), 293-316. http://dx.doi.org/10.1016/j.jsp.2012.01.002

Cho, H., Lee, J., & Kingston, N. (2012). Examining the effectiveness of test accommodation using DIF and a mixture IRT model. *Applied Measurement in Education, 25*(4), 281-304. http://dx.doi.org10.1080/08957347.2012.714682

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey, NJ. Lawrence Erlbaum.

Common Core State Standards. (1/19/2018). Retrieved from: www.k12.wa.us/curriculuminstruct/ccss/ default.aspx

Connolly, A. (2007). *Key Math Revised.* New York, NY: Pearson.

Cook, L., Eignor, D., Sawaki, Y., Steinberg, J., & Cline, F. (2010). Using factor analysis to investigate accommodations used by students with disabilities on an English-language arts assessment. *Applied Measurement in Education, 23*(2), 187-208. http://dx.doi.org/10/1080/08957341003673831

Cormier, D., McGrew, K., & Ysseldyke, J. (2014). The influences of linguistic demand and cultural loading on cognitive test scores. *Journal of Psychoeducational Assessment, 32*(7), 610-623. http://dx.doi.org/10.1177/0734282914536012

Costa, L., Edwards, C., & Hooper, S. (2015). Writing disabilities and reading disabilities in elementary school students: rates of co-occurrence and cognitive burden. *Learning Disability Quarterly, 39*(1), 17-30. http://dx.doi.org/10.1177/0731948714565461

Coutinho, M., & Oswald, D. (2005). State variation in gender disproportionality in special education: Findings and recommendations. *Remedial and Special Education, 26*(1), 7-15. http://dx.doi.org/10.1177/07419325050260010201

Crawford, L., & Ketterlin-Geller, L. (2013). Middle school teachers' assignment of test accommodations. *The Teacher Educator, 48*(1), 29-45. http:dx.doi.org/10.1080/08878730.2012.740152

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests.

Psychological Bulletin, 52(4), 281-302.

Cumming, J. (2008). Legal and educational perspectives of equity in assessment. *Assessment in Education: Principles, Policy & Practice, 15*(2), 123-135.

DiGisi, L., & Fleming, D. (2005). Literacy specialist in the math class! Closing the achievement gap on state math assessments. *Voices from the Middle, 13*(1), 48-52.

Dolan, R., Hall, T., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning, and Assessment, 3*(7), 5-32.

Drost, E. (2011). Validity and reliability in social science research. *Education Research and Perspectives, 38*(1), 105-123.

Dunn, L., & Dunn, D. (2007). *Peabody picture vocabulary test*. New York, NY: Pearson.

Elliott, S., Kratochwill, T., & McKevitt, B. (2001). Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities. *Journal of School Psychology, 39*(1), 3-24. http://dx.doi.org/10.1016/S0022-4405(00)0005-X

Elementary and Secondary Education Act of 1965. Public Law 89-10. 20 U.S.C. § 70 (4/11/1965).

Elementary and Secondary Education Act of 1986. Public Law 89-10. (1991).

Every Student Succeeds Act of 2015. Public Law 114-98 20 U.S.C. § 70 (12/10/2015).

Endrew F. vs. Douglas County School District. RE-1 580 US (2017).

Fegans, L., & Appelbaum, M. (1986). Validation of language subtypes in learning

disabled children. *Journal of Educational Psychology, 78*(5), 358-364.

http://dx.doi/10.1037/0022-0663.78.5.358

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). Los Angeles,

CA: Sage Publications.

Fuchs, L., Fuchs, D., & Capizzi, A. (2005). Identifying appropriate test accommodations

for students with learning disabilities. *Focus on Exceptional Children, 37*(6), 1-9.

Gall, M., Gall, J., & Borg, W. (2006). *Educational research*. Boston, MA: Pearson.

Gallagher, J., & Reid, K. (1981). *The learning theory of Piaget & Inhelder*. Monterey,

CA: Brooks/Cole Publishing Company.

Gazda, G., & Corsini, R. (1980). *Theories of learning: A comparative approach*. Itasca,

IL: F.E. Peacock Publishers, Inc.

Gunzenhauser, M., & Hyde, A. (2007). What is the value of public school accountability?

Educational Theory, 57(4), 489-507.

http://dx.doi.org/10.1111/j.1741-5446.2007.00270

Haag, N., Heppt, B., Roppelt, A., & Stanat, P. (2015). Linguistic simplification of

mathematics items: Effects for language minority students in Germany. *European

Journal of Psychological Education, 30*, 145-167.

http://dx.doi.org/10.1007/s10212-124-023306

Haertel, E., & Lorie, W. (2004). Validating standards-based test score interpretations.

