The Predictive Validity of the Washington Kindergarten Inventory of Developing Skills GOLD's Literacy Domain: Why assessment matters for Washington's earliest readers

Sara Stull

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

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Deborah McCutchen, Chair

Gail Joseph

Sheila Valencia

Program Authorized to Offer Degree:
Education

University of Washington

Abstract

The Predictive Validity of the
Washington Kindergarten Inventory of Developing Skills GOLD's Literacy Domain:
Why assessment matters for Washington's earliest readers

Sara Stull

Chair of the Supervisory Committee:
Professor Deborah McCutchen
Education

This study evaluated the predictive validity of Washington State's recently implemented kindergarten entry assessment, *WaKIDS GOLD* (adapted from *Teaching Strategies* GOLD®), specifically examining the Literacy domain. The primary question of interest was whether teacher-assigned scores from the WaKIDS Literacy domain (assessing phonological awareness, alphabet knowledge, print concepts, reading comprehension, and writing scores) were predictive of later student literacy achievement, as measured by direct, standardized measures of literacy at the end of first grade. Hierarchical linear modeling (HLM) revealed that the WaKIDS Literacy domain was a unique, positive predictor of first grade literacy achievement, despite the concurrent significant effects of first grade teacher and student income. This convergent validity evidence was further supported with hierarchical linear regressions revealing that the individual skill scores within the WaKIDS Literacy domain predicted corresponding literacy skills in first

grade when added to the model together, after controlling for student demographics. When on their own, however, individual skill scores did not demonstrate as much predictive power, with only alphabet knowledge emerging as a unique predictor of first grade literacy skills. An additional HLM unexpectedly revealed that WaKIDS Literacy was a unique predictor of first grade math skill, which did not support discriminant validity, although this could be explained by language demands of the math assessment. Effects of the teacher variable are addressed in terms of instructional implications, and broader implications for the State as well as future validity studies of kindergarten entry assessments are also discussed.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to take this opportunity to acknowledge those who have supported me throughout the dissertation process and my graduate school career. First, I would like to thank my academic advisor and committee chair, Dr. Deborah McCutchen, for her continuous guidance over the past seven years. I have learned so much from you and am grateful for all the time and energy you have put toward assisting me. I would also like to acknowledge my other committee members for their time, expertise, and encouragement: Dr. Gail Joseph, Dr. Sheila Valencia, and Dr. Miriam Bassok – thank you for all of your support.

The data collection for this dissertation would not have been possible if it were not for the help of several individuals. Thank you to Natalie Tuck, Elizabeth Ramirez, Molly Artz, and Jacquelyn Betz for volunteering your time to assist me with administering assessments to first graders. I would have had considerably fewer participants if it were not for you! To Kathe Taylor, Director of Early Learning Assessment at OSPI, and Mary Fickes, Early Learning Coordinator for Seattle Public Schools – thank you for your support of my work and continuing to advocate for me. I am very appreciative of your willingness to help me despite all of the other work demands on your plate. Dr. Eric Anderson, Director of Research, Evaluation, and Assessment for Seattle Public Schools – thank you for allowing me to collect data within your school district and for taking the time out of your busy schedule to help me finalize my data sharing agreement. And to Dr. Walter Trotter, Principal of Greenwood Elementary – thank you for your assistance in working with the district and constant support of my doctoral work.

Also thank you to Julie Lorah and Liz Sanders for offering your statistical expertise and helping me work through my analyses. Janet Soderberg – thank you for always being willing to

offer research advice, as well as your constant encouragement. I am very appreciative of your support and flexibility with my work schedule, particularly during my final weeks of writing.

Finally, I would like to acknowledge my family and friends. Most importantly, my parents – thank you for supporting me every step of the way and encouraging me to complete this degree. I would not have gotten this far if it were not for you pushing me to excel in school from day one and telling me I could when I thought I couldn't. To my sister Kathy and brother Alex – thank you for always listening to me, dealing with my stress, and being such wonderfully weird siblings. To my Grandma and Aunt Jan – thank you for caring so much about me, always checking in, and believing that I could make it through this. To my best friend Erin Gonzales – thanks for being there to listen to me at any time and helping me to not worry but be happy. To all my other childhood friends – thank you for being my friends for so long and making me laugh until it hurts. To Susie Henderson – I don't know what I would have done without you these past few years in Seattle. Thank you for being willing to drop everything for me whenever I needed a friend and cheering me on through every stage in this process. Jamie Phillips – thank you for continuing to tell me I had to graduate and that I could do it. I am grateful to have a friend as loyal and supportive as you. And to my coffee shop buddies, Susanna Eng and Jessi Salvador – thank you for keeping me sane through the writing process and being such a great support network. We made it to the finish line!

CHAPTER 1

Introduction

Purpose of Study

Attention toward the quality of early childhood education in the U.S. has grown in recent years with evidence consistently confirming that children's skill levels during their earliest school years are predictive of later academic success (Duncan et al., 2007; Snow & Van Hemel, 2008). Not only has this led to greater focus on improving instruction in the early years, but also on the assessment systems that inform this instruction through measurement of young students' developing skills. As Graue (1999) stated, "good teaching is linked to knowledge derived from assessed information about each child's status and strategies across time" (p. 134).

The development of literacy skills is one of the critical components in the education of young children today. The ability to read and write is essential for functioning and success in our current, literate society (NELP, 2009; Snow, Burns, & Griffin, 1998; Whitehurst & Lonigan, 2001), and societal demands for literacy are only increasing (Lonigan, Allan, & Lerner, 2011; Snow et al., 1998). In recent years, researchers and educators have come to realize that early literacy learning provides the foundation for later academic success, with consistent evidence of a link between beginning school literacy skills and later literacy, as well as overall achievement in school (Cunningham & Stanovich, 1997; Cunningham, Zibulsky, & Callahan, 2009; Duncan et al., 2007; Snow et al., 1998; Whitehurst & Lonigan, 2001). Unfortunately, not all children experience success with literacy acquisition; children who begin school with lower reading skills than their peers tend to continue to experience difficulty in literacy throughout schooling, without intervention (Cunningham & Stanovich, 1997; Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Juel, 1988). Even more unfortunate, children from low-income and ethnic/racial

minority families along with children with limited English-language proficiency have been found to be most at-risk for experiencing such reading difficulties (McCoach, O'Connell, Reis, & Levitt, 2006; Slavin & Cheung, 2005; Snow et al., 1998; Teale, Paciga, & Hoffman, 2007). For these children, early identification of difficulties and intervention is key in helping them develop the literacy skills they need for success (Spira, Bracken, & Fischel, 2005; Torgesen, 1998).

The challenge lies in early identification of difficulties and understanding children's literacy levels when they enter the formal schooling system. Teachers of kindergarten, the level at which most American children begin formal schooling, are faced with the difficult task each year of determining the incoming developmental abilities of all their students, who often come from varying backgrounds with varying levels of experience (Lonigan et al., 2011;West, Denton, & Germino-Hausken, 2000). This is where the role of early literacy assessment comes into play. While quality literacy instruction is the key to helping young students build the foundational skills and knowledge they need to succeed at reading and writing throughout schooling (Rupley, Blair, & Nichols, 2009; Snow et al., 1998), assessment, when used appropriately, can be a powerful tool for teachers for informing that instruction (Coyne & Harn, 2006; Lonigan et al., 2011). As Lonigan et al. (2011) described, assessments yielding data on "children's developmental achievements in key areas of early literacy can provide teachers with the information they need to provide optimal learning experiences for children" (p. 499). And because early literacy skills are known to predict later literacy achievement, assessments targeted at those skills should be able to accurately forecast children who might face later reading difficulties (Spencer, Spencer, Goldstein, & Schneider, 2013).

With awareness increasing around the importance of early childhood and early literacy assessment, a variety of tools for assessing young children and their environments have become

more available in recent years (Snow & Van Hemel, 2008). While there has been past concern with assessment of young children, particularly regarding assessments of "school readiness" and testing for eligibility, more recent theory has linked assessment more closely with child learning, leading "readiness" to be viewed as more of a measurement of status at kindergarten entry for the purpose of meeting specific child needs (Brown, Scott-Little, Amwake, & Wynn, 2007; Snow, 2011). One approach to measuring this type of readiness that has become more prevalent in recent years is a kindergarten entry assessment program. With children entering kindergarten from a variety of early learning environments and backgrounds (West et al., 2000), and with some, particularly from low-income and racial minority backgrounds, entering with skill levels far behind their peers (Evans & Rosenbaum, 2008; Snow et al., 1998), this first year in the formal schooling system is a "pivotal transition point" for children (Scott-Little, Bruner, & Schultz, 2011, p. 1). Appropriate assessment practices across multiple domains of development, including literacy, can assist with this transition. According to Scott-Little et al. (2011), "data collected at kindergarten entry serve both as a cumulative glimpse into how children's early experiences have (or have not) supported their development and learning and offer a baseline for kindergarten instruction and for measuring future progress" (p. 1).

Development of a kindergarten entry assessment (KEA) program has been encouraged by the U.S. Department of Education through their *The Race to the Top Early Learning Challenge*, launched in 2011 in effort to motivate States to develop plans for improving the quality of their early education systems (ED.gov, 2011). Scott-Little et al. (2011) presented four general goals that States implementing KEAs tend to incorporate in their plans:

> 1) To assess the degree to which children in the state are starting school "ready;" 2) to identify schools and populations of children for which additional efforts are most

> needed to ensure educational success; 3) to provide additional direction to kindergarten teachers in helping their students develop and learn; and 4) to inform parents about their child's learning and development and provide an opportunity to engage parents in supporting their child's learning. (p. 2)

As of January 2014, the Center on Enhancing Early Learning Outcomes (CEELO) reported that a total of 34 states had plans outlined for a KEA in their *Race to the Top* applications, and nine additional states without *Race to the Top* applications were already implementing KEAs (Connors-Tadros, 2014).

Washington State is one of the leaders among these states in the development and implementation of a KEA. The State has adopted *Teaching Strategies* GOLD® for their new kindergarten assessment process known as the *Washington Kindergarten Inventory of Developing Skills* (*WaKIDS*), and it is presently being used by all state-funded full-day kindergarten classrooms across the State for the purpose of providing kindergarten teachers with valuable information about their incoming kindergarteners across six developmental areas, representing the "whole child" (social-emotional, physical, cognitive, language, literacy, and mathematics) (Teaching Strategies, 2011a). GOLD® is an observation-based assessment, and according to Teaching Strategies, GOLD® "blends ongoing, authentic assessment in all areas of development and learning with intentional, focused performance assessment tasks for selected predictors of school readiness in the areas of literacy and numeracy" (Teaching Strategies, 2010a, p. 1). They also stress the importance of remembering that GOLD® "is not intended as a screening or diagnostic measure, an achievement test, or a program-evaluation tool" (Teaching Strategies, 2011b, p. 2). *Teaching Strategies* GOLD® is widely used and has become one of the most commonly selected assessment tools for KEA programs among the states. Burts and Kim

(2014) reported that as of September 2013, approximately one-quarter of the states had agreements with *Teaching Strategies* GOLD® for use as a kindergarten assessment, according to the publishers.

While Teaching Strategies has established GOLD's® reliability and validity with a nation-wide norm sample, and the University of Washington has just completed reliability and concurrent validity studies for the *WaKIDS* version in Washington State, no studies of predictive validity for GOLD® or *WaKIDS GOLD* appear to exist. With Teaching Strategies claiming that GOLD® "measures the knowledge, skills, and behaviors that are predictive or most important for school success" (Teaching Strategies, 2010, p. 2), and with Washington State currently looking to its kindergarten teachers to use *WaKIDS GOLD* results to inform their teaching, verifying the *WaKIDS GOLD*'s predictive validity seems essential. And, to continue improving early literacy instruction and early identification of difficulties with literacy, the predictive validity of the *WaKIDS GOLD's* Literacy domain is particularly important. Therefore, the primary focus of this dissertation was to examine the predictive validity of the *WaKIDS GOLD* assessment's Literacy Domain to confirm that the State's kindergarten entry assessment is a valid measure of the early literacy skills known to be predictive of later literacy achievement.

Objectives

This study used student data collected by Washington kindergarten teachers who participated in WaKIDS during the fall of 2012 to examine the predictive validity of the *WaKIDS GOLD* Literacy domain. Predictive validity is determined by how well performance on an instrument predicts later performance on an established, criterion instrument. (Invernizzi, Landrum, Howell, & Warley, 2008; Rathvon, 2004; Snow & Van Hemel, 2008). For this study, later performance of skills will be specifically defined as literacy achievement at the end of first

grade. Statistical analyses will be used to determine how well WaKIDS literacy scores predict first grade literacy outcomes. Therefore, the *overarching question* for this study is:

Is the WaKIDS GOLD a valid assessment instrument in terms of its ability to predict later elementary student literacy achievement?

To examine the predictive validity of the *WaKIDS GOLD*, a measure of literacy achievement at the end of first grade was needed for students who participated in WaKIDS in fall 2012. Two standardized measures of literacy, each tapping different literacy skills included in the Literacy domain of the *WaKIDS GOLD*, were administered to a sample of first grade students in the Spring of 2014, to obtain the outcome literacy variable. A standardized measure of mathematics skill was also administered to add to the predictive validity findings by providing discriminant validity evidence. The specific objectives of this predictive validity study, encompassing use of these tools, will be articulated further in this study's research questions at the end of the next chapter.

CHAPTER 2

Conceptual Framework and Review of Literature

The purpose of this section is to provide a framework for evaluating the predictive validity of the *WaKIDS GOLD* assessment and its Literacy domain. It is important to establish key elements of early childhood and early literacy assessment, such as *what* assessments of early literacy should assess, as well as *why* they should be used and *how*. It is also necessary to understand the psychometric concepts of validity and predictive validity; specifically how these concepts will be defined for this study and why they need to be evaluated in assessments such as the *WaKIDS GOLD*. Finally, it is essential to review the most appropriate methods and analyses for evaluating the predictive validity of such an assessment. Past theoretical and empirical research will be reviewed in this chapter to establish these important elements of assessment as well as to understand methods for conducting predictive validity studies. This serves to inform this present study's methodologies, analyses, and interpretation of results.

Early Literacy

Literacy is a complex term; as Whitehurst and Lonigan (1998) noted, it has taken on many different meanings over the years and in a broad sense, potentially encompasses any interaction with a symbolic system. My focus, like that of Whitehurst & Lonigan (1998, 2001), is restricted to the *conventional* forms of literacy, which encompasses the reading and writing of alphabetic, school-based texts (it is also referred to as *school-based literacy* by Pelligrini & Galda, 1998). Snow et al. (1998) in their National Research Council Report, *Preventing Reading Difficulties in Young Children*, also specified a focus on conventional forms of literacy (specifically reading). They acknowledged that literacy is "inextricably embedded in educational, social, historical, cultural, and biological realities" (p. 33), but explained that their report needed

to be limited to the conventional forms of reading for the purpose of discussing reading difficulties "as defined by mainstream opinions in the United States" (p. 34). For the same reasons, my discussion of literacy is restricted to conventional forms as well.

It is also important to explain that there are multiple skills and processes involved in the conventional forms of reading and writing that can be defined in many ways, but I will adhere to Snow et al.'s (1998) view of literacy as "not only a cognitive and psycholinguistic activity but also a social activity" (p. 15), requiring "the use of form (the written code) to obtain meaning (the message to be understood), within the context of the reader's purpose (for learning, for enjoyment, for insight)" (p. 33). This is similar to the original "simple view of reading" defined by Gough & Tunmer (1986) as reading equaling the product of decoding and linguistic comprehension, stressing that both, and not just one, are needed for reading comprehension to occur. Snow et al.'s view adds the component of reader purpose, which they believe to be just as necessary. While literacy assessments may primarily focus on measuring students' abilities with their use of form (written code), it cannot be forgotten that meaning and purpose serve a central role in reading and writing as well. This is why some researchers and councils have come to define conventional literacy skills as five essential skills needed to read and write independently: decoding, oral reading fluency, comprehension, writing, and spelling (NELP, 2009; Whitehurst & Lonigan, 1998).

My conceptualization of *early* literacy primarily draws from Whitehurst and Lonigan's (1998) theory of emergent literacy. Whitehurst and Lonigan developed a definition of *emergent literacy* applying the beginning work on this concept of others (Clay, 1966; Fitzgerald, Schuele, & Roberts, 1992; Sulzby & Teale, 1991; Teale & Sulzby, 1986), as: "the skills, knowledge, and attitudes that are presumed to be developmental precursors to conventional forms of reading and

writing and the environments that support these developments" (p. 849). Under this view, the acquisition of literacy is considered a developmental continuum, beginning at the start of a child's life, before formal schooling begins. This stood apart from other theories that existed at the time of "reading readiness," carrying the notion that literacy acquisition does not truly begin until a child reaches a certain cognitive level or definable point in time, typically when formal schooling begins. This theory of emergent literacy recognizes that children can be exposed to literacy-rich environments and social interactions before schooling begins that constitute the beginning development of *conventional* literacy skills. As summarized by Sénéchal, LeFevre, Smith-Chant, and Colton (2001), this emergent literacy perspective has both neo-Piagetian and neo-Vygotskian elements. It is neo-Piagetian in the sense that children learn about literacy prior to the start of school through play and discovery while exploring their surroundings; it is neo-Vygotskian in the sense that children also learn about literacy through their interactions and socializing with more experienced others. Snow et al. (1998) adhered to a similar developmental perspective in their report, stressing that the appropriateness of participation, instruction, and assessment in literacy are all dependent on a child's developmental level.

Whitehurst and Lonigan specified that the skills and knowledge referred to in their emergent literacy definition can be categorized into one of two interdependent domains, which extend into conventional literacy as well: *outside-in* and *inside-out* processes (also referred to as [oral] language and code-related skills, respectively; Lonigan, 2006). Outside-in processes involve "children's understanding of the context in which the writing they are trying to read (or write) occurs" (Whitehurst & Lonigan, 1998, p. 854). Literacy skills falling under this domain include vocabulary, narrative construction, and conceptual and semantic knowledge that can support comprehension (Sénéchal et al., 2001; Whitehurst & Lonigan, 2001). On the other hand,

inside-out processes involve, "children's knowledge of the rules for translating the particular writing they are trying to read into sounds (or sounds into print for writing)" (Whitehurst & Lonigan, 1998, p. 854). Literacy skills classified under this domain include phonological and syntactic awareness, as well as print awareness/concepts (Sénéchal et al., 2001; Whitehurst & Lonigan, 2001). Whitehurst and Lonigan claimed that both outside-in and inside-out processes work together simultaneously and are essential for successful reading.

