2016

Can Teachers Tell Which Students are at Risk? Comparing Teacher Reading Risk Determinations with STAR Reading Risk Determinations

Leigh Anne Whitney Scherer

Can Teachers Tell Which Students are at Risk? Comparing Teacher Reading Risk Determinations with STAR Reading Risk Determinations

by

Leigh Anne Whitney Scherer

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

Doctor of Education
in
Learning and Leading

University of Portland
School of Education

2016

Abstract

The theoretical framework for this study was Michael Polanyi's concept of tacit knowing, that a professional's knowledge is composed of both the things he or she can describe explicitly and a tacit component that is difficult, if not impossible, to define or describe. In the national context of an increasing emphasis on accountability, the use of data, and standardized testing, teachers' judgments, composed as they are of a tacit component that can be difficult to express, are not always valued. A review of the literature revealed a gap in the research related to teachers' abilities to identify individual student risk in reading. The purpose of this research was to determine if, in the context of a Response to Intervention framework, teachers' professional judgments were equally predictive at determining risk level as the results provided by the screening tools in common use in school districts to identify students who would benefit from a reading intervention. This study examined two research questions: (a) what is the relationship between teacher judgment of student reading risk levels and the screening tool risk levels, and (b) are there variations in the relationships related to student characteristics? Using a tracking tool, 31 3rd and 4th grade teachers in a suburban school district in the Pacific Northwest recorded their determinations of their students' reading risk. Those results were then compared with the fall universal screening reading risk determinations from STAR Reading. Percent exact agreement tests were used to determine the concurrent validity of the two measures. Overall there

was an 83% match between STAR Reading and the teacher evaluations of each student's reading risk. When the results were disaggregated, most groups had percent exact agreement rates above 80%. This study might suggest that a teacher's professional judgment could be used as a screening tool, eliminating the need to purchase and maintain a commercially published assessment for the purposes of universal screening.

Acknowledgements

I would not have been able to complete this process without the support of so many. First, I would like to acknowledge my dissertation committee. Thank you to the chair of my committee, Dr. Phyllis Egby, for her ongoing support and guidance along the way. To Dr. Richard Christen for sparking the initial inspiration of my ideas and for guiding my thinking about historical perspectives of school reform and accountability. To Dr. Nicole Ralston, who was a cheerleader from the beginning, provided feedback on preliminary drafts, coached me on what a dissertation should look like, and re-taught me how to use SPSS! Finally, someone I think should be considered an honorary member of my dissertation committee, I would like to acknowledge Dr. Julie Kalnin for her crucial feedback and guidance in getting me to the finish line. In addition to the supportive faculty at the University of Portland, I would like to thank my work colleagues who provided support for the past few years.

Lastly, I want to thank my husband Bruce for his unwavering support, my son Julian for understanding why I could only make a couple of his track meets, cross country meets, or wrestling matches his first two years of high school, and to my daughter Livia for picking up the slack at home in so many ways including teaching herself how to cook so she could make meals for herself and her brother. I love you all!

Dedication

This is dedicated to my family: my husband Bruce and our children Julian and Livia and to my parents Mike and Kay Whitney. Thank you for your support all along the way.

Table of Contents

List of Tables

List of Figures

Chapter One: Introduction

Education means different things to different people. What seems like an ideal school to one person seems totally inappropriate to another. Some think the best schools or classrooms are highly structured with a focus on developing competency in a core set of knowledge while others consider a creative setting dominated by student-led inquiry a more effective model. For some, the best teachers are those who inhabit a central focus in the classroom, whose students look to them for direction and the source of knowledge. For others, students should be at the center of the learning process with the teacher following their lead. However, most everyone agrees that school, by definition, includes both teachers and students, and that teachers play a critical role. Without teachers there would be no school. But what is the teacher's role, particularly related to making decisions about students? How many of the decisions about students' learning should be left up to teachers? Are teachers equipped to make the important decisions that can have a lifelong impact on students' lives? What tools, such as tests, support teachers in their decision-making and when, if ever, should a teacher's judgment about a student supersede that of a test score?