Measurement: Interdisciplinary Research and Perspectives, 2(2), 61-103.

Haladyna, T. (2002). *Standardized achievement testing: Validity and accountability*.

Boston, MA: Allyn & Bacon.

Haladyna, T. (2006). Perils of standardized achievement testing. *Educational Horizons,*

85(1), 30-43.

Haladyna, T., & Downing, S. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17-25. http://dx.doi/10.1111/j.1745-3992.2004.tb00149.x

Hallahan, D., & Mercer, C. (2001, August). *Learning disabilities: Historical perspectives*. Paper presented at the Learning Disabilities Summit, Washington, DC. Retrieved from: https://files.eric.ed.gov/fulltext/ed458756.pdf

Harley, T. (1995). *The psychology of language: From data to theory*. East Sussex, UK: Taylor & Francis.

Herner, L., & Lee, H. (2005). Standards, special education and access to mathematics curriculum. *Teaching Exceptional Children Plus, 1*(6).

Hock, M., Brasseur-Hock, I., Hock, A., & Duvel, M. (2017). The effects of a comprehensive reading program on reading outcomes for middle school students with disabilities. *Journal of Learning Disabilities 50*(2), 195-212. http://dx.doi/10.1177/0022219415618495

Individuals with Disabilities Education Act, 20 U.S.C. §1400 (1997).

Individuals with Disabilities Education Act, 20 U.S.C. §1400 (2004).

Inhelder, B., Sinclair, H., & Bovet, M. (1974). *Learning and the development of cognition*. London, UK: Routledge & Kegan Paul, Ltd.

Jamgochian, E., & Ketterlin-Geller, L. (2015). The 2% transition, supporting access to state assessments for students with disabilities. *Teaching Exceptional Children*, *48*(1), 28-35. http://dx.doi.org/10.1177/0040059915594781

Kallet, R. (2004). How to write the methods section of a research paper. *Respiratory*

*Care, 49(*10).

Kauffman, J., Hallahan, P., & Pullen, C. (2011). *Handbook of special education*. New York, NY: Routledge.

Kettler, R. (2015). Adaptations and access to assessment of common core content. *Review of Research in Education*, *39*(1), 295-330. http://dx.doi.org/10.3102/0091732X1456075

Ketterlin-Geller, L., Yovanoff, P., & Tindal, G. (2007). Developing a new paradigm for conducting research on accommodations in mathematical testing. *Exceptional Children, 73*(3), 331-347.

Kohn, A. (2000). *The case against standardized testing, raising the scores, ruining the schools*. Portsmouth, NH: Heinemann Publishing.

Kohn, A. (2004). *What does it mean to be well educated?* Boston, MA: Beacon Press.

Lamb, J. (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.

Matteson, S. (2006). Mathematical literacy and standardized mathematical assessments. *Reading Psychology, 27*(2), 205-233. http//dx.doi.org/10.1080/02702710600642491

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: The American Council on Education.

Metsisto, D. (2005). Reading in the mathematics classroom. In Kenney, J., Hancewicz, E., Heuer, L., Metsisto, D., & Tuttle, C. (Eds.), *Literacy strategies for improving*

mathematics instruction (pp. 9-23). Alexandria, VA: Association for Supervision and Curriculum Development.

Mokhtari, K., & Velten, J. (2015). Strengthening academic vocabulary with word generation helps sixth grade students improve reading comprehension. *Middle Grades Research Journal, 10*(3), 23-42.

Molina, C. (2012). *The problem with math is English: A language focused approach to helping all students develop a deeper understanding of mathematics*. San Francisco, CA. Josey Bass.

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

National Council on Education Standards and Testing. (1992). *Raising standards for American education.* Washington, DC: Government Printing Office.

No Child Left Behind Act. 20 U.S.C. § 6319. (2001).

Office of Superintendent of Public Instruction. (1/16/18). Retrieved from: http://www.k12.wa.us/

Overton, T. (1992). *Assessment in special education: An applied approach*. New Jersey, NJ: Prentice-Hall, Inc.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research, 31*(1), 117-134. http://dx.doi/org/10.1177/0267658314536435

Parkin, J. (2016). Wechsler Individual Achievement Test-Third Edition Oral Language and reading measures effects on reading. *Journal of Psychoeducational Assessment, 34*(1), 1-16. http://dx.doi/org/10.1177/0734282916677500

Pemberton, J., Rademacher, J., Tyler-Wood, T., & Perez Cereijo , M. (2006). Aligning

assessments with state curriculum. *Intervention in School and Clinic, 41*(5), 283-289.