Longitudinal studies applying this conceptualization of emergent literacy have confirmed that both inside-out and outside-in domains (measured in preschool), predict later reading success (use of conventional literacy skills) in elementary school (Lonigan, Burness, & Anthony, 2000; NELP, 2009; Storch & Whitehurst, 2002). Whitehurst and Lonigan (2001) explained that the strongest links to later reading achievement could be specified as one outside-in skill, *oral language*, and two inside-out skills, *phonological processing* and *print awareness*, based on available evidence. The National Early Literacy Panel (2009) in their *Developing Early Literacy Report* applied Whitehurst and Lonigan's theoretical framework and identified 11 different skills, from review of 299 articles, that consistently predict later literacy achievement for preschoolers and kindergarteners. The six strongest predictors were: alphabet knowledge, phonological awareness, rapid automatic naming (RAN) of letters or digits, RAN of objects or colors, writing/writing name, and phonological memory; followed by five moderate predictors: concepts about print, print knowledge, reading readiness, oral language, and visual processing. While many have called this NELP finding into question because of the greater role they believe oral language skills to play in early literacy development (e.g., Neuman, 2010), research finding different developmental trajectories for code-related and oral language skill domains can help explain this. Storch and Whitehurst (2002) found both code-related and oral language processes

present during preschool, but code-related skills dominated from kindergarten to 2nd grade, with oral language not returning until the 3rd-4th grades. This varying relationship between the two is appropriate considering that kindergarten to 2nd grade instruction is typically more concerned with reading accuracy, which involves code-related skills, while comprehension skills, in the oral language domain, play a greater role in the later elementary grades when more complex texts are used (Whitehurst & Lonigan, 1998). But, these findings confirm that *both* skill-types play a direct role in early literacy development with both being present prior to kindergarten.

Furthermore, Paris (2005) argued that some of these code-related skills, in particular letter knowledge, phonics, and concepts of print, are *constrained* skills because they are mastered completely and universally by most students before more sophisticated literacy skills develop. Therefore, when measured, variability will only be evident during their short periods of development. Unconstrained skills, such as vocabulary and comprehension, follow continuous growth trajectories over time and their measurement will yield normal distribution. Paris claimed that many past interpretations about literacy development are flawed because of constrained skills being analyzed like unconstrained skills. While code-related skills are essential elements of learning to read and write, any analysis that has determined them to be strong predictors of later literacy and assumed a normal distribution may have invalid results. This must be considered when developing and using early literacy assessments.

Another component to Whitehurst and Lonigan's (1998) definition of emergent literacy involves the environments that support outside-in and inside-out skills before school begins. The self-discovery and sociocultural components that shape emergent literacy under this view vary for each child depending on their home environments. Because of differences in literacy practices across family and cultural environments, it is inevitable that there will be wide-ranging

variability in the emergent literacy skills children bring with them to kindergarten, with some children more prepared for the transition to school-based literacy practices than others (Heath, 1983; Lonigan, 2006; Snow et al., 1998). Such experiences with transitioning to the culture of the formal school system can affect children's development of attitudes toward literacy, which is the third component of Whitehurst and Lonigan's definition. If children experience difficulty with learning to read and write in the early grades, this can negatively impact their motivation toward literacy and instruction in general as they progress through school (Snow et al., 1998). It is therefore important that environmental and motivational factors be considered alongside children's skills and knowledge when assessing their early literacy development.

While *emergent literacy* and *early literacy* are sometimes used interchangeably, it is important to establish that my conceptualization of early literacy for this study is not entirely equivalent to the emergent literacy theory just described. The emergent literacy definition put forward by Whitehurst and Lonigan (1998) is specific to children and their literacy learning *before* they begin school and receive formal instruction. For the purposes of discussing early literacy assessment, I will view early literacy skills as encompassing any skills that need to be acquired before a child becomes independently, conventionally literate, with emergent literacy referring to the beginning stages of these skills in the years prior to schooling. Therefore, skills that Whitehurst and Lonigan, as well as others, identify as emergent literacy skills are all early literacy skills, falling into the predominant categories of phonological awareness, print awareness, and oral language. However, because literacy acquisition is a continuous process, there is no clear boundary between early and conventional literacy. And, for assessing children's early literacy *during* the beginning years of school, measurement will include a mix of emergent and conventional literacy skills because of the variation that will exist in development of skills.

Therefore, for the purposes of discussing assessment, early literacy will be defined as a period of time between the preschool ages and approximately 3rd grade in which emergent literacy skills are transforming into conventional, independent literacy abilities. I have chosen to begin with preschool and not birth for the purpose of a range including ages during which formal instruction is likely to occur. While not all children attend preschool, many early literacy assessments are designed for those who do. And 3rd grade was chosen because of the change in reading instruction and assessment between 3rd and 4th grade, when independent reading of more complex, often nonfiction texts is expected (often referred to as going from "learning to read" to "reading to learn") (Snow et al., 1998).

Looking specifically at the early literacy skills assessed in the *WaKIDS GOLD* Literacy domain, the included objectives and dimensions appear to adequately cover these components of early literacy. Table 1 provides the specific dimensions of the WaKIDS Literacy domain linked with one or more of these predominant categories of early literacy (phonological awareness, print awareness, or oral language) with which it best aligns. This table reveals that phonological awareness is slightly more represented across the domain than the other two, but the three are nearly evenly represented. Also, each objective/dimension of the *WaKIDS GOLD* features an expected progression of the development of that skill from infancy to kindergarten, and this aligns with Whitehurst and Lonigan's view of literacy acquisition as a developmental continuum.

Table 1. *WaKIDS GOLD Literacy dimensions aligned with early literacy categories*.

WaKIDS GOLD Literacy Dimension	Early Literacy Category
15a. Notices and discriminates rhyme	Phonological awareness
15b. Notices and discriminates alliteration	Phonological awareness
15c. Notices and discriminates smaller and smaller units of sound	Phonological awareness
16a. Identifies and names letters	Print awareness
16b. Uses letter-sound knowledge	Phonological awareness, Print awareness
17b. Uses print concepts	Print awareness
18a. Interacts during read-alouds and book conversations	Oral language
18b. Uses emergent reading skills	Print awareness, Oral language
18c. Retells stories	Oral language
19a. Writes name	Print awareness
19b. Writes to convey meaning	Phonological awareness, Print awareness, Oral language

Early Childhood Assessment

As articulated in the Introduction, early childhood and early literacy assessment can serve as powerful tools for educators in understanding children's developmental levels when they first enter the school system. Snow and Van Hemel, authors of the comprehensive 2008 report *Early Childhood Assessment: Why, What, and How*, provided a definition of *assessment*, "gathering information in order to make informed instructional decisions" (p. 27), which highlights the beneficial outcome of its use. However, there are many different assessments currently available to educators with varying purposes and formats, and educators must be able to choose the most appropriate assessment for their specific purposes in order to reap the benefits. This section will describe the many different purposes and formats typical of early childhood and early literacy assessment, which will add to the framework to be used for understanding and evaluating the *WaKIDS GOLD* assessment and its Literacy domain.

Kame'enui et al. (2006), stressed the importance of understanding an assessment's purpose in their critical review of beginning reading assessments, as did Snow and Van Hemel, explaining that the purpose is needed to guide all subsequent decisions about the design and use of assessment, such as what to assess and how to interpret data. Snow and Van Hemel presented four main purposes that existing assessments of child learning and development, in general, tend to serve, and also provided more specific purposes associated with each of these main purposes. For example, assessments designed for the first main purpose they presented, *determining an individual child's level of functioning*, can serve the specific purposes of individual-focused screening, community-focused screening, diagnostic testing, or, one of the main purposes of the *WaKIDS GOLD* assessment, establishing readiness for a particular educational program. Their second main purpose, *guiding intervention and instruction*, typically refers to assessments used for planning and monitoring children's progress or response to intervention (RTI). Assessments serving the third purpose of *evaluating the performance of a program or society* are often designed to assess program effectiveness, program impacts, or social benchmarking. And finally, assessments sometimes serve the purpose of *advancing knowledge of child development*, when they are designed and used for research purposes. The second and third main purposes could be considered additional purposes of the *WaKIDS GOLD* because of the effects that individual students' scores are meant to have on instruction, as intended by both Teaching Strategies and WA State, and because of the State's intention to evaluate the developmental levels of their incoming kindergartners each year with the student data. But the purpose of determining an individual child's level of functioning, specifically for establishing school readiness, is the one of these four best aligned with the *WaKIDS GOLD*.

Kame'enui et al. (2006) also defined categories of assessment purpose typically found among assessments of early reading, but described them in terms of the decisions these assessments can be used to make. They defined four "decision-making purposes" (some of which overlap with some of Snow and Van Hemel's specific purposes) as: *screening* (brief assessments targeting skills predictive of future outcomes with the goal of identifying children at-risk of failure and in need of particular instruction), *diagnosis* (thorough assessment of strengths and weaknesses to help teachers shape instruction for progress toward desired outcomes; note that this interpretation of diagnosis is different from the diagnostic testing educators often use for identifying students with disabilities), *progress monitoring* (frequent testing at set points throughout the year to estimate rates of improvement and evaluate effectiveness of instruction), and *outcome evaluation* (end of school-year assessment to determine student progress relative to grade-level expectations). An assessment can serve more than one of these purposes at a time. The *WaKIDS GOLD*'s purpose best aligns with the diagnosis category, as it is designed to help teachers identify strengths and weaknesses of student development in order to inform instruction. Regardless of the purpose(s) that any assessment serves, it needs to be clearly articulated so that assessment users can select the most appropriate tool for their needs (Kame'enui et al., 2006).

Assessments will also vary by format, or mode, which is another crucial component for educators to think about along with purpose when selecting an assessment, according to Snow and Van Hemel. Snow and Van Hemel, along with Snow (2011), both identified two predominant, yet very different formats that are common among early childhood assessments: *direct assessments* (sometimes referred to as *individualized*; involves an adult sitting with one child at a time and asking the child to respond to the various requests of the assessment), and *observation-based* (sometimes referred to as *authentic*; can be individualized but more likely to

happen by observing groups of children or a class as a whole when engaged in various activities). The *WaKIDS GOLD* is a prime example of an observation-based assessment, as it was designed to occur within the everyday instruction of the classroom with teachers collecting observational records on each child over time. The Minnesota Human Capital Research Collaborative (2011) identified similar formats in discussing assessments measuring school readiness: *direct assessments, screening instruments,* and *performance assessments*. Performance assessments are nearly equivalent to observation-based, as they are typically curriculum-embedded and completed via observation in the classroom over time. Screening instruments were not presented by Snow (2011) and Snow and Van Hemel (2008) but were described by the Minnesota Human Capital Research Collaborative as brief assessments measuring specific skills to determine a child's general performance level or if the child needs additional services; this assessment format aligns with Kame'enui et al.'s (2006) decision-making purpose of screening.

Snow (2011) addressed one other important assessment characteristic: the classification of an assessment as *standardized* or not. He noted that there is a common misconception to view standardized assessments for young children as use of "paper and pencil en masse, similar to perceptions of large-scale standardized assessments used for older children" (p. 10). Instead, *standardized* simply refers to the assessment being designed so it can be administered in the same way each time. Violation of this can threaten the assessment's reliability and validity, so it is imperative that assessments used in educational settings are standardized. Although data and documentation methods will differ for each child when using *WaKIDS GOLD*, the assessment is designed to be administered and scored the same way each time by all educators, and so it is considered standardized.

Assessment Validity

Validity overview. Validity is a crucial psychometric property of all types of educational assessment. It has been interpreted in many different ways over time, with much debate around the specific criteria required to determine that an assessment is valid (Moss, 1992; Snow & Van Hemel, 2008). There is general agreement that validity is "evidence showing that the assessment does what it claims to do, namely, that it accurately measures a characteristic or construct" (Snow & Van Hemel, 2008, p. 182). However, many have argued that this is only part of it. One of the predominant validity theorists, Samuel Messick, viewed validity as a "unified, yet faceted, concept" that can be defined as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (Messick, 1989, p. 13-14). It is important to note his emphasis on the need for both empirical evidence and theoretical rationales in evaluating the validity of an assessment; both are essential components for understanding the *construct* being measured, an understanding which Messick viewed as central to the concept of validity. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) present a conceptualization of validity informed by the work of Messick, also stressing the importance of theory and empirical evidence in supporting an assessment's validity.

Messick also asserted that validity is "a matter of degree," (an assessment does not simply have or not have validity) and evolves over time as new findings continuously emerge (Messick, 1989, p. 13). Rathvon (2004) concurred with this, stating "validity is not inherent in a test but is acquired through a continuing process of theoretical and empirical analysis" (p. 53). Messick's view of validity stood apart from others at the time because of its integration of

different "types" of validity into one (entirely focused on the validity of a construct), as well as including implications and social consequences resulting from test use as a necessary component. While validity theorists differ on how consequences of assessment fit into their views of validity, the *Standards for Educational and Psychological Testing* consider consequences to be one of five main sources of evidence for validity and they are important to consider regardless when assessments are used to make decisions about students (Snow & Van Hemel, 2008).

Predictive validity. *Predictive validity* has traditionally been viewed as a type of validity within the broader category of *criterion validity* (along with *concurrent validity*), as it has typically involved the correlation of scores with a *criterion* of interest (Cronbach & Meehl, 1955; Lissitz & Samuelson, 2007). More specifically, it is determined by how well performance on an instrument predicts *later* performance on an established, criterion instrument (Invernizzi, Landrum, Howell, & Warley, 2008; Rathvon, 2004; Snow & Van Hemel, 2008). While not referred to as *predictive validity*, it is represented in another of the *Standards for Educational and Psychological Testing's* main sources of evidence for validity, "evidence based on relations to other variables," emphasizing its importance in a validity argument (AERA, APA, & NCME, 1999). Rathvon (2004) explained that under the view of the *Standards* and Messick, any type of criterion-related validity should not be viewed as a separate type of validity but as "additional sources of evidence that contribute to an understanding of the construct measured by a test," hence the use of "main sources of evidence" defined by the *Standards*, rather than "types of validity" (Rathvon, 2004, p. 52).

Two other "types" of validity can serve as such additional sources of evidence for validating assessments: *convergent* and *discriminant* validity. If an assessment is found to highly

correlate with separate measurement of a theoretically related construct (e.g., a measure of rhyming skill correlating with a measure of overall phonological skill), evidence of *convergent validity* is present and can support the assessment's predictive validity. Alternatively, if an assessment does not correlate with measurement of a construct that is theoretically different (revealing that it does not measure a construct it is not intended to measure; e.g., a measure of rhyming skill not correlating with a measure of math skill), this provides evidence of *discriminant validity* (also known as *divergent* validity), which can also provide support for an assessment's predictive validity (Messick, 1989; Rathvon, 2004).

In their discussion of literacy assessment, Invernizzi et al. (2008) described predictive validity to probably be "the most critical focus for test developers," because of the importance of accurately predicting "real reading outcomes" for students (p. 199). Other researchers have also expressed the necessity for establishing strong predictive validity in early childhood assessments because of the crucial role of these assessments in predicting later educational outcomes, sometimes leading to decisions about student instruction and placement in educational programs (Bordignon & Lam, 2004; Caffrey, Fuchs, & Fuchs, 2008; Coyne & Harn, 2006; LaParo & Pianta, 2000; Pianta & McCoy, 1997; Rafoth, 1997; Snow, 2011). This role of assessment gets to the heart of Messick's (1989) view of validity as encompassing consequences of test use.

Despite this need for strong predictive validity, early childhood and early literacy assessments have often been criticized for having low predictive validity (Bordignon & Lam, 2004; Kame'enui et al., 2006; LaParo & Pianta, 2000), or failing to report any evidence of it altogether (Rathvon, 2004). Another issue lies in the fact that predictive validity is difficult to establish because it requires "longitudinal tracking of children or large retrospective data analyses over time," and for this reason, Rafoth (1997) explained, it is often ignored by test

developers (p. 132). Also, it is rare that studies will attempt to cross-validate results on a different subset of subjects and therefore some assessments only have initial predictive validity findings, potentially reducing the generalizability of the results (Pianta & McCoy, 1997). For these reasons, it is essential that educators, particularly in the early childhood years, carefully consider the predictive validity of an assessment before selecting it for use in schools.

Evaluating Predictive Validity

Predictive validity studies have been conducted on multiple types of early childhood and early literacy assessments, and because of this, the methods for determining predictive validity of an assessment vary considerably. This is not surprising considering the findings of Kim and Suen's (2003) meta-analysis of 44 predictive validity studies of varying types of early assessment, conducted to better understand the generalizability of such studies. They found considerable variability across the studies they reviewed, regarding both study designs and findings, which led them to conclude that predictive power of early assessments is situation-specific and not generalizable. The validity varied based on assessment type, construct, prediction length of time, and administration procedures in the studies they reviewed. This is not to say that previous predictive validity designs and results cannot serve as models for future studies; however, it is important to consider the caution that Kim and Suen prescribed about taking specific testing properties and conditions into account when evaluating the predictive abilities of early assessments, particularly because an assessment could be found valid in one situation but not in another (Rathvon, 2004). Therefore, studies evaluating measures similar to the *WaKIDS GOLD* and the conditions in which it is used (Washington kindergarten classrooms) would serve as the best models for implementation of the present study. Unfortunately, there are a limited number of predictive validity studies of early childhood observational assessments

currently available. This is due to the fact that direct assessments have been used more frequently and for longer than observational assessments, leading to more research of them; regular use of observational assessments in classroom settings has only occurred more recently, explaining why their validity research is limited (Snow & Van Hemel, 2008). Nevertheless, a few predictive validity studies examining observational assessments have been published in recent years and their methods have helped inform the present study.

Predictive validity studies of observational assessments. Sekino and Fantuzzo (2005) studied the predictive validity of the Child Observation Record (COR), an observational assessment measuring a range of skills in children from ages two years, six months, to six years, in 242 Head Start preschool children and looked to predict kindergarten literacy performance. Like *WaKIDS GOLD*, the COR requires teachers to collect anecdotal observations of children across items from six different skill domains which "serve as a basis for teacher's ratings of various classroom competencies on the 5-point developmental sequence" (p. 244). Meisels and colleagues conducted studies (2001, 2008) on two different versions of the Work Sampling System (WSS), a whole-child, observation-based, curriculum-embedded performance assessment, measuring a range of different developmental skills; very similar to *WaKIDS GOLD* as well. Meisels, Bickel, Nicholson, Xue, and Atkins-Burnett (2001) looked at the original WSS assessment in 345 kindergarten through 3rd grade students, looking to predict end-of-year achievement as measured by a standardized psychoeducational battery, and Meisels, Xue, and Shamblott (2008) examined a preschool version of WSS by using the scores of 112 Head Start students to predict end-of-year mathematics and reading performance. Both of these studies were very focused on validating the use of teacher judgments as the basis for measuring students' developing skills in performance assessments.

Gallant's (2009) study of the South Carolina Readiness Assessment, an adaptation of the Work Sampling System for the state of South Carolina, in 1,281 1st graders, across 132 classrooms in one urban school district of the state, provides a third example of a predictive validity study of an early observation-based assessment. Gallant looked to determine the assessment's ability to predict students' 3rd grade language/literacy and mathematics performance on a standards-based, criterion-referenced assessment also used by South Carolina. For all of these early observational assessment studies, evidence supporting predictive validity was present, revealing that assessments using early childhood/elementary teacher observation and report can successfully predict later student achievement.