In the US, the trend of how these questions would be answered has changed over time as the country and its place in the world have changed. There has been an ongoing discussion about the most fundamental question, "what is education for?" (Robinson & Aronica, 2015, p. xix) and not everyone agrees. The current standards and accountability movement in the US focuses on standardization and efficiency with

an emphasis on direct instruction and formal assessment, which includes the use of multiple-choice tests that allow for easy quantification of student learning and a skepticism about teachers' abilities to evaluate their own students' learning, (Robinson & Aronica, 2015).

This chapter will provide an overview and frame for the importance of this study. First there is a brief review of the history of the education accountability movement in the United States followed by a discussion of some of the effects of this movement on teachers. Next is a description of Response to Intervention and universal screening including information related to assessment validity. The chapter concludes with the problem statement and significance of the study.

Accountability

The modern history of school accountability in schools is often dated to the publication of the report, A Nation at Risk: The Imperative for Educational Reform, published in 1983 by the National Commission on Excellence in Education (Vinovskis, 2009). This report examined the quality of public education in US schools and determined that the country was at risk due to the mediocrity of public schools (National Commission on Excellence in Education, 1983), a concern that was shared by many policymakers and the public (Vinovskis, 2009). Following the publication of this report, the United States underwent a period of policy transformation that resulted in the ever increasing importance of the federal government in education (McGuinn, 2006). The original Elementary and Secondary Education Act (ESEA) of 1965 narrowly targeted disadvantaged students by providing additional resources to schools.

the reauthorization of 2002, known as No Child Left Behind, expanded the role of the federal government and represented a culmination of the reform movement that began with the A Nation at Risk report (McGuinn, 2006).

The primary goal of No Child Left Behind, which was passed with broad bi-partisan support, was that every child in every school would achieve grade level standards by the 2013-2014 school year (Hayes, 2008). Key provisions of the reauthorization included annual testing with the goal of academic improvement and a focus on teacher qualification designed to ensure each student would meet the high academic standards adopted by states (Vinovskis, 2009). Schools were to administer annual tests in language arts and math to each student in 3rd–8th grades and once in high school with the results disaggregated into eight groups and reported out publicly (Hayes, 2008). Schools were to reach their annual student performance goals in every group or they were considered to have not met Adequate Yearly Progress and sanctions were applied beginning after the second year of missed targets (Hayes, 2008; Vinovskis, 2009). Consequences included: allowing students to transfer out of schools "in need of improvement" (Hayes, 2008, p. 17) with transportation to the new school provided by the district, supplemental education services (including tutoring) chosen by parents, and ultimately corrective action including replacing the curriculum or school leadership, with a required restructuring or reconstitution as a charter school if there was still no improvement (Hayes, 2008; Vinovskis, 2009). In conjunction with school-level requirements, additional layers of teacher qualification were added under No Child Left Behind. Besides state certification, teachers had to meet additional

requirements to be designated as highly qualified including passing subject area tests and earning college degrees in specific areas (Hayes, 2008). The roots, however, of this school reform effort can be found in the history of public schools in the US.

In the beginning of the 1900s, the progressive movement dominated education throughout the United States. So called administrative progressives (Labaree, 2010) hoped to improve educational outcomes for students by focusing on reforming the systems surrounding the student and teacher and by providing vast quantities of data that teachers could use (Cubberley, 1916). Cubberley (1916) was prescient when he called for an efficient bureaucracy of experts who would examine, quantify, and evaluate schools, districts, classrooms and pupils in order to make recommendations for improvement. Although he was writing 100 years ago, his focus on quantification and standards as a way to measure school effectiveness could be found today in any number of news articles, opinion pieces, or Department of Education guides to best practice.

Cubberley (1916), an administrative progressive, described a need for new data clerks, record keeping systems, and departments and suggested hiring specialists who would study the problems facing schools and essentially figure it all out and provide specific direction for the teachers and school administrators. The administrative progressives believed their efficiency bureaus, record systems, measurement standards, and data clerks would professionalize school administration and describe school accomplishments “in a language which the community could easily understand” (p.