Piaget, J. (1953). *The origins of intelligence in children.* New York, NY: Basic Books.

Piaget, J. (1970). *Structuralism*. New York, NY: Basic Books.

Piaget, J. (1977). *The development of thought: Equilibration of cognitive structures*. New York, NY: Viking Press.

Piechura-Couture, K., Heins, E., & Tichenor, M. (2013). The boy factor: Can single-gender classes reduce the over-representation of boys in special education? *College Student Journal, 47*(2), 235-243. http://dx.doi.org/10.1007/s11191-015-9749-7

Redish, E., & Kuo, E. (2015). Language of physics, language of math: Disciplinary culture and dynamic epistemology. *Science & Education, 24*(5), 561-590. http://dx.doi.org/10.1007/s11101-015-9749-7

Reed, D., Cummings, K., Schaper, A., & Biancarosa, G. (2014). Assessment fidelity in reading intervention research: A synthesis of the literature. *Review of Educational Research, 84*(2), 275-321. http://dx.doi.org/10.3102/0034654314522131

Rehabilitation Act, 29 U.S.C. § 701 et seq. (1973).

Rene v. Reed, 751 N.E. 736 (Ind. Ct. App. 2001).

Rhodes, K., Branum-Martin, L., Morris, R., Romski, M., & Sevcik, R. (2015). Testing math or testing language? The construct validity of the KeyMath-Revised for children with intellectual disability and language difficulties. *American Journal on Intellectual and Developmental Disabilities, 120(*6), 542-568. http:/dx.doi.org/10.1352/1944-7558-120.6.542

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105-126.

Shananhan, T. (2006). Relations among oral language, reading and writing development. In MacAuthur, C., Graham, S., & Fitzgerald, J. (Eds.) *Handbook of writing research*. New York, NY: Guilford Press.

Shapiro, E., Dennis, M., & Fu, Q. (2015). Computer adaptive and curriculum-based measures of math in progress monitoring. *School Psychology Quarterly, 30*(4), 470-487. http://dx.doi.org/10.1037/spq0000116

Shepard, L. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405 – 450.

Smarter Balanced Assessment Consortium. (2/6/2018). Retrieved from: www.k12.wa.us/smarter/

Smarter Balanced Content Specifications. (2015). Retrieved from: https://www.smarterbalanced.org/wp-content/uploads/2015/08/Mathematics-Content-Specifications.pdf

Smarter Balanced Reporting System User Guide. (3/29/2016). Retrieved from: https://portal.smarterbalanced.org/library/en/reporting-system-user-guide.pdf

Smarter Balanced 2014-15 Technical Report. (10/5/2016). Retrieved from: https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf

Swanson, H., Jerman, O., & Zheng, X. (2009). Math disabilities and reading disabilities, can they be separated? *Journal of Psychological Assessment, 27*(3), 173-196. http://dx.doi.org/10.1177/073428908330578

Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics*.(6th ed.) New Jersey, NJ: Pearson.

Thomas, C., Van Garderen, D., Scheuermann, A., & Lee, E. (2015). Applying a universal design for learning framework to mediate the language demands of mathematics. *Reading and Writing Quarterly, 31*(3), 207-234. http:dx.doi.org/10.1080/10573569.2015.1030988

Thurlow, M., Albus, A., & Lazarus, S. (2015). *Graduation policies for students with disabilities who participated in states' general assessments.* Synthesis report 98. National Center on Educational Outcomes.

Thurlow, M., House, A., Scott, D., & Ysseldyke, J. (2000). Students with disabilities in large-scale assessments: State participation and accommodation policies. *The Journal of special Education, 34*(3), 154-163.

Thurlow, M., Lazarus, S., Thompson, S., & Morse, A. (2005). State policies on assessment participation and accommodations for students with disabilities. *The Journal of Special Education, 38*(4), 232-240.

Thurlow, M., Quenemoen, R., & Lazarus, S. (2011). *Meeting the needs of special education students: Recommendations for the Race to the Top consortia and states*. National Center on Education Outcomes.

Tileson, D. (2004). *What every teacher should know about student assessment*. Thousand Oaks, CA: Corwin Press.