While these studies just described all have their own unique designs, which Kim and Suen (2003) emphasized must be remembered, there are some trends in their methods that are important to consider. First, all studies used a very large sample size (the smallest was Meisels et al. (2008) with 112 students), which is likely important in validity studies to ensure adequate variability necessary for analyses. Second, while the predictive utility of these assessments was determined in different aged students, the length of time between administration of the predictor and criterion measures ranged from one to two years, with the exception of Sekino and Fantuzzo's (2005) prediction occurring from end of preschool to beginning of kindergarten (specific months were not specified). A one to two year prediction period seems to be a good range for yielding predictive strength. Kim and Suen's meta-analysis also looked into this particular element of predictive validity studies and found that those using a one-year prediction length had the strongest predictive validity coefficients, although a second analysis did not reveal a difference between one and two year studies.

Best practices for predictive validity studies. Other predictive validity studies looking at non-observational early literacy assessments have revealed that particular early literacy skills will yield stronger predictions depending on the time of year students are assessed. Catts, Petscher, Schatschneider, Bridges, and Mendoza (2009) and Hosp, Hosp, and Dole (2011) both discovered this in examining the potential impact of bias on the predictive validity of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). In Catts et al., performance of kindergarten students for all five subtests of DIBELS was used to predict 3rd grade reading achievement (determined by the Oral Reading Fluency DIBELS measure and the Reading Comprehension subtest of the SAT-10), and in Hosp et al., performance of 1st through 3rd graders on the Nonsense Word Fluency and Oral Reading Fluency was used to predict performance on the English/Language Arts component of a state criterion-referenced test at the end of the school year. Bias was evident in the findings of both studies in the form of floor effects. Hosp et al. explained that floor effects occur when an abundance of scores fall near the lower end of the distribution because of the measurement scale not extending low enough to capture the full range of abilities. This then weakens the predictive validity of the measure because of its inability to capture the actual distribution of scores. Interestingly, Catts et al. and Hosp et al. similarly found that these floor effects were consistently present for the initial administrations of the measures (in Catts et al. for kindergarten, and in Hosp et al. for 1st and 2nd, but not 3rd, grades), and then went away in later administrations. Both studies attributed this pattern of floor effect occurrence to the measure being administered too early. With students bringing in differing prior knowledge and experiences at the beginning of the year, some had not yet gained the experience through instruction they needed with particular early literacy skills, such as alphabet knowledge and phonological awareness. As instruction progressed through the year, students had enough

experience with learning these skills for assessments to accurately measure their ability, leading the floor effects to disappear.

Linklater, O'Connor, and Palardy (2009) came across a similar finding in their study of phonemic awareness measures' ability to predict end-of-year kindergarten performance in both ELL and non-ELL students. Their discussion presented ideas similar to those posed by Catts et al. and Hosp et al. above:

> Collecting measurements near the beginning and end of the period of consideration (kindergarten for the present study) is important for obtaining reliable growth estimates. However, in cases where the child is developmentally unprepared for testing with a given instrument, reliability of the growth estimate can be improved by collecting additional measures after he or she has had more exposure to instruction in language and reading domains. (p. 391)

These findings of bias may be further explained by Paris' (2005) theory of constrained variables. As described previously, literacy skills such as alphabet knowledge, phonemic awareness, and oral reading fluency are constrained because they are mastered completely and universally by most students before more sophisticated literacy skills develop. If these skills are analyzed assuming a normal distribution, variability will only be evident during their short periods of development. Paris therefore advised researchers to not use constrained variables such as these as predictors unless analyses account for the non-normal distribution. So, before additional measures are planned later in the year, researchers need to consider their use of constrained variables as predictors and possible adjustments that might need to be made. *WaKIDS GOLD* is administered by teachers at the beginning of the kindergarten year in Washington so this bias could be present and may need to be accounted for in analyses.

Another factor to consider in the design of a predictive validity study is selection of the most appropriate outcome measure(s). There are many factors involved in making this choice, however, one that is of more concern for this study's purposes regards the *format* of outcome measure. Many of the predictive validity studies reviewed seemed to use the same format of measure (often direct assessment) for both predictor and outcome measures. But, with observation-based assessments, this cannot always occur because of the limited availability of other observation-based assessments with psychometric properties already well-established. The Meisels et al. studies provide examples of a predictive validity studies using an outcome measure of a different format than the predictive measure, which is referred to as *method bias* or *method variance*. Specifically, the outcome measures chosen were direct, norm-referenced assessments (Meisels et al. (2008) used the *TERA-3* for reading and *TEMA-3* for math; Meisels et al. (2001) used the Woodcock Johnson Psychoeducational Battery Revised), which record student performance quite differently from the WSS observation/teacher report system. Both discussed the potential limitation of this method variance because, while both were designed to measure the same skills, Meisels et al. (2008) argued that "differences in results using the two types of assessments are highly probable simply because the assessments measure different though overlapping parameters in different ways" (p. 977). Meisels et al. (2001) explained that their choice of a direct, standardized measure was because it was the best available option. Despite this, both studies aimed "to maximize the overlap between the two indicators of achievement" (Meisels et al., 2008, p. 977), and were in fact successful in finding a significant relationship between the WSS scores and later standardized measures. Sekino and Fantuzzo (2005) made a similar decision using DIBELS subtests as their outcome measure, and Gallant (2009) likewise with a multiple-choice state assessment. These studies not only provide evidence that a direct,

standardized assessment can effectively serve as a criterion when assessing the predictive validity of an observation-based assessment, but also contribute a model for future predictive validity study designs using similar types of assessments.

Statistical analyses. Looking at statistical analyses used in predictive validity studies, linear regressions (both multiple and hierarchical) and hierarchical linear modeling are most commonly used to determine the strength of the relationship between predictor and criterion variables. Thorkildsen (2005), in her text book on measurement, claimed that investigators tend to rely on multiple regression equations when conducting predictive validity studies. Four studies reviewed used only simultaneous multiple linear regressions (Augustyniak et al., 2004; MacDonald, Sullivan, & Watkins, 2013; Missall et al., 2007; Sekino & Fantuzzo, 2005). Others used step-wise multiple linear regressions or hierarchical regressions, either alone or along with other types of regression analyses (Betts, Pickart, & Heistad, 2009; Bridges & Catts, 2011; Caffrey et al., 2008; Hecht & Greenfield, 2001; Jitendra, Sczesniak & Dearline-Buchman, 2005; Linklater et al., 2009; Meisels et al., 2008; Munger & Blachman, 2013). These types of analyses were used in order to understand how much of the outcome variance could be uniquely explained by not only the main predictor variable of interest but also a second factor for comparison purposes. For example, in Betts et al.'s (2009) predictive validity study of the Minneapolis Kindergarten Assessment (a standardized, direct assessment battery of early literacy and numeracy skills in kindergarten), a step-wise regression was used to understand the unique contributions of kindergartners' performance on the literacy versus numeracy subtests to their reading and math achievement in 2nd grade, in addition to using standardized regression coefficients to understand the overall predictive validity of the assessment. The researchers chose to combine measurement of both literacy and numeracy skills, rather than assess separately,

because of the potential to increase overall information regarding the development of formative early learning skills at the start of schooling. Including both skills also allowed for discriminant validity to be examined within their analyses. As described previously, discriminant validity adds to the overall validity of an assessment with evidence that the assessment does not measure a construct it is not intended to measure (Messick, 1989). Betts et al. tested for this validity by correlating kindergarten literacy scores with 2nd grade math scores, and kindergarten math scores with 2nd grade literacy scores, to ensure that literacy scores predicted literacy outcomes more strongly than math scores, and vice versa.

Hierarchical and step-wise regression models have also allowed researchers to control for variance contributed by demographic variables, such as age, gender, race/ethnicity, and ELL status (Can, Ginsburg-Block, Golinkoff, & Hirsh-Pasek, 2013; Linklater et al., 2009; Meisels et al., 2008; Meisels et al., 2001), as well as to establish the predictive validity abilities of a particular test type by comparing it to a more traditionally-used type assessing the same skill (such as dynamic compared to static assessment) (Bridges & Catts, 2011; Hecht & Greenfield, 2001). In studies with individuals nested within group variables, such as students within classrooms or schools, a multi-level model approach, often hierarchical linear modeling (Raudenbush & Bryk, 2002), was used to account for any variance explained by the group variable in the predictive relationship (Downer, Lopez, Grimm, Hamagami, Pianta, and Howes, 2012; Gerde, Bingham, and Pendergast, 2015; Henry et al., 2013; Musu-Gillette, Barosky, and List, 2015).

Present Study

Considering the purpose and format of the *WaKIDS GOLD* assessment, along with the importance of the early literacy skills assessed by the Literacy Domain, it is clear that the

predictive validity of Washington kindergartners' literacy scores needs to be understood in order for the State to truly benefit from the data. If strong evidence of predictive validity does exist, the literacy scores of this assessment will have great potential to assist the State with improving their understanding of incoming kindergarten students' literacy skill levels, as well as with informing literacy instruction and potentially providing needed interventions for certain students; all of which will help with predicting and then improving later student literacy outcomes.

Therefore, for the present study, the following research questions were formulated to reveal the clearest evidence possible for predictive validity of the *WaKIDS GOLD* Literacy domain. These questions are summarized below:

Research Question 1a: Do kindergarten student scores from the WaKIDS Literacy Domain have a statistically significant *predictive* relationship with scores of student literacy achievement at the end of first grade, taking student demographic and teacher variables into account? Is the relationship strong enough to provide evidence of predictive validity?

Research Question 1b: Are the individual literacy skills measured by objectives within the WaKIDS Literacy Domain uniquely predictive of corresponding literacy skills measured at the end-of-first grade? Subsequently, are particular WaKIDS Literacy objectives stronger predictors of end-of-first-grade literacy achievement scores than others? How do the revealed relationships affect the Literacy Domain's overall predictive validity?

Research Question 2: Is there evidence of *discriminant validity* to support this predictive relationship? Specifically, are WaKIDS Literacy scores *not* uniquely predictive of end of first grade achievement scores in a conceptually/theoretically different domain of learning (i.e., mathematics)?

Applying the methods and analyses reviewed above, the following chapters will outline the plan for carrying out this predictive validity study, including participant recruitment, administration of measures, procedures, and data analyses.

CHAPTER 3

Method

Participants

Participants were 64 first grade students, 48.4% female and 51.6% male, from six public schools in a large urban district in Washington State. Across the six schools, the student participants were recruited and sampled from 14 different classrooms. Students needed to have been assessed with the *WaKIDS GOLD* during the fall of their kindergarten year in order to participate. The racial composition of this sample was 25% Asian, 20.3% Hispanic, 20.3% White, 17.2% Black or African American, 15.6% two or more races, and 1.6% Native Hawaiian. The majority of the sample qualified for free-or-reduced price meals (75%), and 6.3% of the students had an individualized education program (IEP). At the time of end-of-first-grade assessment, students ranged from 6.8 to 8.4 years of age, with a mean age of 7.4 years. This demographic information is summarized in Table 1, and is presented in comparison to the demographics of the WaKIDS statewide sample (N = 21,811) for the same 2012-2013 year.

Students were unevenly distributed among first grade classrooms and schools due to the study's convenience sample: low return rates for parent consent forms in some classrooms led to very few participating students, while other classrooms had many. Also, not all first grade teachers at each school agreed to participate, which led to a varying numbers of participating classrooms at each school. Table 2 breaks down the number of participating first grade classrooms and students within each school, revealing the nested structure of the data, as well as

Table 2

Student demographic information: Present Study vs. WaKIDS WA State 2012-2013

Demographic	Present Study	WA State (2012-13)
	(N = 64)	*(N = 21,811)*
Gender		
Female	48.4%	48.5%
Male	51.6%	51.5%
Race/Ethnicity		
Asian	25.0%	4.7%
Black/African American	17.2%	6.9%
Hispanic	20.3%	38.4%
Native Hawaiian	1.6%	1.2%
Two or more races	15.6%	5.7%
White	20.3%	34.2%
Amer.Indian/Alaskan Native	0.0%	1.8%
Not provided	0.0%	7.1%
Free/Reduced Lunch (Income Status)		
No FRL (0)	25%	31.1%
FRL (1)	75%	68.9%
English Language Status		
English First Language (0)	56.3%	69.7%
Other First Language (1)	43.8%	30.3%
Special Education (IEP)		
No IEP	93.7%	91.7%
IEP	6.3%	8.3%
Age (1st grade)		
Mean	7.4 years	N/A
Range	6.8 - 8.4 years	N/A

Table 3

Frequencies: First Grade 2014 and Kindergarten 2012 Students, grouped by School and Teacher.

	School ID	Teacher ID	Number of Students
First Grade			
	1	1	7
		2	8
		3	6
	2	4	3
		5	1
	3	6	6
		7	2
		8	5
	4	9	1
		10	1
	5	11	12
		12	6
		13	5
	6	14	1
Kindergarten			
	1	1	4
		2	12
		3	5
	2	4	3
		5	1
	3	6	4
		7	4
		8	4
	4	9	2
	5	10	8
		11	7
		12	6
	6	13	1
	7	14	1
	8	15	1
	9	16	1

Note. *N = 64* for both grades. School and Teacher IDs do not represent the same schools and teachers at between grade levels.

the number of kindergarten schools and classrooms from which the first grade students' WaKIDS data originated.

Measures

The measures reviewed in this section are the sources for the predictive and outcome variables of this study. The first measure presented was administered in the fall of 2012: *WaKIDS GOLD* (a modified version of *Teaching Strategies* GOLD®). Student data from this assessment were retrieved from a state database shared by OSPI, only after district, school, and parent consent was obtained. The other two instruments presented here served as the outcome measures for this study and were administered in the Spring of 2014 to participating first grade students: *Test of Early Reading Ability-3* (*TERA-3*) and *Woodcock Johnson III Tests of Achievement* (*WJ-III ACH*).

Teaching Strategies GOLD®/ WaKIDS GOLD. *Teaching Strategies (TS)* GOLD® is a whole-child, observation-based assessment designed for use by teachers of children from birth to kindergarten. It includes a total of ten domains of development and learning, which can be broken down into 38 different objectives. Many of these objectives include two to five dimensions guiding teachers to observe and then score children on separate skills within the objective (e.g., Objective 16 is "Demonstrates knowledge of the alphabet," and is composed of two dimensions, each of which receive a separate score: "16a: Identifies and names letters," and "16b: Uses letter sound knowledge.") (Teaching Strategies, 2013). For every dimension, or objective without dimensions, *TS* GOLD® provides a ten-point scale for rating child skills across a particular progression of development and learning, represented by a "colored band" system marking "widely held expectations" for a range of chronological ages and grades (Teaching Strategies, 2011b, p. 3). Specifically, six chronological age levels make up a progression of the

particular skill being measured by an objective/dimension, and each level is represented by a different colored band (kindergarten is purple, at the end of the spectrum) to help guide teachers in understanding the expectations for the age range being assessed (see Table 3).

Table 4

WaKIDS Color Bands.

Color	Age Range
Red	Birth to 1 year
Orange	1 to 2 years
Yellow	2 to 3 years
Green	3 to 4 years
Blue	4 to 5 years
Purple	Kindergarten

Some colored bands of an objective/dimension are longer or shorter than others because the expected progression of development and learning varies from age to age. Using indicators and examples provided by *TS* GOLD® with each progression, teachers can observe student knowledge, skills, and behaviors for each objective/dimension, document observations in student portfolios over a period of time, and then use the collected information to assign a value between 0 (“not yet”) and 9 (“child exceeds kindergarten level expectations”) to a student’s level on a particular progression. Each progression has “in-between boxes,” assigned to score values in between those with assigned indicators and examples. These provide a way for teachers to “document any skills that are emerging but not yet fully developed” (Teaching Strategies, 2013, p. 8). An example of a dimension’s progression with colored bands is included in Appendix A.

TS GOLD® is designed for use with all children, including English and dual-language learners, children with disabilities, and children who demonstrate competencies beyond typical developmental expectations (Lambert, Kim, & Burts, 2014; Teaching Strategies, 2013). Teaching Strategies has established the complete GOLD® assessment's reliability and validity with a large nation-wide norm sample ranging the entire age-span included in *TS* GOLD®. In their inter-rater reliability study, domain-level ratings of students by a master trainer were compared to ratings of the same students assigned by teachers new to *TS* GOLD®. The resulting correlations supported reliability with all correlation coefficients greater than .90, with the exception of one at .80. Teaching Strategies also examined GOLD's® internal consistency and revealed reliability estimates with a mean of .97 (Teaching Strategies, 2011b).

Construct validity was determined with a six-factor model corresponding with six main domains of the instrument (Social-emotional, Physical, Language, Cognitive, Literacy, and Mathematics). The results provided strong evidence for construct validity with a Comparative Fit Index (CFI) = .931, a Root Mean Square Error of Approximation (RMSEA) = .066, and a Standardized Root Mean Square Residual (SRMR) = .033; all of these analyses were statistically significant at $p < .001$ (Teaching Strategies, 2011b). In a more recent review of *TS* GOLD's® psychometric properties, the authors presented additional reliability and validity evidence, leading them to conclude that, "the measure is psychometrically sound, culturally and linguistically responsive, and sensitive to children with disabilities" (Burts & Kim, 2014, p. 131).

Washington State worked with Teaching Strategies to develop a tailored version of *TS* GOLD® for their assessment component of *WaKIDS*, only requiring teachers to focus on the following six domains: Social-emotional, Physical, Language, Cognitive, Literacy, and Mathematics. Within these six domains, only 19 of the possible 38 objectives, and 31 of the

possible 39 dimensions within these objectives, were selected by Washington State for *WaKIDS*. All objective and dimension numbers for this modified version correspond with the numbers used by the complete *TS* GOLD® assessment, and training and administration follow the same procedures as the complete *TS* GOLD®, just focused on the selected objectives and dimensions (Soderberg, Stull, Cummings, Nolen, McCutchen, & Joseph, 2013). The full list of domains, objectives, and dimensions for the *WaKIDS GOLD* version can be found in Appendix B.

In the fall of 2012, participation in WaKIDS was required by all state-funded full-day kindergarten classrooms, and teachers in these classrooms were required to attend a two-day *WaKIDS GOLD* training offered by the State during the summer prior to using WaKIDS. Teachers were also offered the option of completing Teaching Strategies' Inter-rater Reliability Certification. All WaKIDS teachers were instructed to begin collecting student data at the beginning of the school year and were given until the end of their seventh complete week of the school year to finish collecting and entering data in the WaKIDS online tool. After data were entered and finalized in the online system, reports on all students in the class became available to the teacher, school principal, and other district administrators online; these reports included a class "snapshot" displaying student developmental levels across the class as well as individual student reports for parents. While teachers were only required to collect and enter data using the *WaKIDS GOLD* during this beginning of the year period, they were offered the option of using the *WaKIDS GOLD* up to two additional times during the remainder of the school year (*TS* GOLD® is designed to be administered in the fall, winter, and spring of each school year), and some school districts in the state required this of their teachers (Soderberg et al., 2013).