327), the language of numbers. This would free teachers up to focus on curriculum and instruction (Cubberly, 1916).

This vision of a data-driven focus on accountability for schools is a central theme of K-12 education today. School reform in response to inequities, a need to produce a better prepared workforce or to improve our nation's global competitiveness, or in order to ensure proper use of public funds are all reasons put forth to make the case for an accountability system. Accountability has been a primary driver affecting schools for the past few decades (Taubman, 2009). Education has been described as, "trapped in a language of schooling that stresses economics, accountability, and compliance" (Rose, 2009, p. 25).

While many have raised concerns about the negative effects of accountability on schooling (for example: Rose, 2009, Senechal, 2013, Taubman, 2009), there have been real, positive benefits for many traditionally underserved students. When schools are forced to confront the reality of the disparate opportunities available for students based on their race/ethnicity, language, socio-economic status, or disability, there are often changes in practices that positively affect those students. Attention is paid when a spotlight is applied to a problem or issue. But how did education get to the point where a quality school became synonymous with high student test scores? Where the purpose of an education seemed to be narrowly defined as passing a series of tests?

Over time in the United States, the prominence of the purposes of education for social cohesion and democratic equality have given way to those of social mobility and competition (Labaree, 1997; Robinson & Aronica, 2015). The emphasis on an

economic, market-driven model of education as it exists today has led to the development of a hierarchical structure for individual students as well as for schools and districts. If the goal of education for individual social mobility, moving up, is to be achieved, that individual must be able to move ahead of someone else. There must be winners and losers. In order for parents, as consumers, to advantage their own children, they must have a way to differentiate the relative merits of various educational opportunities. There must be so called good and bad schools, programs, and teachers so parents can choose the best for their children (Robinson & Aronica, 2015). The focus on competition in schools created a need for data generated in part to enable comparisons of different schools, districts, or teachers in a purportedly objective way. The data is often based on student achievement on standardized assessments, and has led to a belief that the only useful or true information about students is that which can be quantified, graphed, and reported numerically.

In contrast to this hierarchical, competitive view of education with winners and losers there is an oft-stated belief that all students are capable of achieving the highest levels of understanding and skill, that all students are capable of mastering the same content standards, and that it is the responsibility of schools to ensure this for each student (Mintrop, 2012). This belief has become a focal point for school accountability systems. Schools and districts are given ratings on the percentage of their students (as a whole as well as in groups disaggregated by race, socio-economic status, or participation in education programs) that are proficient as measured by standardized assessments. The idea is all students can learn, schools are held accountable for the

performance of their students, and the rating systems allow parents to make informed decisions about the best placement for their children.

School improvement and reform has largely focused on the how of education rather than the why. Changes in instruction, curriculum, materials, or the structure of the school day, have all been made in various attempts to identify effective ways of meeting the expectation that all students master high academic standards. A variety of approaches have been implemented nationally to address disparities in achievement and outcomes for students that in many cases can be predicted based on race, socio-economic status, or participation in programs such as Title 1, Special Education, or English as a Second Language/Bilingual Education (Robinson, 2015). Data is collected, analyzed, and shared to try to quantify the work teachers are doing, and to predict future student outcomes based on their current performance (Cowan, 2014). The importance of predictability of outcomes has become increasingly central to decision-making in schools. This emphasis comes straight from the top. During a visit to a charter school in New York, US Education Secretary Arne Duncan stated, "We should be able to look every second grader in the eye and say, 'You're on track, you're going to be able to go to a good college, or you're not'" (Hernández, 2009, para.7). But how do schools know which second graders will be going to good colleges?