Trochim, W. (10/20/2006). *Research methods knowledge base 2nd edition*. Retrieved from: http://www.social researchmethods.net/kb/index.php

Uccelli, P., Galloway, E., Barr, C., Menses, A., & Dobbs, C. (2015). Beyond vocabulary:

Exploring cross-disciplinary academic-language proficiency and its association

with reading comprehension. *Reading Research Quarterly, 50*(3), 337-356.

http://dx.doi.org/10.1002/rrq.104

U.S. Office of Education. (1968). First annual report of National Advisory Committee on

Handicapped Children. Washington, DC: U.S. Department of Health, Education

and Welfare.

Veldkamp, B. (2016). On the issue of item selection in computerized adaptive testing

with response times. *Journal of Educational Measurement, 53*(2), 212-228.

Vogt, W. (2005). *Dictionary of statistics and methodology, a nontechnical guide for the

social sciences*. Thousand Oaks, CA: Sage Publications.

Wakefield, D. (2000). Math as a second language. *The Educational Forum, (64)*3.

http://dx.doi.org/10.1080/00131720008984764

Washington State Board of Education. (3/1/2017). Retrieved from

http://sbe.wa.gov/GradRequirements/ClassOf2017.php

Weekly, S. (1979). *The theories of Jean Piaget – what they mean for special education*

(Unpublished master's thesis). University of Wisconsin, Lacrosse, WI.

Yell, M., Katsiyannis, A., Collins, J., & Losinski, M. (2012). Exit exams, high stakes

testing, and students with disabilities; A persistent challenge. *Intervention in

School and Clinic; 48*(1), 60-64. http://dx.doi.org/1177/105345122449740

Ysseldyke, J., & Bielinski, J. (2002). Effect of different methods of reporting and

reclassification on trends in test scores for students with disabilities. *Exceptional

Children, 68*(2), 189-200.

Ysseldyke, J., Christenson, S., & Thurlow, M. (1987). *Instructional factors that influence*

student achievement: An integrative review. Instructional alternatives project.
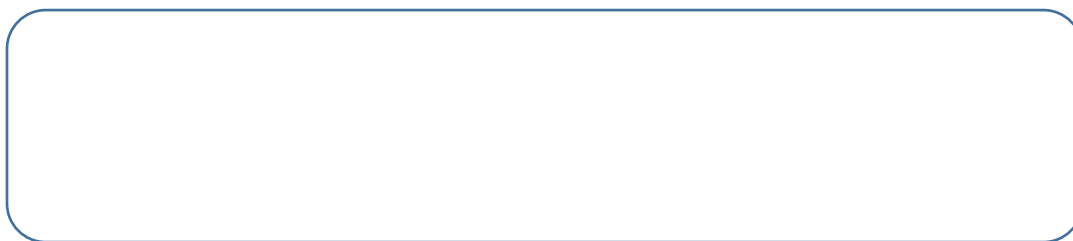
Retrieved from https://files.eric.ed.gov/fulltexted303977.pdf

Appendix A

Sample SBA Math Questions 6th Grade

<u>Question sample 1478 – 6th grade</u>

Design a new cereal box for this company. All cereal boxes are rectangular prisms. Then explain why your design is better for the company, based on the requirements.

In your response, give the dimensions of your box, explain how your box meets each of the requirements for the new box.
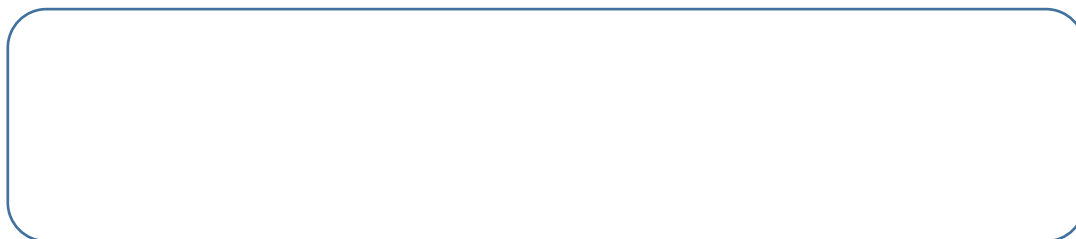
Attached reading sample:

> Cereal Boxes: a cereal company uses cereal boxes that are rectangular prisms. The boxes have the dimensions shown. 12 inches high/8 inches wide/2 inches deep. The managers of the company want a new size for their cereal boxes. The new boxes have to be rectangular prisms. You will evaluate one box design the company proposed. Then you will create and propose your own design for the company. Requirements for the new boxes: the new boxes have to use less cardboard than the original boxes. The new boxes have to hold the same or a greater volume of cereal as the original boxes.