Although Teaching Strategies has established reliability and validity of their assessment as described above, the findings are supportive of the complete instrument, with all 38 objectives

included, and cannot be generalized to the abbreviated WaKIDS version. Therefore, a research team from the University of Washington was contracted by Washington's Office of Superintendent of Public Instruction (OSPI) to conduct inter-rater reliability and concurrent validity studies with WaKIDS teachers and students across the state during the fall of 2012. In the inter-rater reliability study, teachers across the state scored the same students as a master-trainer, similar to the *TS* GOLD's® reliability study, and analyses comparing the scores revealed that the *WaKIDS GOLD* has moderate inter-rater reliability, with certain domains (particularly cognitive) having lower levels of agreement with the master code than others, and with discrepancies also varying by the student scored. For concurrent validity, student WaKIDS scores from across the state were compared with scores on standardized assessments aligned with the skills measured by each of the domains. Results varied again by domain, with Literacy showing the strongest levels of concurrent validity, followed by Math and Language, providing evidence of concurrent validity for at least these domains (Soderberg et al., 2013).

Test of Early Reading Ability-3 (TERA-3). The *TERA-3* (Reid, Hresko, & Hammill, 2001) is a test of early reading skills that can be individually administered to children between the ages of three years, six months and eight years, six months. The *TERA-3* has two parallel forms, which both have three subtests, each one designed to measure a separate type of reading skill: (1) alphabet knowledge, (2) print conventions, and (3) meaning. Participating students completed all three subtests as part of the first grade battery. The assessment provides scores for each subtest as well as a single, standardized composite score, referred to as a "Reading Quotient" (M = 100, SD = 15). There are basals and ceilings which are both defined as three consecutive correct or incorrect items. The examiner records scores as the child progresses through each item, with dichotomous scoring (0, 1) for all subtests. Raw scores from each subtest

can be converted into age-based standard scores as well as percentiles and age and grade equivalents. The norming sample included 875 children across 22 states, determined to be representative of many demographic characteristics of the 2000 school-age population (Rathvon, 2004).

Reliability evidence is available for the *TERA-3*, with alternate-form reliability estimates for the three subtests ranging from .82 to .92 across all age groups, and coefficient alphas for eight subgroups ranging from .91 to .99 for the three subtests and composite, for both forms. Validity evidence is also available: correlations between the older version, the *TERA-2*, and the *TERA-3* ranged from .85 to .98, with the lowest correlations yielded from the Meaning subtests; confirmatory factor analysis results supported each subtest measuring a separate component of early reading ability, while high intercorrelations were still revealed among the subtests, supporting that all three together measure a unitary construct (Rathvon, 2004; Reid et al., 2001).

The *TERA-3* was one of two standardized literacy measures used in the concurrent validity study of the *WaKIDS GOLD*, serving as a comparison for the WaKIDS Literacy domain. It was chosen largely for its ability to assess kindergarten students' print awareness; there are few standardized print awareness measures available to educators and researchers, and the *TERA-3* is noteworthy in that it requires students to read as part of the assessment (Farrall, 2012). The concurrent validity study's results revealed that WaKIDS Literacy scores were a positive, statistically-significant predictor of the TERA-3 scores. I chose the *TERA-3* as one of my study's outcome measures for these reasons of print awareness measurement and previous findings, as well as because of its assessment of a wide variety of early literacy skills, with its three subtests aligning with Objectives 16 (Demonstrates knowledge of the alphabet), 17 (Demonstrates knowledge of print and its uses), and 18 (Comprehends and responds to books and other texts) of

WaKIDS GOLD. Furthermore, the *TERA-3* has been used as an outcome measure in other validity studies of observational assessments (e.g., Meisels et al., 2008; Sekino & Fantuzzo, 2005).

Woodcock Johnson III Tests of Achievement (WJ-III ACH). The *WJ-III ACH* (Woodcock, McGrew, & Mather, 2001) is a comprehensive assessment of academic achievement composed of two batteries, *Standard* and *Extended*, with several subtests in each measuring various "narrow abilities." The assessment can be administered in full for a comprehensive assessment of achievement, or individual subtests can be selected when focusing on assessing specific abilities, and it can be used with subjects from preschool to geriatric levels (McGrew, Schrank, & Woodcock, 2007). There are two parallel forms available that are matched for content, and norms provided are both age-based (two to 90+ years) and grade-based (kindergarten through graduate school). (Rathvon, 2004). A normative update was published in 2007 and provides a recalculation of the normative data based on statistics from the 2005 U.S. Census (McGrew et al., 2007).

Two subtests were used from the *WJ-III ACH* with the first grade students, one measuring literacy skills and the other measuring math skills: Letter-Word Identification (ID) and Applied Problems. The Letter-Word ID subtest is part of the *WJ-III ACH*'s Standard Battery and tests a child's ability to identify isolated letters and read real words, measuring reading decoding skill (McGrew et al., 2007; Rathvon, 2004). This subtest aligns with Objectives 15 (Demonstrates phonological awareness) and 16 (Demonstrates knowledge of the alphabet) of *WaKIDS GOLD*. Letter-Word ID has a split-half reliability coefficient of .98 for age six and .97 for age seven (McGrew et al., 2007). There is evidence for criterion-related validity for this subtest as well, with moderate to high correlations between scores from students in first through

eighth grade on the *WJ-III ACH* reading clusters (which the Letter-Word ID falls under) and reading scores on two other prominently-used achievement assessments (coefficient values ranged from .44 to .82) (Rathvon, 2004).

Because phonological skill, particularly decoding, is typically a main focus of first grade reading development (Snow et al., 1998), and because phonological skill is not well-represented in the *TERA-3* (Farrall, 2012), I decided to use the Letter-Word ID subtest in addition to the *TERA-3* so that a more complete and developmentally appropriate picture of student literacy achievement at the end of first grade could be obtained. Additionally, the Letter-Word ID subtest is widely used in the early literacy research, has "psychometric soundness," and provides a range of score types, including nationally normed and standardized scores (M = 100, SD = 15) (Rathvon, 2004). It has been commonly used (as well as an earlier version in the *WRMT-R*) as an outcome measure in predictive validity studies seeking to predict literacy achievement (Bridges & Catts , 2011; Can et al., 2013; Denton et al., 2006; Duncan & Rafter, 2005; Linklater et al., 2009; MacDonald et al., 2013; Oslund et al., 2012; Smolkowski & Gunn, 2012; Welsh et al., 2010).

The other subtest from the *WJ-III ACH* that was used was the Applied Problems subtest, which is a part of the Standard Battery and tests a child's ability to analyze and solve orally presented mathematical problems, measuring math achievement, math knowledge, and quantitative reasoning (McGrew et al., 2007; Rathvon, 2004). This subtest served as the mathematics measure in the WaKIDS concurrent validity study and aligns with the objectives and dimensions of the Mathematics Domain. The Applied Problems subtest has a split-half reliability coefficient of .88 for age six and .91 for age seven (McGrew et al., 2007). Like the Letter-Word ID subtest, Applied Problems also yields nationally normed and standardized scores

(M = 100, SD = 15) and has been widely used in educational research, including validity studies predicting math ability (Downer et al., 2012; Jordan et al., 2007; Welsh et al., 2010). While this study's focus is on the predictive validity of the WaKIDS Literacy Domain, this short measure of math skill was used for examining discriminant validity as part of the predictive validity analysis.

Procedure

Recruitment and Sampling. After receiving approval from the school district to recruit and work with students in their schools, I began the recruitment process by emailing principals of schools that participated in WaKIDS in the fall of 2012, asking to contact their first grade teachers about recruiting and assessing their students for my study. Once principals granted permission, I emailed first grade teachers of those schools to ask if they would send consent letters home to parents/guardians of their students, collect the letters upon return, and schedule times with me for assessing the consented students. The parent consent letter was available in both English and Spanish; both versions can be found in the appendices.

As discussed in the Participants section, a total of 64 eligible students were granted parent permission, and all 64 were selected to participate and receive the assessments. I had originally hoped to recruit at least 100 students from which to sample, but due to low response rates from schools and parents, use of a convenience sample was necessary.

Data collection. Prior to school visits, I recruited "researchers" to assist with administration of the assessments. These additional researchers were undergraduate students studying in the field of early childhood and had previous experience working with children. All researchers were trained on all three assessments of the first grade battery, as well as day-of-visit protocols, including establishment of student rapport and verbal assent. On the scheduled days, at

least one other researcher and I met participating first grade teachers in their classrooms and asked consented students, one at a time, to accompany us for the assessments. Each researcher took her assigned student to a quiet, distraction-free space outside of the classroom to administer the assessments. Before beginning the first assessment, the researcher explained a little about the study and what the student would be asked to do during their time working together, as part of the verbal assent process. Students had the option to verbally decline participation if they did not want to continue. The assessments took approximately 30 minutes to complete with each student. When finished, students were praised for their efforts and returned to their classrooms.

Once all students had been completely assessed, I provided teachers with detailed reports of each participating student's performance on the assessments as a thank you for their assistance. These reports provided teachers with a table of each student's raw and standard scores for the assessments and subtests, as well as age-equivalents, grade-equivalents, percentile ranks, and a "description" category aligning with the student's results (e.g., *advanced*, *average*). A teacher-friendly guide to interpreting the scores was also provided. These reports and guides were self-developed using information from the assessment manuals, and a de-identified example is included in Appendix C.

Data Analysis

After data collection, statistical analyses were needed to interpret the data and address this study's research questions. Planning the analyses required consideration of all possible variables as well as selection of the most appropriate analyses for each research question.

Variables. Student WaKIDS scores across all domains were accessed from a database maintained by OSPI. Individual dimension scores within the domains were available, as well as composite scores for each domain. For these domain scores, OSPI summed the scores of all

individual dimensions within that domain to create a composite variable. Each student therefore had a composite score for each domain, in raw and standard form. The Literacy Domain score was the most important for analysis, as it was the variable for which predictive ability was being measured, but the Math domain composite scores were obtained as well for use in examining discriminant validity. The Literacy Domain contains a total of 11 dimensions, and therefore, the composite raw scores could range from 0 to 99. The raw composite scores for all domains were converted to standard scores by OSPI. For the purposes of this study, I converted these standard scores into z-scores to serve as the primary **predictor variable**.

Other predictor variables were created from combinations of the Literacy dimension scores to align with the five objectives of the Literacy domain. The five objectives, phonological awareness, alphabet knowledge, print concepts, reading comprehension, and emergent writing, were chosen to define distinct "literacy skill variables" for the purposes of research question 1b. Composite variables were created from raw scores means of the dimensions within each objective. Each objective and its dimensions are presented in Table 4, aligned with the first grade literacy outcomes measuring the corresponding skill. Note that the print concepts variable is not actually a composite but one dimension on its own; this is because the *WaKIDS GOLD* only selected one of the dimensions from the *TS* GOLD's® print concepts objective.

Table 5

WaKIDS Literacy objectives and dimensions aligned with First Grade outcomes.

WaKIDS Objective	First Grade Outcomes
15. Phonological Awareness (15a. Rhyme, 15b. Alliteration, 15c. Units of Sound)	Letter-Word ID
16. Alphabet Knowledge (16a. Letter ID, 16b. Letter-Sound Knowledge)	TERA Alphabet
17b. Uses Print Concepts	TERA Conventions
18. Reading Comprehension (18a. Interacts Read-Alouds, 18b. Emergent Reading, 18c. Retells Stories)	TERA Meaning
19. Emergent Writing (19a. Writes Name, 19b. Writes Meaning)	All (no directly aligned measure)

Note. *N = 64*. WaKIDS dimensions under each objective in parentheses; 17b. Print Concepts is a WaKIDS dimension, there is no objective.

Measurement of literacy achievement at the end of first grade was needed for the study's primary **outcome, or criterion, variable**. The *TERA-3* and *WJ III Letter-Word ID* assessments each measured different components of the literacy skills that define literacy achievement at this age and grade level in school, as previously discussed. Therefore, the scores from the *TERA-3* and *Letter-Word ID* assessments were combined into a composite variable, using the mean of their z-scores, to provide one end-of-first-grade literacy achievement score for each student. The TERA-3's Reading Quotient was used to represent student literacy skills concerning print awareness and oral language, or more specifically, WaKIDS dimensions measuring alphabet knowledge (16a, b), print knowledge (17b), and comprehension and responding to books and other texts (18a-c). Standard scores from the Letter-Word ID were used to represent skills

concerning phonological awareness, or more specifically, WaKIDS dimensions measuring rhyming, alliteration, and phonemic skill (15a-c) and use of letter-sound knowledge (16b). WaKIDS Objective 19, *Demonstrates emergent writing skills*, could not be directly represented by scores from either of these outcome measures because neither directly assessed students' writing ability. However, as mentioned in Chapter 2's conceptual framework/literature review, WaKIDS dimension 19a (*writes name*) measures a skill related to the early literacy category of print awareness, and 19b (*writes to convey meaning*) measures skills related to all three early literacy categories (phonological awareness, print awareness, and oral language). Therefore, these two emergent writing dimensions were still included in analyses to determine their role in the predictive relationship. The literacy achievement composite variable therefore represented measurement of all literacy skills assessed in the WaKIDS Literacy Domain at some level.

As previously explained, the *WJ-III Applied Problems* first grade student scores were obtained as an additional outcome variable for the purpose of supporting the WaKIDS Literacy Domain's validity with evidence of discriminant validity (research sub-question 1a). The Applied Problems subtest measured math skills that align with most of the WaKIDS Mathematics Domain: dimensions measuring use of number concepts and operations (20a-c) and comparing and measuring skills (22). The standard scores from this subtest were the scores selected to measure end of first grade math achievement.

Student demographics, as well as teacher (kindergarten and first grade) served as **control variables** in analyses for the purpose of determining whether the relationship between WaKIDS Literacy and first grade literacy scores would differ for particular types of students, or for students with particular teachers. The three demographic variables included in analyses, student race/ethnicity, first language, and income level, were specifically selected because of the

evidence that students of these backgrounds are at greater risk for reading difficulties (McCoach et al., 2006; Slavin & Cheung, 2005; Snow et al., 1998; Teale, Paciga, & Hoffman, 2007). These variables were recoded as dichotomous variables to represent more meaningful categories: race/ethnicity was converted to *minority status* with students categorized as white assigned a "0" and students of all other race/ethnicity categories assigned a "1," while first language was converted to *English language status*, having students with English as their first language assigned a "0," and all other language categories assigned a "1." Student income level was determined by their free and/or reduced lunch (FRL) status (yes/no) and therefore was already dichotomous – FRL students were assigned a "1" and non-FRL students a "0."

Student's kindergarten teacher and first grade teacher were nested variables that needed to be considered as well to determine whether students with the same teachers had more similar scores than students with different teachers. Students were assigned teacher identification numbers for both years so that they could be partitioned into their appropriate classroom groupings in the analyses used to account for this nested data.

Selected analyses. The statistical analyses that would use these variables were selected according to the research question being assessed. The majority of predictive validity studies reviewed in Chapter 2 used various types of regression and multi-level analyses to determine the strength of the relationship between predictor and criterion variables, as well as the existence of a uniquely predictive relationship. Likewise, this study primarily used hierarchical linear regressions and hierarchical linear modeling (HLM) (Raudenbush & Bryk, 2002) to answer its research questions. First, zero-order correlations between predictor, control, and outcome variables were calculated, followed by a simple linear regression to take an initial look at the predictive relationship between the WaKIDS Literacy domain and first grade literacy

achievement. For research questions 1a and 2, unconditional and full models using HLM were selected because of the nested structure of the data. Research question 1a's HLM examined the predictive relationship between WaKIDS Literacy (predictor at level 1) and first grade literacy (outcome), while controlling for the demographic variables at level 1, as well as teacher (both kindergarten and first grade) at level 2. This allowed for the predictor and control variables to explain variance in the first grade literacy scores, both separately and together, independent of the effect of students grouped by teacher.

For the discriminant validity analysis in research question 2, the HLM allowed for the predictive relationship between WaKIDS Literacy and first grade math (Applied Problems) to be examined, independent of the demographic and teacher variables once again. Applied Problems served as the outcome to verify that WaKIDS Literacy was not predictive of first grade math skill. WaKIDS Literacy and demographic variables were still nested within teacher as before, so an HLM was considered the most appropriate analysis.

In research question 1b, the predictive validity of the individual WaKIDS objective scores were examined with individual first grade literacy measures serving as outcome variables. Hierarchical linear regressions were used to determine the unique predictive effects of the WaKIDS objective variables on the corresponding first grade measures (see Table 4), while controlling for all other objective variables as well as demographic variables. Four regression analyses were conducted, each with demographic variables only entered into the first block, and demographics along with the WaKIDS objective variables entered into the second. Each regression analysis used one of the four first grade subtest measures as the outcome: TERA Alphabet subtest, TERA Conventions subtest, TERA Meaning subtest, and *WJ III* Letter-Word ID subtest.

For the HLM analyses, a more liberal significance level ($p < .10$, instead of .05) was used for all tests of significance. Because HLM is used to account for the random ness in coefficients, Nezlek (2008) stated that the model should reflect this randomness "to the extent that it is possible," (p. 851). He therefore recommended using the significance level of .10, which is typically used for multi-level models.

CHAPTER 4

Results

Preliminary Analyses

Descriptives. As explained in the previous chapter, kindergarten and first grade data of 64 students were attained for several statistical analyses to address this study's research questions. Descriptive statistics for all kindergarten and first grade measures are reported in Table 6. Means, standard deviations, and the range of standard scores for the student sample are given when possible; only raw scores were available for WaKIDS dimension scores. Because the first grade literacy composite was formed by combining z-scores, its descriptives were not included in this table.

Looking at the individual dimension scores, it is important to note that the ranges of scores, while typical for these dimensions across the state, are nearly as large as possible for some (the score range is 0 to 9), indicating that for many of these dimensions, there were students in this sample scoring as low as the 0-1 year age level, and as high as the kindergarten level. Interestingly, the score range for dimension 19a (writes name) was the smallest with scores ranging from 3 to 7.

Regarding the first grade statistics, this sample's mean TERA-3 Reading Quotient was slightly lower than that of the national norms, but within one standard deviation (M = 100, SD = 15). On the other hand, the sample's means for both *WJ III* subtests were slightly higher than the reported national norms (M = 100, SD = 15), with the Letter-Word ID mean nearly two thirds of a standard deviation above. For the raw scores of the TERA-3 subtests, the Meaning subtest had the greatest dispersion of scores (SD = 6.23), while the Alphabet subtest had the least dispersion (SD = 2.42).