The Effect of Systems of Accountability on Teachers

How do teachers decide which students need extra help and which will be fine in the classroom without different instruction or materials, extra time, or an

intervention from a specialist? What training or experience gives teachers the knowledge they need to make such important decisions? Do we even trust teachers to make those decisions? These ratings systems and focus on school and district accountability themselves have had unforeseen consequences for teachers, and the push to be able to accurately predict individual student performance is a different expectation altogether. Increases in levels of stress and perceived decreases in efficacy (Berryhill, Linney, & Fromewick, 2009) or agency (Robinson, 2012) are results of the systems of accountability and constraints of education policies outside of teachers' control. The effects of the accountability reform movement have been described as devastating and even a threat to the foundation of public education (Taubman, 2009).

Another unintended consequence of the accountability reform movement and the proliferation of student data has been a decreased trust in the ability of teachers to make professional decisions about students (Mausethagen, 2013; Robinson & Aronica, 2015). The increased scrutiny on teachers and schools, reporting requirements, and the availability of student assessment data has led to a reliance on data review processes when making decisions about students. Rather than utilizing multiple pieces of information about a student, teachers feel pressured to trust a number rather than their own instinct or judgment which is based on much more than results on a test or assignment (Rowe, Witmer, Cook, & daCruz, 2014). In no place is this truer than when schools identify students who might be at risk, particularly when considering a special education placement.

Response to Intervention

One method of identifying students who might be at risk is the Response to Intervention (RTI) framework. RTI is a framework integrating instruction and assessment in a multi-level system designed to maximize student achievement and identify at risk students early. Many districts across the United States utilize a Response to Intervention (RTI) framework of instruction including universal screening for individual student risk. In fact, 17 states require utilization of RTI data in some form for identification of students with Specific Learning Disability for special education services, six states require districts to submit an RTI plan as part of the special education process, four have established timelines for adoption of RTI, and an additional 12 states provide information within state regulations for districts that choose to utilize RTI (Hauerwas, Brown, & Scott, 2013). The National Center on Response to Intervention, or NCRTI, (2010a) describes RTI as a "prevention oriented approach" (p. 4) with essential components including: a school-wide system for preventing school failure, universal screening, progress monitoring, and data-based decision making. Levels of prevention (of failure), or tiers as they are often referred to in the field, designate various levels of intervention (National Center on Response to Intervention, 2010a).

The primary level or first tier includes core instruction that is research-based and is designed to meet the needs of 80% of students without additional supports. Instructional practices should be culturally and linguistically responsive to the unique needs of the students, but some districts purchase packaged programs that do not

match the cultural and linguistic background of the students they serve. Universal screening, at least annually but often 2-3 times per year, identifies students' current levels of performance and also serves to identify students at risk for not achieving grade level standards. In the primary level accommodations are integrated to ensure all students have appropriate access to instruction (National Center on Response to Intervention, 2010a).

Students whose universal screening results indicate they may be at risk for school failure are further assessed, according to the NCRTI (2010a), to determine if they would benefit from a more intensive, secondary level of instruction in addition to the core. This second tier typically involves small-group instruction utilizing evidence-based practices or programs designed to intervene early before students fall too far behind. Students participate in the intervention for a specified amount of time with regular progress monitoring to determine the effectiveness of the intervention. It is expected that this secondary level of prevention will benefit the majority of students who require additional intervention to the first tier or level of core instruction. Students may remain in the second tier, return to the core, or, if their progress monitoring results indicate a failure to benefit, may proceed to the third tier, a more intensive or individualized instructional program (National Center on Response to Intervention, 2010a).

According to the NCRTI (2010a), the tertiary level of prevention, or tier three, is the most intensive level of support with instruction targeted to an individual student's needs. More intensive intervention and more frequent (at least weekly)

progress monitoring allow the teacher to identify the student's rate of improvement over time. When the progress monitoring results indicate the student is unlikely to achieve the goal, the teacher modifies components of the intervention and continues monitoring to determine which components enhance the student's learning. This cycle of instruction, monitoring, and modification allows teachers to design effective instructional programs for individual students (National Center on Response to Intervention, 2010a).