<u>Question sample 3565 – 6th grade</u>

The expressions 4(42)(8.2) and 45 are equivalent.

Show that the two expressions are equivalent. Describe the steps that can be applied to create the equivalent expression 45.

Type your answer in the space provided.

Appendix B

Sample SBA Math Questions 7th Grade

<u>Claim 3 item 2064 – 7th grade</u>

People can save water by taking some proactive steps. Consider and average American household of 4 people. Explain how much water, on average, can be saved each day if they implement the following plan

- They fix one leaky faucet in the hone that drips about 3 drips per second.
- Each person reduced the time in the shower by 3 minutes
- Each person does **not** leave the water running while brushing teeth, washing hands and shaving

Support your answer by including the average amount of water saved by implementing each part of the play, as well as the total amount saved.

*Attached reading passage: **Using Water Wisely***

Water is a valuable resource that can easily be wasted. In this task you will investigate how much water the average American uses each day. You will then investigate how much water a family of 4 could save using different strategies.

According to some estimates, the average American uses 80 – 100 gallons of water daily. Of this total the averages American uses about:

- 27% by flushing toilets
- 25% while taking showers/baths
- 10% by running the faucet while brushing teeth, washing hands, and shaving

Water is also used for various other purposes (cooking, drinking water watering plants, washing clothes, etc.) that account for the remaining percentage of water used by the average American.

Table 1 shows the average amount of water used during some activities.

Table 1. Water used by Activity and duration

Table 1. Water Used by Activity and Duration

Activity	Duration	Amount of Water Used
Showering	11 minutes	20–25 gallons
Running water in bathtub while waiting for water to get hot	$\frac{1}{2}$ minute	3 gallons
Leaky faucet	1 second	2 drips*

*There is no standard definition of the volume of a faucet drip, but the USGS Water Science School uses $\frac{1}{4}$ milliliter (mL) as the volume of a faucet drip. So, by these drip estimates:
- One gallon = 15,140 drips
- One liter = 4,000 drips
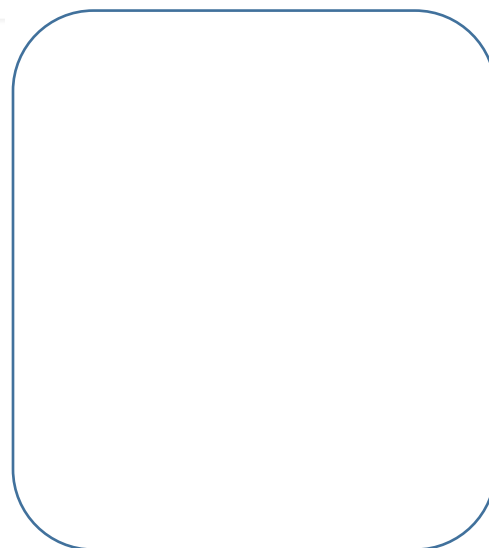
<u>Sample item 3635 – 7th Grade</u>

Alfonso went to Famous Sam's Appliance store and purchased a refrigerator and a stove. The sale price of the refrigerator was 40% off the original price and the sale price of the stove was 20% off the original price.

Which statement must be true to conclude that Alfonso received a 30% overall discount on the refrigerator and stove together?

(A) The sale price of the refrigerator and the stove were the same.

(B) The original prices of the refrigerator and the stove were the same.

(C) The sale price of the refrigerator was twice the sale price of the stove.

(D) The original price of the refrigerator was twice the original price of the stove.

Appendix C

Sample SBA Math Questions 8th Grade

Item 3575 – 8th Grade

The ratios a:b and b:c are equivalent to one another.

Select **all** the statements that must be true.

- a = c
- b/c = c/b
- b-a = c-b
- a < b and b < c
- if a = b, then b = c

Item 1518 – 8th Grade

In this task, you will use data to create a model that shows the relationship between animal body weight and pulse rate measures. Then you will examine additional data to evaluate your model.

A study shows that the relationship between an animal's pulse rate and body weight is approximately linear. The study data are below.

Table 1. Average Body Weight and Average Pulse Rate of Seven Animals

Animal	Average Body Weight (in kilograms)	Average Pulse Rate (in beats per minute)
Cat	3	130
Goat	28	75
Sheep	56	75
Pig	192	95
Ox	362	48
Cow	465	66
Horse	521	34

Interpret the slope of the line from Item 1 in the context of the situation.