Table 6

Means and Standard Deviations: WaKIDS Literacy scores and First Grade Literacy scores

	M	*(SD)*	*n*
WaKIDS Lit[a]	664.78	(58.09)	64
PA(15a)	4.73	(1.95)	64
PA(15b)	5.06	(1.82)	64
PA(15c)	4.64	(1.86)	64
ABC(16a)	5.44	(2.20)	64
ABC(16b)	3.80	(2.40)	64
PC(17b)	5.23	(1.55)	64
RC(18a)	4.75	(1.66)	64
RC(18b)	5.23	(1.53)	64
RC(18c)	5.14	(1.69)	64
W(19a)	5.67	(0.69)	64
W(19b)	4.25	(1.16)	64
TERA RQ[a]	94.06	(15.11)	64
TERA Alphabet	24.41	(2.42)	64
TERA Conventions	15.09	(3.18)	64
TERA Meaning	16.34	(6.23)	64
Letter-Word ID[b]	109.03	(14.19)	64
Applied Problems	105.05	(15.49)	64

Note. [a]Standard Score (M = 100, SD = 15); [b]Raw Score; Word ID = WRMT-R word identification; Word Attack = WRMT-R word attack (nonword reading); Morph = Siegel morphological assessment task.

Research Question 1a

Do kindergarten student scores from the WaKIDS Literacy Domain have a statistically significant predictive relationship with scores of student literacy achievement at the end of first grade? If so, is the relationship strong enough to provide evidence of predictive validity? It was hypothesized that WaKIDS Literacy Domain scores would positively and uniquely predict the end of first grade literacy achievement scores, while controlling for student demographics and teacher effects.

Zero-order correlations between the predictor, outcome, and control variables were first examined to gain an initial understanding of the relationship. The correlation matrix provided in Table 7 displays the resulting correlation coefficients, with all significant at either the $p < .05$ level or $p < .01$ level. The main correlation of interest for this question, that between the WaKIDS and first grade composite variables, was strong and positive as anticipated ($r = .72$, $p < .01$). Also worth noting is the strong correlation between the two literacy outcome assessment scores, TERA and Letter-Word ID ($r = .83$, $p < .01$), which confirmed their measurement of related literacy constructs (alphabet knowledge, print conventions, meaning, and reading decoding skill).

Table 7

Zero-order correlations: Predictor, outcome, and control variables.

	1.	2.	3.	4.	5.	6.	7.	8.
1. WaKIDS Lit	--							
2. First Grade Lit	0.72**	--						
3. TERA RQ	0.68**	0.95**	--					
4. Letter-Word ID	0.58**	0.93**	0.83**	--				
5. Applied Problems	0.58**	0.73**	0.72**	0.73**	--			
6. FRL	-0.43**	-0.46**	-0.40**	-0.43**	-0.53**	--		
7. Minority	-0.21	-0.31*	-0.26*	-0.31*	-0.42**	0.70**	--	
8. Language	-0.30*	-0.24	-0.25*	-0.15	-0.18	0.36**	0.29*	--

Note. N = 64. WaKIDS Lit = WaKIDS Literacy Composite (standard); First Grade Lit = First Grade Literacy Composite (z-score mean of TERA RQ & Letter-Word ID); TERA RQ = TERA Reading Quotient; FRL = Free/Reduced Lunch Status (student income). *$p < .05$, **$p < .01$

A simple linear regression revealed that WaKIDS Literacy accounted for a significant amount of variance (approximately 51%) in first grade literacy as predicted, $t(62) = 8.10$, $p < .001$, $R^2 = 0.51$. Specifically, for every one point increase on the WaKIDS Literacy composite score, there was an estimated mean increase of .01 points on the first grade literacy composite. This implied that the higher a student's WaKIDS Literacy score, the higher their first grade literacy skill. While this finding supported the domain's predictive validity, the simple linear regression did not account for the presence of other variables or the nested structure of the data. Therefore, more stringent analyses were needed to provide the most accurate model of this relationship.

Because student-level variables were nested within kindergarten and first grade teachers, hierarchical linear modeling (Raudenbush & Bryk, 2002) was the most appropriate type of analysis. Use of a multi-level model allowed for the relationship between WaKIDS Literacy and first grade literacy scores to be more precisely examined at both the student and teacher level. As discussed in the Method chapter, student level control variables included student income,

minority, and language status, and teacher level control variables included kindergarten teacher and first grade teacher.

Prior to conducting the full hierarchical models, unconditional models were used to examine the total proportion of variance in the outcome variable (first grade literacy) existing between teachers (either kindergarten or first grade), in the absence of predictor variables. It was important to calculate and then examine the intraclass correlation (ICC) coefficients as well as tests of significance to confirm that first grade literacy scores did indeed vary by teacher, justifying the need for hierarchical linear modeling (HLM).

$$Y_{(\text{First Gr Lit})} = \gamma_{00} + u_0 + r. \text{ (unconditional)}$$

When the variance in first grade literacy was examined between kindergarten teachers, the resulting ICC coefficient was .04, indicating that only 4% of the variance in first grade literacy scores was accounted for by kindergarten teacher, while 96% remained at the student level. The unconditional model of these variables revealed that this small proportion of variance did not carry statistical significance, suggesting that there might not be enough variance between kindergarten teachers to continue with HLM.

The unconditional model exploring the variance in first grade literacy by first grade teacher resulted in an ICC of 0.11, indicating that 11% of the variance in first grade literacy scores was accounted for by students' first grade teacher, and 89% was at the student level. This proportion of variance nearly reached significance with $x^2(13) = 21.31$, $p = .067$. While still small, this significant proportion of variance had greater potential than that of kindergarten teacher to have an effect on the relationship between WaKIDS Literacy and first grade literacy. Therefore, first grade teacher was selected as the group level variable for the HLM analysis. This seemed like a logical choice because of the more recent experience students had with this teacher

at the time of assessment, along with the substantial amount of reading instruction known to occur during the first grade year (Snow et al., 1998).

A full model assessed the overall predictive relationship between WaKIDS Literacy and first grade literacy achievement with inclusion of all student-level predictors: WaKIDS Literacy Domain composite scores (group-mean centered) and student demographic variables (income, language, and minority status), as well as the teacher-level variable, first grade teacher. This resulted in the following model:

$$Y_{(\text{First Gr Lit})} = \gamma_{00} + \gamma_{10(\text{WaKIDS L})} + \gamma_{20(\text{FRL})} + \gamma_{30\ (\text{minority})} + \gamma_{40\ (\text{language})} + u_0 + r. \text{ (full)}$$

The model estimate of the intercept showed that the mean estimate of the first grade literacy scores was .45 points (*SE* = .23) which was significantly greater than zero, due to the increased significance level, $t(13) = 1.94$, $p < .10$. The predictors together accounted for 40.5% of the total variance in first grade literacy. WaKIDS Literacy was revealed to be a unique, positive predictor of first grade literacy while the demographic variables were held constant, $b = .01$ (*SE* = .002), $t_{(46)} = 4.81$, $p < .001$. Specifically, there was an estimated increase of .01 standard deviations in first grade literacy scores for every one point increase on the WaKIDS Literacy scores; or put another way, a 1.0 standard deviation increase for every 100 point increase in WaKIDS scores. Although this small increase was in part due to the first grade literacy scores being in z-score format (because it was the composite of TERA and Letter-Word ID scores), its size still indicates a small improvement in scores from kindergarten to first grade that, while still significant, should be noted. Student income status was a significant negative predictor of first grade literacy, with students receiving free and/or reduced lunch (lower-income students) tending to have lower first grade literacy scores, $b = -.58$ (*SE* = .31), $t_{(46)} = -1.88$, $p < .10$ (see Table 8). The other two demographic variables did not uniquely predict the first grade

outcome on their own. When the predictive relationship was analyzed earlier with a simple linear regression, which does not consider the data's nested structure, WaKIDS Literacy was found to account for 51% of the variance in first grade literacy. While the HLM resulted in a slightly reduced proportion of variance accounted for by predictors, the significant relationship between WaKIDS and first grade literacy remained, thereby confirming that use of HLM with multiple predictors helped provide more precise results.

Table 8

Two-level model of First Grade Literacy.

	Unconditional Model					Full Model				
Fixed Effect	*Coeff*	*SE*	*t*	*df*	*p*	*Coeff*	*SE*	*t*	*df*	*p*
Intercept	-0.02	0.15	-0.15	13	>.05	0.45	0.23	1.94	13	0.08*
WaKIDS Literacy						0.01	0.002	4.81	46	< .001***
FRL						-0.58	0.31	-1.88	46	0.06*
Minority						-0.10	0.32	-0.34	46	>.05
Language						0.10	0.19	0.50	46	>.05
Random Effect	*Var*		*chi*	*df*	*p*	*Var*		*chi*	*df*	*p*
Intercept	0.10		21.31	13	0.07*	0.09		24.69	13	0.03*
Level 1	0.80					0.48				

Note. N = 64 . WaKIDS Literacy = WaKIDS Literacy Composite (standard), entered into model group mean centered;
FRL = Free/Reduced Lunch Status (student income); FRL, Minority, & Language entered into model mean uncentered; *p < .10, ***p < .001

The variance between teachers was significantly different from zero as well, indicating that the strength of the relationship between WaKIDS and first grade literacy significantly varied across first grade teachers. Specifically, first grade teacher accounted for 15.5% of the variance in WaKIDS Literacy's prediction of first grade literacy. Despite this significant teacher effect, WaKIDS Literacy scores were still uniquely predictive of first grade literacy achievement, providing strong support for this WaKIDS domain's predictive validity.

Research Question 1b

Are the individual literacy skills measured by objectives within the WaKIDS Literacy Domain predictive of corresponding literacy skills measured at the end-of-first grade?

Subsequently, are particular WaKIDS Literacy objectives stronger predictors of end-of-first-grade literacy achievement scores than others? How do the revealed relationships affect the Literacy Domain's overall predictive validity? While a multi-level analysis contributed favorable results to answer the first research question, these results only pertained to the overall WaKIDS Literacy domain and had not taken into account the predictive abilities of the individual WaKIDS Literacy objectives and dimensions. Therefore, the second half of this research question focused on the role of the individual WaKIDS Literacy objectives in the predictive relationship between WaKIDS and first grade literacy.

The correlation matrix in Table 9 provides the zero-order correlations between the different WaKIDS literacy skill variables (as defined by objective composite scores described in Chapter 3) and first grade literacy subtests. All resulting correlations between the WaKIDS and first grade variables were positive and statistically significant. The coefficients of the WaKIDS writing variable, as well as those of the TERA Alphabet subtest, were generally smaller in size when compared with the coefficients resulting from the other literacy skill variables and first grade measures.

Table 9
Zero-order correlations: WaKIDS Literacy Skill Variables and First Grade Outcomes.

	1	2	3	4	5	6	7	8	9
1. Phon Awareness	--								
2. ABC Knowledge	0.78**	--							
3. Print Concepts	0.77**	0.74**	--						
4. Reading Comp	0.80**	0.69**	0.74**	--					
5. Writing	0.50**	0.54**	0.55**	0.53**	--				
6. TERA A	0.34**	0.49**	0.39**	0.30*	0.26*	--			
7. TERA C	0.60**	0.68**	0.66**	0.55*	0.52**	0.55**	--		
8. TERA M	0.63**	0.60**	0.55**	0.62*	0.40**	0.55**	0.64**	--	
9. Letter-Word ID	0.58**	0.67**	0.62**	0.61*	0.43**	0.68**	0.69**	0.79**	--

Note. N = 64 . All raw scores. WaKIDS Literacy Skill Variables (#1-5) align with WaKIDS Literacy Objectives; TERA A = TERA Alphabet Subtest; TERA C = TERA Conventions Subtest; TERA M = TERA Meaning Subtest; *p < .05, **p < .01

Hierarchical linear regressions (in contrast to an HLM approach) were then conducted to determine the effect of each WaKIDS Literacy objective on the first grade outcome measures, as well as all together, while controlling for student demographics. Because of the strong predictive ability of the WaKIDS domain revealed in the previous HLM analysis, which persisted even when controlling for student's first grade teacher, simpler regression analyses were deemed appropriate for examining the predictive abilities of the individual WaKIDS literacy skill variables, with the nested structure no longer a major concern. Four hierarchical regressions were conducted, one for each of the four first grade outcome measures. For each of these analyses, demographic variables (FRL, minority, and language) were controlled for by entering all three together in the first block, followed by addition of all five literacy objectives together in the second block. All four analyses resulted in a significant change in R^2 with the addition of the WaKIDS variables to the second block.

Specifically, for the Letter-Word ID regression, the demographic predictors together in the first block accounted for 23% of the variance in Letter-Word ID scores, which was significant, $F(3, 60) = 5.83$, $p < .01$. Of the predictors, FRL was the only one to hold a significant unique prediction ($p < .01$). The addition of the WaKIDS objectives explained an additional 36% of the variation in Letter-Word ID and this change in R^2 was significant, $F(8, 55) = 8.78$, $p < .001$. The only predictor revealed to uniquely predict Letter-Word ID with all of the other predictors together in the model was alphabet knowledge ($p < .01$). In the next regression, the demographic predictors together accounted for 13% of the variance in TERA Alphabet scores, with significance, $F(3, 60) = 3.00$, $p < .05$; however, none of the three were able to uniquely predict on their own. When the WaKIDS objectives were added, an additional 23% of the variation in TERA Alphabet was explained, which was significant, $F(8, 55) = 3.93$, $p < .01$.

Again, alphabet knowledge emerged as the only unique predictor out of all of the predictors present ($p < .01$).

For the TERA Conventions regression, the demographic predictors explained 15% of the variance in the TERA Conventions scores, which was significant $F(3, 60) = 3.60$, $p < .05$, and FRL was the only unique predictor of the three ($p < .01$). The WaKIDS objective explained an additional 40% of the variance when added in the second block, a significant amount $F(8, 55) = 8.55$, $p < .001$. Alphabet knowledge was once again the only unique predictor in the second block ($p < .05$). Finally, for the fourth regression, the demographic predictors explained 15% of the variance in TERA Meaning scores, yet again a significant amount, $F(3, 60) = 3.42$, $p < .05$, although none of the predictors uniquely explained variance on their own. When added in the second block, the WaKIDS objectives accounted for a significant amount variance (34%) in the TERA Meaning scores, $F(8, 55) = 6.40$, $p < .001$. Unlike the other three regressions, however, none of the individual predictors in the second block, including alphabet knowledge, were found to uniquely explain any variance, although alphabet knowledge was close with a p-value of .08. These results are summarized in Tables 10-13.

Table 10

Summary of Hierarchical Regression Analysis for Student Demographics and WaKIDS Literacy Skills Predicting WJ III Letter-Word ID

Variable	*B*	*SE*	β	*t*	R^2
Model 1					0.23**
(Constant)	0.84	0.26		3.27 **	
Free/Reduced Lunch	-1.05	0.37	-0.46	-2.83 **	
Minority	0.01	0.39	0.01	0.03	
Language	-0.09	0.24	-0.05	-0.38	
Model 2					0.56***
(Constant)	0.56	0.23		2.49 *	
Free/Reduced Lunch	-0.39	0.31	-0.17	-1.25	
Minority	-0.33	0.32	-0.14	-1.05	
Language	0.04	0.20	0.02	0.19	
Phon Awareness	-0.06	0.20	-0.06	-0.32	
ABC Knowledge	0.49	0.16	0.45	3.04 **	
Print Concepts	0.19	0.16	0.20	1.20	
Reading Comp	0.07	0.21	0.06	0.34	
Writing	0.03	0.13	0.03	0.28	

Note. $N = 64$. Letter Word ID converted to z-scores. $R^2_{change} = 0.36$***

*p < .05, **p < .01, ***p < .001

Table 11

Summary of Hierarchical Regression Analysis for Student Demographics and WaKIDS Literacy Skills Predicting TERA Alphabet Subtest

Variable	*B*	*SE*	β	*t*	R^2
Model 1					0.13*
(Constant)	0.64	0.27		2.34 *	
Free/Reduced Lunch	-0.48	0.40	-0.19	-1.13	
Minority	-0.16	0.42	-0.06	-0.38	
Language	-0.39	0.26	-0.19	-1.48	
Model 2					0.36**
(Constant)	0.72	0.27		2.61 *	
Free/Reduced Lunch	-0.09	0.38	-0.04	-0.23	
Minority	-0.55	0.38	-0.22	-1.43	
Language	-0.47	0.25	-0.23	-1.88	
Phon Awareness	-0.08	0.24	-0.07	-0.33	
ABC Knowledge	0.62	0.20	0.56	3.15 **	
Print Concepts	0.26	0.20	0.26	1.31	
Reading Comp	-0.49	0.26	-0.40	-1.89	
Writing	-0.02	0.16	-0.02	-0.11	

Note. $N = 64$. TERA Alphabet converted to z-scores. $R^2_{change} = 0.23$**

*p < .05, **p < .01, ***p < .001

Table 12

Summary of Hierarchical Regression Analysis for Student Demographics and WaKIDS Literacy Skills Predicting TERA Conventions Subtest

Variable	*B*	*SE*	β	*t*	R^2
Model 1					0.15*
(Constant)	0.31	0.27		1.15	
Free/Reduced Lunch	-1.13	0.39	-0.49	-2.89 **	
Minority	0.80	0.41	0.32	1.96	
Language	-0.20	0.26	-0.10	-0.79	
Model 2					0.55***
(Constant)	0.04	0.23		0.17	
Free/Reduced Lunch	-0.40	0.32	-0.18	-1.26	
Minority	0.39	0.32	0.16	1.20	
Language	-0.07	0.21	-0.04	-0.34	
Phon Awareness	0.07	0.20	0.06	0.35	
ABC Knowledge	0.38	0.16	0.35	2.31 *	
Print Concepts	0.30	0.17	0.30	1.84	
Reading Comp	-0.11	0.22	-0.09	-0.49	
Writing	0.18	0.13	0.15	1.33	

Note. N = 64 . TERA Conventions converted to z-scores. $R^2_{change} = 0.40$***

*p < .05, **p < .01, ***p < .001

Table 13

Summary of Hierarchical Regression Analysis for Student Demographics and WaKIDS Literacy Skills Predicting TERA Meaning Subtest

Variable	*B*	*SE*	β	*t*	R^2
Model 1					0.15*
(Constant)	0.72	0.27		2.66 *	
Free/Reduced Lunch	-0.53	0.39	-0.23	-1.35	
Minority	-0.22	0.41	-0.09	-0.53	
Language	-0.31	0.26	-0.15	-1.20	
Model 2					0.48***
(Constant)	0.31	0.25		1.28	
Free/Reduced Lunch	0.20	0.34	0.09	0.58	
Minority	-0.51	0.35	-0.21	-1.47	
Language	-0.10	0.22	-0.05	-0.45	
Phon Awareness	0.31	0.21	0.28	1.48	
ABC Knowledge	0.32	0.18	0.29	1.81	
Print Concepts	-0.01	0.18	-0.01	-0.08	
Reading Comp	0.18	0.23	0.15	0.77	
Writing	0.03	0.14	0.02	0.20	

Note. N = 64 . TERA Meaning converted to z-scores. $R^2_{change} = 0.34$***
*p < .05, **p < .01, ***p < .001

With the addition of WaKIDS objectives explaining statistically significant amounts of variance, ranging between 23-40%, in all four of the first grade literacy outcomes, there is clear evidence that the objectives together were predictive of different types of literacy skills in first grade. However, when examined individually, these objectives were not able to predict any of these skills on their own, with the exception of alphabet knowledge. The alphabet knowledge objective uniquely predicted three of the four first grade outcomes: Letter-Word ID, TERA Alphabet, and TERA Conventions. The other four objectives did not contribute any unique

variance on their own, revealing that their contributions to the literacy outcomes were only within the shared variance explained by all of the objectives together.