Although RTI has been emphasized as a framework appropriate for all students, an additional purpose is to identify students with learning disabilities (Fuchs & Fuchs, 1998). The 2004 reauthorization of the Individuals with Disabilities Education Act (IDEA) allows RTI to identify children with Specific Learning Disabilities (SLD) for services rather than relying on the traditional discrepancy model (Speece, et al, 2011). In the RTI framework a student's response to instruction and interventions, along with screening and progress monitoring assessment results, would be considered during the evaluation for SLD (NCRTI, 2010a).

Universal screening. Within an RTI framework, or independently from it, schools implement universal screening of all students. The goal of universal screening is early identification of potential at-risk students through the utilization of brief assessment measures focused on target skills that are predictive of future outcomes (Hughes & Dexter, 2011). Universal screening three times a year for risk identification has been utilized successfully with academics such as reading or math (e.g., Ardoin & Christ, 2008; Jenkins, Hudson, & Johnson, 2007; Kilgus, Methe, Maggin, & Tomasula,

2014) or for behavior (Dowdy, Doane, Eklund, & Dever, 2011; Eklund, Rensahw, Dowdy, Jimerson, Hart, Jones, & Earhard, 2009; Greer, Wilson, DiStefano, & Liu, 2012). An analysis of 13 studies found that the use of universal screening had a positive effect on early reading and math (Hughes & Dexter, 2011). The screening measures are typically administered as an initial filter for identifying students who might be considered at-risk and would benefit from an intervention in order to improve their performance, catch them up to the other students in the class, or as an identifier for tier two in an RTI framework.

While there are many benefits, one of the challenges of universal screening is the misidentification of students. Floor effects, when large numbers of students score at the lowest level (Catts, Petscher, Schatschneider, Sittner Bridges, & Mendoza, 2009), atypical outcomes for students with different demographics (e.g., Hosp, Hosp, & Cole, 2011; Kilgus, Methe, Maggin, & Tomasula, 2014), or different results depending on the tool used to identify risk (Parker et al., 2015) all affect the predictive validity of a universal screening process. This increase in the likelihood of false positives, students identified as at risk when they actually are not, or false negatives, students identified as not at risk when they actually are, is a challenge that can have negative repercussions on students, schools, and systems.

If the purpose for universal screening is system evaluation, the misidentification of individual students is less important, but identification of which individual students are at risk is the primary purpose of many universal screening processes in schools (Curtis, 2012; Frontera & Horowitz, 1995; Kilgus et al., 2014;

Parker et al., 2015). Taking additional student risk factors into consideration when making decisions about intervention has been found to be a more efficient method of identifying at risk students (VanDerHeyden, 2013). This raises the question, if the screening tools are inefficient, and if additional input from teachers is required to maintain the predictive validity of a universal screening process for risk identification, why not just ask the teachers in the first place, particularly since they are usually familiar with the non-academic or out-of-school factors that contribute to or take away from learning?

Researchers have investigated the effectiveness of universal screening as a method of identifying students at risk (e.g., Greer et al., 2012; Hughes & Dexter, 2011; Parker et al., 2015; Rowe et al., 2014) but including teacher determination of risk as an aspect of the screening process was most common in studies related to behavior (Dowdy et al., 2011; Eklund et al., 2009; Greer et al., 2012) and rarely a consideration in studies of academic risk. Research on teachers' professional decision-making about students' academic risk has been very limited in scope. Curriculum-based measures have been extensively studied (for example Fuchs & Fuchs, 1998; Hosp et al., 2011; Kilgus et al., 2014; McGlinchey & Hixson, 2004; Parker et al., 2015), but the inclusion of teacher judgment was typically limited to teachers' abilities to match the specific score a student received on the assessment (Martin & Shapiro, 2011). Further, the use of computer adaptive assessments as universal screening tools in schools are fairly new and a review of published, peer-reviewed literature identified only a single study comparing the predictive and diagnostic accuracy of computer

adaptive and curriculum-based measures (Shapiro & Gebhardt, 2012). There do not appear to be studies that compare a teacher's ability to identify student risk with a computer adaptive assessment.