Based on the equation from Item 2, predict the average pulse rate in beats per minute of an animal that weights 6000 kilograms.
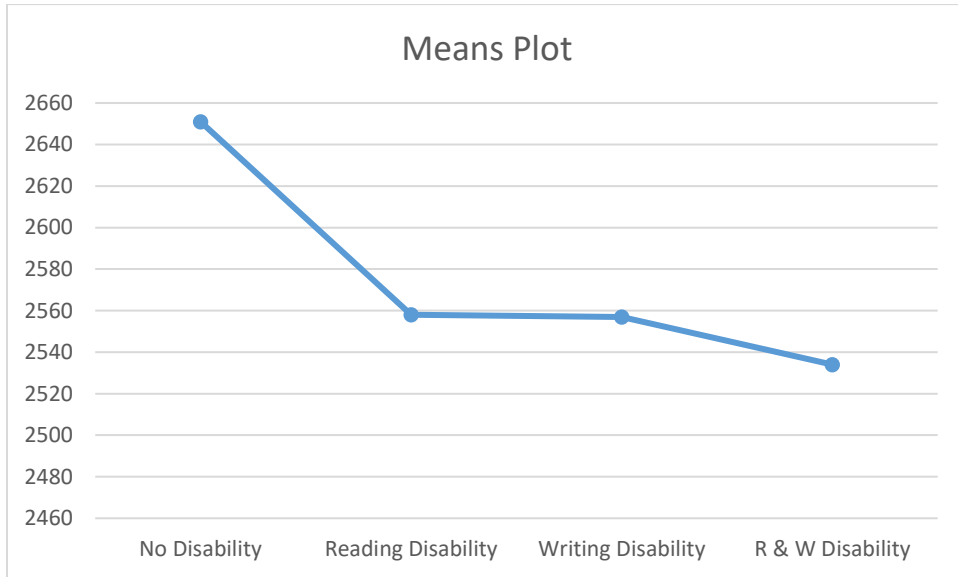
Appendix D

SBA Overall Reliability and Validity Statement

TABLE 2.6 OVERALL SCORE AND CLAIM SCORE PRECISION/RELIABILITY: MATHEMATICS

Grade	Overall Mathematics					Claim 1					Claim 2/4					Claim 3				
	ave # items	$SD(\bar{\theta})$	mean SEM	RMSE	$\bar{p}$	ave # items	$SD(\bar{\theta})$	mean $SE(\bar{\theta})$	RMSE	$\bar{p}$	ave # items	$SD(\bar{\theta})$	mean $SE(\bar{\theta})$	RMSE	$\bar{p}$	ave # items	$SD(\bar{\theta})$	mean $SE(\bar{\theta})$	RMSE	$\bar{p}$
3	39.7	1.0	.25	.25	.94	20.0	1.1	.35	.35	.89	9.9	1.2	.52	.63	.74	9.8	1.3	.61	.79	.63
4	39.2	1.1	.28	.28	.93	20.0	1.1	.38	.39	.88	9.6	1.3	.57	.74	.69	9.6	1.3	.62	.74	.67
5	39.7	1.2	.35	.36	.91	20.0	1.3	.48	.50	.84	9.8	1.6	.64	1.01	.61	9.9	1.4	.65	.87	.63
6	38.8	1.3	.35	.34	.93	19.0	1.3	.47	.46	.88	9.8	1.6	.67	.91	.67	10.0	1.6	.76	.94	.64
7	39.4	1.4	.44	.44	.90	20.0	1.5	.58	.61	.83	10.0	1.8	.81	1.14	.60	9.4	1.7	.95	1.14	.57
8	38.8	1.5	.46	.45	.91	20.0	1.5	.60	.60	.85	9.1	2.0	.86	1.28	.58	9.7	1.7	.88	.98	.66
11	41.3	1.6	.52	.50	.90	22.0	1.6	.69	.68	.83	9.3	2.1	.95	1.31	.60	10.0	1.9	1.04	1.10	.66

From: Smarter Balanced Technical Report, 2015

Appendix E

Histogram of Normal Distribution

Means Plot

Appendix G

Confidence Interval Adjustment Grades 6-8

Sixth Grade Student Scores with Confidence Intervals

Seventh Grade Student Scores with Confidence Intervals

Eighth Grade Student Scores with Confidence Intervals