Additionally, student income (FRL) was a unique, negative predictor of Letter-Word ID and TERA Conventions scores before addition of the WaKIDS variables, indicating that the FRL students tended to score lower on these measures than the higher-income students. However, this demographic predictor was not able to retain its unique prediction when the WaKIDS variables were added to both the Letter-Word ID and TERA models, revealing that the WaKIDS objectives were able to predict the first grade outcomes despite the effect of student income. Nevertheless, student income's unique prediction without the WaKIDS variables was not trivial and should still be considered with the other findings of these regression analyses.

Research Question 2

Is there evidence of discriminant validity to support this predictive relationship? Specifically, are WaKIDS Literacy scores not uniquely predictive of end of first grade achievement scores in a conceptually different domain of learning (i.e., mathematics)? The first research question sought to provide evidence of predictive validity by confirming that the WaKIDS Literacy domain was predictive of students' later performance of a skill that it should predict: literacy achievement at the end of first grade. As discussed in Chapter 2, this type of validity evidence is known as *convergent* validity. Confirming that a measure is *not* predictive of another measure that it is *not* meant to predict also supports the measure's predictive validity; hence this second research question sought to evaluate WaKIDS Literacy's *discriminant* validity.

To do this, a measure of a skill theoretically different from the literacy skill measured by WaKIDS was needed. The *WJ-III* Applied Problems subtest, which had also been administered at the end of first grade, provided measurement of student mathematic skill to serve as this

conceptually-different, *discriminant* variable. Hierarchical linear modeling was used again to test the relationship between WaKIDS Literacy domain scores and first grade math scores. HLM was chosen in order to account for the student scores nested within first grade teachers just as before, as well as to remain consistent in analysis of the WaKIDS *domain* scores, as opposed to objectives/dimensions. The student-level variables for the discriminant validity model included WaKIDS Literacy scores, as well as the three demographic variables (income, language, and minority). The teacher level was assessed by first grade teacher, and Applied Problems scores (first grade math) served as the outcome variable.

Zero-order correlations between the Applied Problems scores and predictor /control variables are provided alongside the other correlations in Table 6. A moderate, positive correlation resulted between WaKIDS Literacy and first grade Applied problems scores ($r = .58$, $p < .01$), revealing a stronger relationship than anticipated. The unconditional model, using only the teacher and outcome variables, was once again examined first to determine the proportion of variance in first grade math scores existing between first grade teachers. The resulting ICC coefficient of .001 and the lack of statistical significance, $x^2(13) = 10.37$, $p > .50$, revealed that first grade teacher had no effect on Applied Problems scores.

$$Y_{(\text{Applied Prob})} = \gamma_{00} + u_0 + r. \quad \text{(unconditional)}$$

Despite this lack of teacher effect, it was important to continue with the full model to assess the relationship between WaKIDS Literacy and first grade math because of the still-existing nested structure. The full model included all student and teacher variables, with the WaKIDS Literacy composite scores again centered around the group mean:

$$Y_{(\text{Applied Prob})} = \gamma_{00} + \gamma_{10(\text{WaKIDS L})} + \gamma_{20(\text{FRL})} + \gamma_{30\ (\text{minority})} + \gamma_{40\ (\text{language})} + u_0 + r. \text{ (full)}$$

The model estimate of the intercept showed that the mean estimate of the first grade math scores was .77 points (*SE* = .21) which was significantly greater than zero, $t(13) = 3.65$, $p < .01$ (see Table 13). Unexpectedly, WaKIDS Literacy was found to be a unique, positive predictor of first grade math with the demographic variables held constant. Specifically, there was an estimated increase of .01 standard deviations in first grade math scores for every one point increase on the WaKIDS Literacy scores, $b = .01$ (*SE* = .002), $t_{(46)} = 5.67$, $p < .001$. Again, this very small increase was in part due to the Applied Problems scores being in z-score format; they were converted to z-scores in order to be comparable to the HLM literacy analyses. While the relationship is still significant, there is small improvement in scores from kindergarten literacy to first grade math that should be recognized. Student income level was a unique, negative predictor of first grade math, holding all other predictors constant, with FRL students scoring an estimated -.60 standard deviations in first grade math lower than students not receiving FRL, $b = -.60$ (*SE* = -.51), $t_{(46)} = -1.98$, $p < .10$. Minority and language were once again not significantly predictive; however, the predictors all together accounted for 51.5% of the variance in the first grade math scores. Not surprisingly, because of the findings from the unconditional model, variance between first grade teachers was not significant, with teacher only accounting for 1.2% of the variance in WaKIDS Literacy's prediction of first grade math. This unexpected predictive relationship between WaKIDS Literacy and first grade math did not provide the evidence needed to support the discriminant validity of the WaKIDS Literacy domain.

Table 14

Two-level model of Applied Problems.

	Unconditional Model					Full Model				
Fixed Effect	*Coeff*	*SE*	*t*	*df*	*p*	*Coeff*	*SE*	*t*	*df*	*p*
Intercept	0.02	0.13	0.14	13	>.05	0.77	0.21	3.65	13	.003**
WaKIDS Literacy						0.01	0.002	5.67	46	< .001***
FRL						-0.60	0.30	-1.98	46	0.05*
Minority						-0.51	0.31	-1.64	46	>.05
Language						0.21	0.19	1.10	46	>.05
Random Effect	*Var*		*chi*	*df*	*p*	*Var*		*chi*	*df*	*p*
Intercept	0.001		10.37	13	>.05	0.02		12.23	13	>.05
Level 1	0.99					0.48				

Note. N = 64 . WaKIDS Literacy = WaKIDS Literacy Composite (standard), entered into model group mean centered; FRL = Free/Reduced Lunch Status (student income); FRL, Minority, & Language entered into model mean uncentered; *p < .10, **p < .01,***p < .001

Summary of Results

This results chapter provided evidence to support the WaKIDS Literacy domain's predictive validity using hierarchical linear modeling and linear regression to examine the strength of the predictive relationship between WaKIDS Literacy and end of first grade literacy achievement. In the first research question, an HLM analysis revealed that WaKIDS Literacy domain scores were uniquely and positively predictive of first grade literacy achievement, while controlling for student demographics and a significant teacher effect. This provided evidence of the domain's convergent validity. Student income level, as well as the nesting variable, first grade teacher, both also carried statistical significance. For student income, there was a negative relationship indicating that low income (free/reduced lunch status) was predictive of lower first grade scores, independent of the effects of WaKIDS Literacy scores and first grade teacher. The effect of first grade teacher revealed that a student's assigned teacher was also predictive of first grade scores, independent of the effects of any of the student-level predictor variables.

Hierarchical linear regressions were used to understand this relationship further by testing the abilities of the WaKIDS Literacy objectives to predict each of the first grade literacy subtests, while controlling for student demographics. The five WaKIDS objectives together explained additional variance in all four of the first grade literacy outcomes with statistical significance, with amounts ranging from 23-40%. At the individual level, alphabet knowledge was the only objective of the five able to hold its own unique prediction of Letter-Word ID, TERA Alphabet, and TERA Conventions. No individual objectives were uniquely predictive of the TERA Meaning subtest. Student income was a unique predictor of Letter-Word ID and TERA Conventions scores, but was not able to stand on its own as a unique predictor in the presence of the WaKIDS objectives.

For the second research question, the WaKIDS Literacy domain's discriminant validity was evaluated. First grade math skill served as the discriminant variable that WaKIDS Literacy was not meant to predict. HLM was used again to test this relationship in the presence of student demographic variables and first grade teacher. Surprisingly, WaKIDS Literacy uniquely and positively predicted first grade math skill, even with the other variables taken into account. Student income, as measured by FRL, was also a unique predictor of math, although negative, as it had been with first grade literacy, implying that low income was predictive of lower math scores, regardless of WaKIDS Literacy scores and teacher effects. Unlike the model from the convergent analysis, first grade teacher did not have a significant effect, indicating that a student's first grade teacher did not predict first grade math scores. These results together did not yield the desired support for the WaKIDS Literacy's discriminant validity; however, the emphasis on language in the Applied Problems subtest may have contributed to the unexpected

predictive relationship. This possibility will be discussed further, along with interpretation of these other convergent and discriminant validity results in the following chapter.

CHAPTER 5

Discussion

Main Findings

The purpose of the present study was to evaluate the predictive validity of the *WaKIDS GOLD*, Washington State's new kindergarten entry assessment, specifically answering the question: *Is the WaKIDS GOLD a valid assessment instrument in terms of its ability to predict later elementary student literacy achievement?* To determine this, student scores from the WaKIDS Literacy domain were evaluated in relation to measurement of student literacy achievement at the end of the first grade year. Specific research questions were targeted at obtaining both convergent and discriminant evidence of predictive validity through use of hierarchical linear modeling and linear regression. The results of these analyses contributed new sources of evidence to the evolving understanding of the WaKIDS Literacy's overall validity.

Considering the findings from the previous chapter together, the most compelling evidence resulted from the hierarchical linear model results of research question 1a: WaKIDS Literacy was a unique and positive predictor of literacy achievement at the end of first grade. As predicted, the WaKIDS Literacy domain scores were able to explain a statistically significant amount of variance in the first grade literacy scores, independent from variance explained by student demographics (income, minority, and English language status) and first grade teacher. This was particularly telling because of the significant effect of first grade teacher, the nesting variable, also resulting from the model, as well as a significant effect of student income. Even though first grade teacher was predictive of student first grade literacy achievement, indicating a possible effect of teacher instruction, the predictive strength of WaKIDS Literacy remained for all students, regardless of teacher. Student income status was a unique negative predictor of first

grade literacy, supporting a well-documented finding in reading research that lower income students were more likely to have lower scores of first grade achievement than higher income students (McCoach et al., 2006). Despite this effect, the WaKIDS unique prediction of first grade literacy was strong enough to persist across students of all income levels.

Another source of evidence was revealed from examination of the specific literacy skills measured within the WaKIDS Literacy domain. While the predictive validity of the Literacy domain as a whole had been evaluated, the contributions of the five objectives within the domain (phonological awareness, alphabet knowledge, print concepts, comprehension, and emergent writing) to the predictive relationship between kindergarten and first grade literacy needed to be considered as well. It was of interest to determine whether the objectives had an effect all together on measures of corresponding first grade literacy skills, and/or if certain objectives were uniquely explaining variance in the outcome measures. Hierarchical linear regressions revealed that the five objectives together accounted for significant amounts of variance in all four of the individual first grade literacy subtests (Letter Word ID, TERA Alphabet, Conventions, and Meaning) while controlling for demographic variables – which in two of the models included a significant effect of student income. The addition that explained the largest amount of variance, 40%, was for the TERA Conventions subtest. The predictive ability of the objectives together was not a surprise, considering the significant findings for the Literacy domain in the HLM analysis. However, when looking at the individual contributions of the objectives after being entered into the model, only one objective emerged as a unique predictor for three of the four first grade literacy outcomes: alphabet knowledge. No other objective made any unique contribution to any of the outcomes. This implied that the remaining variance in the first grade literacy outcomes was shared among the objectives, possibly due to their measurement of one

common literacy construct predictive of the first grade literacy outcomes. Each objective may still have the ability to measure the literacy skill that it was designed to measure, but when in the presence of the other literacy objectives, the predictive abilities of this shared construct likely dominate, explaining the strength of the objectives in predicting the outcomes together. This provides further evidence to support the predictive validity of the Literacy domain as a whole, but if users of WaKIDS are looking to measure one or more of the individual literacy skills within the domain, beyond alphabet knowledge, this finding of shared variance should be taken into consideration.

The unique predictive ability of alphabet knowledge on its own suggests that this WaKIDS objective, composed of the two dimensions, 16a (identifies and names letters), and 16b (uses letter-sound knowledge), might provide more precise measurement of the skill it is intended to measure (alphabet knowledge) than the other objectives in the Literacy domain for their own skills. This finding is not unexpected considering that alphabet knowledge has been documented in early literacy research as one of the strongest and most consistent predictors of later literacy achievement (Lonigan et al., 2000; MacDonald et al., 2013; NELP, 2009; Schatschneider, Fletcher, Francis, Carlson, & Foorman, 2004; Smith, Scott, Roberts, & Locke, 2008; Wagner et al., 1997). At the same time, Paris (2005) suggested that caution be taken when interpreting the predictive ability of a constrained skill such as alphabet knowledge due to its short period of development. Because alphabet knowledge can be mastered early in reading development, there is greater likelihood that its measurement will result in a ceiling effect, depending on when in a child's development it is measured (Paris, 2005). However, the unique predictive validity revealed for alphabet knowledge in the present study, along with the lack of

ceiling effects in the scores from both WaKIDS alphabet dimensions, prevent it from being a concern here and support its use as a predictor in this situation.

To further support the predictive validity argument, WaKIDS Literacy's discriminant validity was evaluated in addressing the second research question. As discussed in Chapter 2, discriminant, or divergent, validity is present when an assessment is found to not measure a construct it is not intended to measure (Messick, 1989). For this study, the construct that the WaKIDS Literacy Domain intended to measure was early literacy skill, while it did not intend to measure early math skill. Scores of first grade math achievement were used as the outcome in another HLM analysis, again selected to control for the nesting teacher variable, in order to confirm that little to no predictive relationship existed between the first grade math and WaKIDS Literacy scores. Unexpectedly, WaKIDS Literacy accounted for a significant amount of variance in the first grade math scores, a finding which did not yield evidence for discriminant validity as hoped.

The significant predictive relationship that resulted between WaKIDS Literacy and first grade math scores is not entirely surprising, however. First, the Applied Problems subtest, used to assess first grade math, required the assessor to ask questions and gather information from students verbally, while the child simultaneously looked at a visual representation. As students progressed through the assessment, items changed from providing pictorial representations to displaying text of actual sentences of the problem. While the assessor still read these sentences aloud, students had the option to read them on their own. Because of the heavy verbal demands (both listening comprehension and production), it seems logical that an assessment with language-based items such as the Applied Problems might predict language and related literacy skills, in addition to math.

Betts et al. (2009) also examined discriminant validity as part of their predictive validity study and experienced similar findings, with their predictor literacy and math variables highly correlating with both literacy and math outcome variables, when they had expected to see lower correlations between literacy and math. Betts et al. conjectured that in the early developing years of these skills, there may not be a clear distinction between the two skill types that is typically seen in later elementary school outcomes. Rather, a more broad range of skills that contribute to development of both literacy and math skills (including skills such as learning letters and numbers, sounds, and counting) may exist. This could potentially explain the predictive relationship that resulted between between WaKIDS Literacy and first grade math. Measurement of this broad skill range could potentially be a better predictor of later literacy and math achievement, and therefore, Betts et al. recommend use of a multivariate approach in future research predicting later achievement.

The use of hierarchical linear modeling in this study revealed two opposing and noteworthy findings regarding the group-level variable of first grade teacher: in the first HLM analysis, the grouping of students by first grade teacher had a statistically significant effect separate from the WaKIDS – first grade literacy predictive relationship, while the discriminant validity's HLM did not reveal a significant effect of first grade teacher on students' math performance. Both of these findings have instructional implications. For the first, the presence of the teacher effect suggests that the similar literacy performance within first grade classrooms could be due to the literacy instruction provided by the teacher. First grade instruction in the U.S. is known to be heavily focused on students' reading development, with instruction typically targeting phonological skill, decoding, and fluency to support comprehension (Common Core State Standards, 2015; Snow et al., 1998). Furthermore, in a study using data from the Early

Childhood Longitudinal Study – Kindergarten (ECLS-K), McCoach et al. (2006) found that reading gains were greater for students in first grade than in kindergarten, which also supports this study's finding of a teacher effect in first grade and not kindergarten. Therefore, the presence of this significant teacher effect indicates that first grade literacy instruction is having an impact on student learning to some degree, as we would expect.

On the other hand, the ability of WaKIDS Literacy to predict first grade Applied Problems scores was no more similar within classrooms than between. This can potentially be explained by teacher instruction as well. The lack of a first grade teacher effect implies that the literacy and math instruction provided in first grade, at least for this sample, did not contribute to how a student performed on the Applied Problems. In other words, even though the "math" skill measured by Applied Problems clearly overlaps with literacy skill, it does not seem to be a skill that is a direct focus of first grade instruction, whether math or literacy. This contributes further evidence that the discriminant validity results are not a concern for the WaKIDS Literacy's validity. The Applied Problems subtest, while backed with substantial research for its valid measurement of math knowledge, math achievement, and quantitative reasoning (Rathvon, 2004), may not be the most appropriate measure of a skill that can be clearly distinguished from literacy skill in first grade.

One other factor to consider in interpreting these teacher effects: the coefficient for WaKIDS Literacy resulting from the full models of both the Literacy and Math HLM analyses was very small in size, specifically .01 points for both. As mentioned in the Results, this can partially be attributed to the z-score format of both outcome measures, but its size nevertheless implies that the predictive power of scores from kindergarten to first grade was small. While the size of the coefficient does not change the predictive strength of the WaKIDS Literacy domain, it

could be indicative of the instruction that was provided during the time between the beginning of kindergarten and end of first grade – potentially either ineffective instruction or instruction affecting student performance to a small degree. Regardless, this is not the only study to find a strong effect of a predictor with a small coefficient. Musu-Gillette and colleagues (2015) also used z-scores for their outcome measures of kindergarten reading achievement and had small coefficients result for many student and teacher level predictors used in their multi-level models, some as low as the thousandths range. They explained these low coefficients as standard deviation increases (or decreases) in the outcome measure and did not factor them into any conclusions about their significant predictors.

One last variable of this study to address is student demographics. As is well documented in early literacy research, a student's income level, race/ethnicity, and first language are all predictive of student's later literacy achievement (McCoach et al., 2006). There is deep concern for students from low-income families, as well as students of a racial or ethnic minority, and students with limited English-language proficiency, as they have been found to be most at-risk for experiencing reading difficulties throughout schooling (Slavin & Cheung, 2005; Snow et al., 1998; Teale, Paciga, & Hoffman, 2007). In the present study, student income level was the one demographic variable that consistently held unique predictive ability when serving as a control variable in the HLM and regression analyses. It was also consistently a negative predictor, with affirmative FRL status predicting lower literacy scores. This confirmed what has been continuously documented in the literature – that students from lower-income families tend to have lower literacy skills than higher income students (McCoach et al., 2006). While WaKIDS Literacy was always able to explain significant variance in first grade literacy independent of student income, the unique predictive effects of income persisted for this sample (which was

disproportionally low income status). It is therefore essential that the socioeconomic backgrounds of students be considered when interpreting WaKIDS Literacy scores and their predictive abilities, considering the risk for future reading difficulty.

Broader Implications

This study served to evaluate the WaKIDS Literacy's predictive validity and add to the overall understanding of the *WaKIDS GOLD's* validity as a kindergarten entry assessment. The concurrent validity of the Literacy domain had already received support from the University of Washington's study (Soderberg et al., 2013), but its ability to predict skills over time had yet to be examined. The present study may be one of the first to empirically examine the predictive validity of any domain of the *WaKIDS GOLD*, and possibly the *TS* GOLD®. It can be concluded from the analyses and interpretations above that this study has sufficiently provided convergent evidence to support the WaKIDS Literacy domain's predictive validity, and thus contributes to the overall validity of the WaKIDS assessment tool.

Several positive implications emerge from this validity evidence. Many pertain to the students and teachers who are most directly affected by use of the *WaKIDS GOLD*. As discussed in Chapter 2, *WaKIDS GOLD* strives to fulfill two of Snow and Van Hemel's (2008) main purposes of early childhood assessment: *determining an individual child's level of functioning* and *guiding intervention and instruction*. With knowledge that the WaKIDS Literacy scores are valid, teachers can rely on this assessment tool to effectively serve these purposes. Specifically, the WaKIDS Literacy domain can provide kindergarten teachers with accurate information regarding each of their students' incoming literacy skill levels at the start of kindergarten. Such information can benefit not only teachers in supporting their students, but also students with the

overall kindergarten transition process (Scott-Little et al., 2011). Kindergarten teachers can then use this information as a "baseline" for literacy instruction to meet the needs of both individual students and the classroom as a whole (to the extent that the curriculum/school policies allow) (Scott-Little et al., 2011, p. 1).

Furthermore, teachers can use data from the WaKIDS Literacy domain to guide them in providing effective intervention to students when necessary. It is well known that many students, particularly those from low-income families, minority ethnic/racial backgrounds, and limited experience with speaking English, enter kindergarten at risk for later literacy difficulty (McCoach et al., 2006; Slavin & Cheung, 2005; Snow et al., 1998; Teale, Paciga, & Hoffman, 2007), and in order for them to experience success, early identification of the difficulties and subsequent intervention is essential (Spira, Bracken, & Fischel, 2005; Torgesen, 1998). For these at-risk students, the WaKIDS Literacy Domain's predictive abilities can help facilitate identification of their difficulties with a teacher documenting their skills over time. However, it is important to remember that Teaching Strategies did not intend for GOLD® to be used as an individual screening or diagnostic measure, and because of this, teachers should not rely on these scores alone when making decisions about students.

The predictive validity findings also have implications for Washington State and its ongoing development of the WaKIDS assessment program. The State's primary intent for the data collected by *WaKIDS GOLD* is to acquire a "snapshot" of the developmental skill levels of students across the state upon kindergarten entry in order to "inform state and district-level decisions about educational policy and investments, and classroom decisions about individualized learning" (OSPI WaKIDS Website, 2015). Additional aims align with general goals common among kindergarten entry assessments, such as determining the overall school

readiness of students entering kindergarten, targeting where greater efforts and supports are needed to help both teachers and students succeed, and providing parents with pertinent information regarding their child's learning and development (Scott-Little et al., 2011). With this study's evidence of the WaKIDS Literacy domain's predictive validity, the State can have more confidence that these goals are successfully being met, at least for use of the Literacy domain.

Finally, this validity evidence can contribute to the growing knowledge and research surrounding the use of observation-based assessment tools and kindergarten entry assessments, specifically with regard to the measurement of early literacy. The present study, in addition to the previous concurrent validity study, confirmed that observation-based assessments can be effectively evaluated in validity studies through comparison to a criterion measure. While it is preferable that this criterion measure be of the same assessment format as the predictive measure (Meisels et al., 2008), other predictive validity studies of observation-based assessments, in addition to the present one, have successfully used direct, standardized assessments, with well-established reliability and validity, as criterion measures (Downer et al., 2012; Gallant, 2009; Sekino & Fantuzzo, 2005). Additionally, the unique predictive abilities of the WaKIDS alphabet knowledge objective, as well as all of the literacy objectives together, call for closer examination of multivariate literacy measures and the predictive strengths of the individual skill measures within. Therefore this WaKIDS Literacy predictive validity study could potentially serve as a model for future validity studies of early literacy, as well as kindergarten entry assessments, but not without considering some of the study's limitations, which are detailed next.

Limitations and Future Research

While this study succeeded in contributing convergent validity evidence to the WaKIDS Literacy domain's overall predictive validity, there are limitations that must be addressed before considering the generalizability of the findings. One of the most obvious concerns was the small sample size and uneven distribution of students across classrooms and schools. This was a result of the need for a convenience sample; low response rates from principals and teachers, as well as parent consent, led to very few students able to participate. The uneven distribution of teachers across schools was due to voluntary participation; not all schools had all first grade teachers volunteer. Furthermore, the sample was drawn from only one school district. The State would benefit from another predictive validity study with a much larger student sample (ideally large enough to randomly select participants) drawn from multiple school districts from across the State.

Users of the *WaKIDS GOLD* should be careful not to generalize these results across the other developmental domains of the assessment, as the study only examined data from the Literacy domain. It is possible that the other domains may have similar evidence of predictive validity as well; however, in the 2013 concurrent validity study, only three of the six domains, Literacy, Math, and Language were found to demonstrate strong concurrent validity (Soderberg et al., 2013). The State may want to explore this further if some of the WaKIDS domains are more valid than others, particularly as they continue to rely on the different domain scores for accurately predicting later educational outcomes.

Regarding the study's selection of outcome assessments, the results of the discriminant validity analysis indicated that the *WJ-III* Applied Problems subtest as a measure of first grade

math skill might not have been the most appropriate selection considering its purpose in the study. It is likely that this measure's dependence on language in its administration led it to correlate so highly with literacy skill. Because of the possibility of a common, more general construct linking these two skill sets, it would be advisable to stay away from measures of math skill and instead choose a measure of developmental skill known to be even less related, such as gross motor ability, for evaluation of a literacy assessment's discriminant validity.

Finally, one other concern not addressed in this study was the reliability of the WaKIDS scores, particularly considering kindergarten teachers as raters and the subjectivity that comes with an observation-based assessment such as WaKIDS. The University of Washington conducted an inter-rater reliability study with kindergarten teachers in conjunction with the concurrent validity study in 2013 to determine how reliable teachers were in their scoring of student abilities across the WaKIDS domains. The findings for the Literacy domain in particular were questionable, with inconsistency in teachers' ratings of the same children across the domain; the most inconsistency was found in scores of the reading comprehension objective. Waterman, McDermott, Fantuzzo, and Gadsden (2012) warned that assessor variance is more common among teachers in their ratings of students due to subjective judgments, which can threaten an assessment's validity. Therefore, future studies of the *WaKIDS GOLD's* validity need to take teacher's reliability in scoring into account to ensure that the variability in student scores is attributable to the students themselves and not teacher subjectivity.

CONCLUSION

Predictive validity is a crucial yet often overlooked psychometric property of most educational assessments, including early literacy measures, that is essential to evaluate, particularly when considering student educational outcomes. The main goal of this study was to provide evidence for Washington State and its teachers that their kindergarten entry assessment, *WaKIDS GOLD*, is a valid instrument in terms of its ability to predict later elementary student literacy achievement. Through statistical comparisons of kindergarten and first grade literacy performance, this study was able to demonstrate with a small number of students that scores resulting from the WaKIDS Literacy domain validly predicted measurement of students' literacy achievement at the end of first grade, nearly two years later. Although evidence of discriminant validity was not revealed, the convergent evidence that resulted held greater weight with effects of teacher and student income level taken into account. This finding can give the State more assurance that their use of the *WaKIDS GOLD* can provide important information needed to inform instruction, as well as effectively predict and improve later student literacy outcomes. This exemplifies the powerful role that assessment can have for Washington's youngest readers who need it most.

REFERENCES

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

Augustyniak, K.M., Cook-Cottone, C.P., & Calabrese, N. (2004). The predictive validity of the Phelps Kindergarten Readiness Scale. Psychology in the Schools, 41, 509–516.

Betts, J., Pickart, M., & Heistad, D. (2009). Construct and predictive validity evidence for curriculum-based measures of early literacy and numeracy skills in kindergarten. *Journal of Psychoeducational Assessment, 27,* 83-95.

Bordignon, C. M., & Lam, T. C. M. (2004). The early assessment conundrum: Lessons from the past, implications for the future. *Psychology in the Schools, 41*(7), 737-749.

Bridges, M. S., & Catts, H. W. (2011). The use of a dynamic screening of phonological awareness to predict risk for reading disabilities in kindergarten children. *Journal of Learning Disabilities*, *44,* 330-338.

Brown, G., Scott-Little, C., Amwake, L., & Wynn, L. (2007). *A review of methods and instruments used in state and local school readiness evaluations*. (Issues and Answers Report, REL 2007-No. 004.) Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.

Burts, D. C., & Kim, D. (2014). The Teaching Strategies GOLD® assessment system: Measurement properties and use. *Dialog*, *17*(3), 122-135.

Caffrey, E., Fuchs, D., & Fuchs, L. S. (2008). The predictive validity of dynamic assessment: A review. *The Journal of Special Education*, *41*(4), 254-270.

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2009). Floor

effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities, 42*, 163–176.

Clay, M. M. (1966). *Emergent reading behavior.* Unpublished doctoral dissertation, University of Auckland, New Zealand.

Connors-Tadros, L. (2014). *Information and resources on developing state policy on kindergarten entry assessment (KEA)* (CEELO FASTFacts). New Brunswick, NJ: Center on Enhancing Early Learning Outcomes.

Coyne, M. D., & Harn, B. A. (2006). Promoting beginning reading success through meaningful assessment of early literacy skills. *Psychology in the Schools*, *43*(1), 33-43.

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

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental psychology*, *33*(6), 934-945.

Cunningham, A. E., Zibulsky, J., & Callahan, M. D. (2009). Starting small: Building preschool teacher knowledge that supports early literacy development. *Reading and Writing*, *22*(4), 487-510.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... & Japel, C. (2007). School readiness and later achievement. *Developmental psychology*, *43*(6), 1428-1446.

ED.gov. (2011). We can't wait: Nine states awarded Race to the Top-Early Learning Challenge grants awards will help build statewide systems of high quality early education programs. Retrieved from: http://www.ed.gov/news/press-releases/we-cant-wait-nine-states-awarded-race-top-early-learning-challenge-grants-awards.

Evans, G. W., & Rosenbaum, J. (2008). Self-regulation and the income-achievement gap. *Early Childhood Research Quarterly, 23*(4), 504-514.

Farrall, M. L. (2012). *Reading assessment: Linking language, literacy, and cognition*. Hoboken, NJ: John Wiley & Sons.

Fitzgerald, J., Schuele, C. M., & Roberts, J. E. (1992). Emergent literacy: What is it and what does the teacher of children with learning disabilities need to know about it? *Reading and Writing Quarterly, 8,* 71-85.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88*(1), 3.

Gallant, D. J. (2009). Predictive validity evidence for an assessment program based on the Work Sampling System in mathematics and language and literacy. *Early Childhood Research Quarterly, 24,* 133-141.

Gough, P. B., and Tunmer, W. E. (1986). Decoding, reading, and reading disability. *RASE: Remedial and Special Education, 7,* 6-10.

Graue, E. (1999). Diverse perspectives on kindergarten contexts and practices. In R. C. Pianta & M. J. Cox (Eds.), *The transition to kindergarten* (pp. 109-142). Baltimore: Brookes.

Heath, S. B. (1983). *Ways with words: Language, life and work in communities and classrooms*. Cambridge: Cambridge University Press.

Hecht, S. A., & Greenfield, D. B. (2001). Comparing the predictive validity of first grade teacher ratings and reading-related tests on third grade levels of reading skills in young children exposed to poverty. *School Psychology Review, 30*(1), 50-69.

Hosp, J. L., Hosp, M. A., & Dole, J. K. (2011). Potential bias in predictive validity of universal

screening measures across disaggregation subgroups. *School Psychology Review*, *40*(1), 108.

Invernizzi, M. A., Landrum, T. J., Howell, J. L., & Warley, H. P. (2008). Toward the peaceful coexistence of test developers, policymakers, and teachers in an era of accountability. In R. D. Robinson & M. C. McKenna (Eds.), *Issues and trends in literacy education* (pp. 198-209). San Francisco: Pearson.

Jitendra, A. K., Sczesniak, E., & Dearline-Buchman, A. (2005). An exploratory validation of curriculum-based mathematical word problem-solving tasks as indicators of mathematics proficiency for third graders. *School Psychology Review*, *34*, 358-371.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, *80*, 437−447.

Kame'enui, E. J., Fuchs, L., Francis, D. J., Good III, R., O'Connor, R. E., Simmons, D. C., Tindal, G., & Torgesen, J. K. (2006). The adequacy of tools for assessing reading competence: A framework and review. *Educational Researcher*, *35*, 3-11.

Kim, J., & Suen, H. K. (2003). Predicting children's academic achievement from early assessment scores: A validity generalization study. *Early Childhood Research Quarterly, 18*, 547–566.

Lambert, R. G., Kim, D. H., & Burts, D. C. (2014). Using teacher ratings to track the growth and development of young children using the Teaching Strategies GOLD® assessment system. *Journal of Psychoeducational Assessment*, *32*(1), 27-39.

LaParo, K. M., & Pianta, R. C. (2000). Predicting children's competence in the early school years: A meta-analytic review. *Review of Educational Research, 70,* 443–484.

Linklater, D. L., O'Connor, R. E., & Palardy, G. J. (2009). Kindergarten literacy assessment of English only and English language learner students: An examination of the predictive

validity of three phonemic awareness measures. *Journal of School Psychology*, *47*, 369-394.

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

Lonigan, C. J. (2006). Development, assessment, and promotion of pre-literacy skills. *Early Education and Development, 17*, 91–114.

Lonigan, C. J., Allan, N. P., & Lerner, M. D. (2011). Assessment of preschool early literacy skills: Linking children's educational needs with empirically supported instructional activities. *Psychology in the Schools*, *48*(5), 488-501.

Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent-variable longitudinal study. *Developmental Psychology, 36*, 596–613.

MacDonald, H. H., Sullivan, A. L., & Watkins, M. W. (2013). Multivariate screening model for later word reading achievement: Predictive utility of prereading skills and cognitive ability. *Journal of Applied School Psychology*, *29*, 52-71.

McGrew, K. S., Schrank, F. A., & Woodcock, R. W. (2007). *Woodcock Johnson III Normative Update: Technical Manual*. Rolling Meadows, IL: Riverside Publishing.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten– grade 3. *American Educational Research Journal, 38*, 73–95.

Meisels, S. J., Xue, Y., & Shamblott, M. (2008). Assessing language, literacy, and mathematics skills with Work Sampling for Head Start. *Early Education and Development, 19*, 963-981.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (13-104). Washington DC: American Council on Education.

Minnesota Human Capital Research Collaborative. (2011). Assessing the validity of Minnesota school readiness indicators: Summary report. Retrieved from: http://humancapitalrc.org/mn_school_readiness_indicators.pdf

Missall, K., Reschly, A., Betts, J., McConnell, S., Heistad, D., Pickart, M., ... & Marston, D. (2007). Examination of the predictive validity of preschool early literacy skills. *School Psychology Review*, *36*(3), 433-452.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*(3), 229-258.

Munger, K. A., & Blachman, B. A. (2013). Taking a "simple view" of the dynamic indicators of basic early literacy skills as a predictor of multiple measures of third-grade reading comprehension. Psychology in the Schools, 50(7), 722-737.

National Early Literacy Panel (NELP). (2009). *Developing Early Literacy*. Washington, DC: National Institute for Literacy.

Neuman, S. B. (2010). Lessons from my mother: Reflections on the National Early Literacy Panel Report. *Educational Researcher*, *39*(4), 301-304.

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly, 40*, 184-202.

Pelligrini, A., & Galda, L. (1998). The development of school-based literacy: A social ecological perspective. New York: Routledge.

Pianta, R. C., & McCoy, S. J. (1997). The first day of school: The predictive validity of early school screening. *Journal of Applied Developmental Psychology, 18*, 1–22.

Rafoth, M. A. (1997). Guidelines for developing screening programs. *Psychology in the*

Schools, *34*(2), 129-142.

Rathvon, N. (2004). *Early reading assessment: A practitioner's handbook*. New York: Guilford.

Reid, D. K., Hresko, W. P., & Hammill, D. D. (2001). *Test of Early Reading Ability–Third Edition Form A*. Austin, TX: PRO-ED.

Rupley, W. H., Blair, T. R., & Nichols, W. D. (2009). Effective reading instruction for struggling readers: The role of direct/explicit teaching. *Reading & Writing Quarterly*, *25*(2-3), 125-138.

Scott-Little, C., Bruner, C., & Schultz, T. (2011). Discussion guide to responding to focused investment area (E)(1) and competitive priority 3: Kindergarten entry assessment. Retrieved from: http://www.elccollaborative.org/illinois/doc_view/17-kindergarten-entry-assessment-discussion-guide.html

Sekino, Y., & Fantuzzo, J. (2005). Validity of the Child Observation Record: An investigation of the relationship between COR dimensions and social-emotional and cognitive outcomes for Head Start children. *Journal of Psychoeducational Assessment, 23*(3), 242–260.

Sénéchal, M., LeFevre, J. A., Smith-Chant, B. L., & Colton, K. V. (2001). On refining theoretical models of emergent literacy the role of empirical evidence. *Journal of School Psychology*, *39*(5), 439-460.

Slavin, R. E., & Cheung, A. (2005). A synthesis of research on language of reading instruction for English language learners. *Review of Educational Research, 75*(2), 247-284.

Smith, S. L., Scott, K. A., Roberts, J., & Locke, J. L. (2008). Disabled readers' performance on tasks of phonological processing, rapid naming, and letter knowledge before and after Kindergarten. *Learning Disabilities Research & Practice*, *23*, 113-124.

Snow, K. (2011). Developing Kindergarten Readiness and Other Large-Scale Assessment Systems: Necessary Considerations in the Assessment of Young Children. Washington,

DC: National Association for the Education of Young Children.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, D.C.: National Academy Press.

Snow, C. E., & Van Hemel, S. B.. (2008). *Early Childhood Assessment: Why, What, and How*. Washington, D.C.: The National Academies Press.

Soderberg, J. S., Stull, S., Cummings, K., Nolen, E., McCutchen, D., & Joseph, G. (2013). Inter-rater reliability and concurrent validity study of the Washington Kindergarten Inventory of Developing Skills (WaKIDS). Unpublished report prepared for the State of Washington Office of Superintendent of Public Instruction.

Spencer, E. J., Spencer, T. D., Goldstein, H, & Schneider, N. (2013). Identifying early literacy needs: Implications for child outcome standards and assessment systems. In T. Shanahan & C. J. Lonigan (Eds.), *Early childhood literacy: The National Early Literacy Panel and beyond*. Baltimore: Brookes.

Spira, E. G., Bracken, S. S., & Fischel, J. E. (2005). Predicting improvement after first-grade reading difficulties: The effects of oral language, emergent literacy, and behavior skills. *Developmental Psychology*, *41*, 225−234.

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology*, *38*, 934 – 947.

Sulzby, E., & Teale, W. (1991). Emergent literacy. In R, Barr. M. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. II, pp. 727-758). New York: Longman.

Teaching Strategies. (2010). Research foundation: Teaching Strategies GOLD® assessment system. Retrieved from: http://www.teachingstrategies.com/content/pageDocs/Research-

Foundation-GOLD-2010.pdf

Teaching Strategies. (2011a). Press release: Teaching Strategies GOLD® assessment system selected for use in the Washington Kindergarten Inventory of Developing Skills program. Retrieved from: http://www.teachingstrategies.com/content/pageDocs/Teaching-Strategies-WaKIDS-Press-Release-10-2011.pdf

Teaching Strategies. (2011b). GOLD® assessment system: Technical summary. Retrieved from: http://www.teachingstrategies.com/content/pageDocs/GOLD-Tech-Summary-8-18-2011.pdf

Teaching Strategies. (2013). Teaching Strategies GOLD®, birth through kindergarten: Touring guide. Retrieved from: http://www.teachingstrategies.com/content/pageDocs/GOLD-Touring-Guide_5-2013.pdf

Teale, W. H., Paciga, K. A., & Hoffman, J. L. (2007). Beginning reading instruction in urban schools: The curriculum gap ensures a continuing achievement gap. *The Reading Teacher*, *61*(4), 344-348.

Teale, W. H., & Sulzby, E. (Eds.). (1986). *Emergent literacy: Writing and reading.* Norwood, NJ: Ablex.

Thorkildsen, T. A. (2005). *Fundamentals of measurement in applied research*. San Francisco: Pearson.

Torgesen, J. K. (1998). Catch them before they fall: Identification and assessment to prevent reading failure in young children. *American Educator*, *22*, 32-41.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., et al. (1997). Changing relations between phonological processing abilities and word-level

reading as children develop from beginning to skilled readers: A 5-year longitudinal study. *Developmental Psychology, 33,* 468–479.

Waterman, C., McDermott, P. A., Fantuzzo, J. W., & Gadsden, V. L. (2012). The matter of assessor variance in early childhood education—Or whose score is it anyway?. *Early Childhood Research Quarterly*, *27*(1), 46-54.

West, J., Denton, K., & Germino-Hausken, E. (2000). *America's kindergartners: Findings from the Early Childhood Longitudinal Study, kindergarten class of 1998-99, fall 1998.* Washington, DC: National Center for Education Statistics, U. S. Department of Education, Office of Educational Research and Improvement.

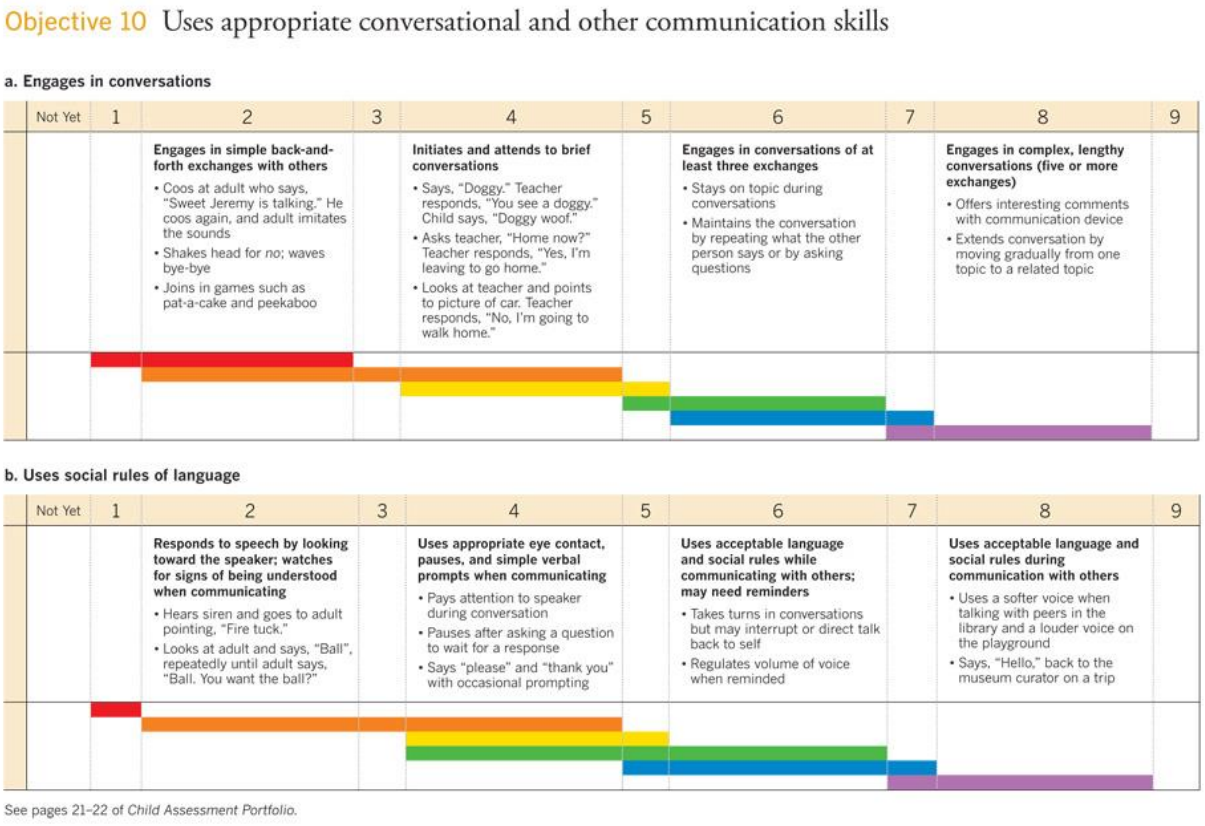
Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development*, *69*, 848 – 872.

Whitehurst, G. J. & Lonigan, C. J. (2001). Emergent literacy: Development from prereaders to readers. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research*. New York: Guilford Press.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement*. Rolling Meadows, IL: Riverside Publishing.

APPENDIX A

Teaching Strategies GOLD® Colored Bands Progression, Objective 10: Uses appropriate conversational and other communication skills. Image obtained from *Teaching Strategies* GOLD® *online*: http://www.teachingstrategies.com/page/*GOLD*-assessment-online.cfm.



Objective 10 Uses appropriate conversational and other communication skills

a. Engages in conversations

Not Yet	1	2	3	4	5	6	7	8	9
		Engages in simple back-and-forth exchanges with others • Coos at adult who says, "Sweet Jeremy is talking." He coos again, and adult imitates the sounds • Shakes head for *no*; waves bye-bye • Joins in games such as pat-a-cake and peekaboo		**Initiates and attends to brief conversations** • Says, "Doggy." Teacher responds, "You see a doggy." Child says, "Doggy woof." • Asks teacher, "Home now?" Teacher responds, "Yes, I'm leaving to go home." • Looks at teacher and points to picture of car. Teacher responds, "No, I'm going to walk home."		**Engages in conversations of at least three exchanges** • Stays on topic during conversations • Maintains the conversation by repeating what the other person says or by asking questions		**Engages in complex, lengthy conversations (five or more exchanges)** • Offers interesting comments with communication device • Extends conversation by moving gradually from one topic to a related topic	

b. Uses social rules of language

Not Yet	1	2	3	4	5	6	7	8	9
		Responds to speech by looking toward the speaker; watches for signs of being understood when communicating • Hears siren and goes to adult pointing. "Fire tuck." • Looks at adult and says, "Ball", repeatedly until adult says, "Ball. You want the ball?"		**Uses appropriate eye contact, pauses, and simple verbal prompts when communicating** • Pays attention to speaker during conversation • Pauses after asking a question to wait for a response • Says "please" and "thank you" with occasional prompting		**Uses acceptable language and social rules while communicating with others; may need reminders** • Takes turns in conversations but may interrupt or direct talk back to self • Regulates volume of voice when reminded		**Uses acceptable language and social rules during communication with others** • Uses a softer voice when talking with peers in the library and a louder voice on the playground • Says, "Hello," back to the museum curator on a trip	

See pages 21–22 of *Child Assessment Portfolio.*

APPENDIX B

WaKIDS GOLD Objectives and Dimensions.

Social–Emotional

1. Regulates own emotions and behaviors
    - b. Follows limits and expectations
    - c. Takes care of own needs appropriately
2. Establishes and sustains positive relationships
    - c. Interacts with peers
    - d. Makes friends

Physical

4. Demonstrates traveling skills
5. Demonstrates balancing skills
6. Demonstrates gross-motor manipulative skills
7. Demonstrates fine-motor strength and coordination
    - a. Uses fingers and hands
    - b. Uses writing and drawing tools

Language

9. Uses language to express thoughts and needs
    - a. Uses an expanding expressive vocabulary
    - b. Speaks clearly
    - c. Uses conventional grammar
    - d. Tells about another time or place
10. Uses appropriate conversational and other communication skills
    - a. Engages in conversations
    - b. Uses social rules of language

Cognitive

11. Demonstrates positive approaches to learning
    - c. Solves problems
    - d. Shows curiosity and motivation
    - e. Shows flexibility and inventiveness in thinking
12. Remembers and connects experiences
    - a. Recognizes and recalls
13. Uses classification skills

Literacy

15. Demonstrates phonological awareness
    - a. Notices and discriminates rhyme
    - b. Notices and discriminates alliteration
    - c. Notices and discriminates smaller and smaller units of sound
16. Demonstrates knowledge of the alphabet
    - a. Identifies and names letters
    - b. Uses letter–sound knowledge
17. Demonstrates knowledge of print and its uses
    - b. Uses print concepts

18. Comprehends and responds to books and other texts
    a. Interacts during read-alouds and book conversations
    b. Uses emergent reading skills
    c. Retells stories
19. Demonstrates emergent writing skills
    a. Writes name
    b. Writes to convey meaning

Mathematics

20. Uses number concepts and operations
    a. Counts
    b. Quantifies
    c. Connects numerals with their quantities
21. Explores and describes spatial relationships and shapes
    b. Understands shapes
22. Compares and measures

APPENDIX C

Example of First Grade Student Score Report and Teacher Guide.

Example student report of scores, Spring 2014:

Child's Name						
Date of Assessment: 06/09/14						
Age at Assessment Time:						
7 years, 9 months						
TERA-3		**Reading Quotient**	**Percentile Rank**	**Description**		
Overall		94	35%	Average		
	Raw Score	**Age Equivalent**	**Grade Equivalent**	**Standard Score**	**Description**	
Alphabet	25	7y, 9m	late 2nd grade	10	Average	
Conventions	14	6y, 7m	mid 1st grade	7	Below Average	
Meaning	23	7y, 9m	late 2nd grade	10	Average	
WJ III Tests of Achievement						
	Raw Score	**Age Equivalent**	**Grade Equivalent**	**Standard Score**	**Percentile Rank**	**Description**
Letter-Word ID (reading decoding)	41	8y, 3m	early 3rd grade	107	67%	advanced
Applied Problems (math knowledge)	35	10y, 6m	late 4th grade	125	95%	very advanced

Teacher Guide for student assessment scores:

TERA-3:

The *Test of Early Reading Ability*, *3rd edition*, assesses children's mastery of early developing reading skills. It is designed to measure early reading skill in children between the ages of 3 years, 6 months to 8 years, 6 months. The assessment is divided into 3 subtests, each measuring a "different but highly interrelated" component of reading: **Alphabet**, **Conventions**, and **Meaning**. Descriptions of each subtest from the TERA-3 Examiner's Manual are provided below:

Subtest I, Alphabet: measures children's knowledge of the alphabet and sound-letter correspondence. The items in this subtest also measure letter name knowledge, the ability to determine the initial and final sounds in printed words, and the awareness of letters printed in different fonts.

Subtest II, Conventions: measures children's familiarity with the conventions of print. Items in this subtest measure book handling (e.g., knowing the correct orientation of a book, knowing where to begin reading); print conventions (e.g., letter orientation, case, presentation of print, text genre); and knowledge of punctuation, capitalization, and spelling.

Subtest III, Meaning: measures children's ability to comprehend the meaning of printed material. Items in this subtest measure comprehension of words, sentences, and paragraphs. Items also measure relational vocabulary, sentence construction, and paraphrasing.

Raw scores from each subtest (the total number of items correct) can be converted into standard scores, based on the child's age, and the standard scores from each subtest can be combined to form a composite, known as the **Reading Quotient**. The Reading Quotient reflects the child's ability across a variety of reading activities involving conventions, alphabet, and meaning. *It is the best indicator of overall reading ability*.

For each subtest, the child's raw score, age and grade equivalents, percentile, and subtest and composite standard scores are recorded. Below is an example of a table of scores for a student who completed the TERA-3. You can use this example to help interpret the scores in the tables for your own students:

Mya Henry					
Date of Assessment: 11/01/13					
Age at Assessment Time: 6 years, 1 month					
TERA-3		**Reading Quotient**	**Percentile Rank**	**Description**	
Overall		104	61%	Average	
	Raw Score	**Age Equivalent**	**Grade Equivalent**	**Standard Score**	**Description**
Alphabet	23	6y, 10m	late 1st grade	13	Above Average
Conventions	14	6y, 7m	mid 1st grade	12	Average
Meaning	9	5y, 4m	early kindergarten	7	Below Average

This is Mya's reading composite score, representing performance across all 3 subtests.

61% of others at Mya's age level scored at or below her score.

Mya's Reading Quotient falls in the *average* performance range.

Mya's Alphabet score of 23 is typical of children 6 years and 10 months old, or, in terms of grade level, is typical of children at the end of their first grade year.

Mya's Meaning Standard Score is 7, which falls in the *below average* range. Standard scores allow for comparisons across subtests. Looking at her Alphabet and Conventions standard scores, they are much higher than her Meaning score, and so they are considered *above average* and *average*, respectively. See the table on the following page to understand the ranges of all possible standard scores.

From this table, we can tell that Mya's overall reading skill can be considered average, reflecting appropriate skill for her age. She shows above average performance with Alphabet skill and average performance with Conventions of print, but her skill in comprehending printed material (Meaning) is weaker. Reading instruction for Mya could focus less on letter-sound correspondence, and possibly more on print conventions as well as comprehension.

TERA-3's Guide to Interpreting Subtest Standard Scores:

Standard Scores	Description
17-20	Very Superior
15-16	Superior
13-14	Above Average
8-12	Average
6-7	Below Average
4-5	Poor
1-3	Very Poor

**Standard scores are based on a distribution with a mean of 10 and a standard deviation of 3. They can only fall within the range of 1-20.*

WJ-III ACH:

The Woodcock Johnson III, Tests of Achievement is a comprehensive assessment of academic achievement with several subtests measuring various abilities. It can be used with subjects from preschool level up to adulthood. The assessment can be administered in full for a comprehensive assessment of achievement, or individual subtests can be selected when focusing on assessing specific abilities. For my purposes, I administered only two of these subtests: **Letter-Word Identification** (to measure reading decoding) and **Applied Problems** (to measure math knowledge). Descriptions of each and what they measure are provided below:

Letter-Word Identification (Letter-Word ID): measures a child's ability to identify isolated letters and read real words, measuring the child's reading decoding skill.

Applied Problems: measures a child's ability to analyze and solve orally presented mathematical problems, measuring math achievement, math knowledge, and quantitative reasoning.

For both of these subtests, age and grade equivalents, standard scores, and percentile ranks can all be determined from the child's raw score. The standard scores allow for comparison across subtests by using a mean of 100 and a standard deviation of 15. With the mean of 100, if a child's standard score is 100, his or her performance is right at the average level. Below is an example of a table of scores for a student who completed both the Letter-Word ID and Applied Problems subtests of the WJ-III. You can use this example to help interpret the scores in the tables for your own students:

Josef Mayers						
Date of Assessment: 04/28/14						
Age at Assessment Time: 7 years, 5 months						
Woodcock Johnson III Tests of Achievement						
	Raw Score	**Age Equivalent**	**Grade Equivalent**	**Standard Score**	**Percentile Rank**	**Description**
Letter-Word ID (reading decoding)	31	7y, 4m	early 2nd grade	96	41%	limited to average
Applied Problems (math knowledge)	29	8y, 4m	late 2nd grade	116	86%	advanced

Josef's Letter-Word ID Standard Score is 96, indicating that 41% of others at his age level scored at or below Josef's score. This is considered to fall in the *limited to average* range.

Josef's raw score of 31 for Letter-Word ID is typical of children 7 years and 4 months old, or, in terms of grade level, is typical of children in early 2nd grade. His raw score of 29 for Applied Problems is typical of children 8 years and 4 months old, or in terms of grade level, is typical of children at the end of 2nd grade.

For Applied Problems, Josef's Standard Score is 116, indicating that 86% of others at his age level scored at or below Josef's score. This is considered to fall in the *advanced* range.

From this table, we can tell that Josef's reading decoding skill is just about on par for his age, but there is room for improvement. His reading instruction should involve continued practice with decoding words and learning sight words. His math knowledge skill is stronger, with his performance considered advanced for his age. This suggests that he may need to be challenged more during math instruction.

Guide to Interpreting Letter-Word ID Standard Scores:

Standard Scores	Description
112 +	Very Advanced
106-111	Advanced
103-105	Average to Advanced
98-102	Average
95-97	Limited to Average
88-94	Limited
87 -	Very Limited

**Standard scores are based on a distribution with a mean of 100 and a standard deviation of 15.*

Guide to Interpreting Applied Problems Standard Scores:

Standard Scores	Description
125 +	Very Advanced
111-124	Advanced
105-110	Average to Advanced
97-105	Average
90-96	Limited to Average
80-90	Limited
79 -	Very Limited

**Standard scores are based on a distribution with a mean of 100 and a standard deviation of 15.*

VITA

Sara Stull was born in Boston, Massachusetts and lived most of her life in southern New Hampshire. She earned a Bachelor of Arts degree in Psychology at Dickinson College in 2006. After teaching elementary school for two years in New York City, she moved to Seattle in 2008 to pursue graduate study in Educational Psychology at the University of Washington. There, she earned a Master of Education degree in 2010, followed by a Doctor of Philosophy in Education in 2015.