Assessment validity. It is important to consider the characteristics of effective screening tools used in a Response to Intervention or universal screening framework. Generally speaking, validity refers to the accuracy of the inferences that can be made based on the assessment results (Mellard, McKnight, Woods, 2009). In other words, whether or not the assessment is measuring what it is intended to measure, or construct validity (Messick, 1980). Validity, as it relates to screening tools used in RTI or universal screening, can be described in two other ways: criterion validity and consequential validity (Parker et al., 2015). Criterion validity refers to correlations between the scores that are generated by the tool and the variable of interest (i.e., year-end test scores, future academic performance, etc.) while consequential validity is an evaluation of the technical characteristics and impact of the tool as it is used (Parker et al., 2015). Consequential validity is a critical consideration in the practical application of screening tools. Diagnostic accuracy, or how well the data produced by the tool predict student proficiency including correct classification in a universal screening process, is important as the information is used to make decisions regarding student placement, allocation of teacher or curricular resources, or the need for further testing (Parker et al., 2015). The sensitivity, or accurate identification of students at risk, and specificity, or identification of students truly not at risk, affect the practicality and

appropriateness of the screener when put in practice (Lembke, McMaster, & Stecker, 2009).

Problem Statement

The purpose of this research was to determine if, in the context of a Response to Intervention framework, teachers' professional judgments were equally predictive at determining risk level as the data provided by the screening tools in common use in school districts to identify students who would benefit from a reading intervention. The study examined two research questions: (a) what is the relationship between teacher judgment of student reading risk levels and the screening tool risk levels, and (b) are there variations in the relationships related to student characteristics including identification as having limited English proficiency, receiving special education services, race/ethnicity, gender, or being economically disadvantaged?

Significance

This study is significant because it fills a gap in the literature by examining the accuracy of teachers' professional judgment about reading risk as compared to a computer adaptive screening tool being used by more than 46,000 schools (Renaissance Learning, 2016). Taking this gap in the literature into consideration, this study examined teachers' ability to identify which of their students were at risk in reading based on their own knowledge of their students without the benefit of a universal screening assessment and found that generally teachers were able to match the risk levels determined by the universal screening tool. Universal screening is widely used in schools across the United States, and the tools can be expensive. This

study suggests that teachers are mostly able to identify the same students as the universal screening tool using information gained through their interactions with students and other adults. If teachers are able to accurately identify which students are at risk, there may be less need to purchase the expensive screening assessments for every student. The time spent testing, and reviewing the results, for every student could instead be used focused on instruction

Summary

This chapter introduced the need for a study examining the efficacy of teacher judgment about their students' reading risk in a Response to Intervention framework. In the national context of an increasing emphasis on accountability, the use of data, and standardized testing, teachers' judgments are not always valued and teachers themselves sometimes doubt their own abilities to make decisions about their students. Instead, systems have been designed which encourage relying on the results of screening tools or standardized tests sometimes without even reference to teacher professional judgment. Chapter 2 includes a review of the literature related to ways of knowing, risk determination, computer adaptive testing, and the STAR Reading assessment, Chapter 3 describes the study including the purpose, the researcher, the subjects and setting, instrumentation, the research design and procedures, data analysis, limitations, and ethical considerations, Chapter 4 includes the results of the analysis, and Chapter 5 concludes with a discussion of the implications and recommendations for further study.

Chapter Two: Literature Review

This literature review will first discuss the theoretical framework for this study, tacit knowing as conceived by Michael Polanyi. Next, it will explore the various ways of knowing teachers utilize in their interactions with students categorized by personal knowledge, aesthetic knowledge, emancipatory and ethical knowledge, and empirical knowledge. Third, the chapter will discuss the literature on risk determination in schools. Finally, the literature review will describe the development and use of computer adaptive tests, with a particular emphasis on the STAR Reading Assessment.

Theoretical Framework: Tacit Knowing

Michael Polanyi was a Jewish scientist and philosopher born in Budapest in 1891 (Mead, 2007). He earned a medical degree and a PhD in physical chemistry from the University of Budapest and served as a medical officer in the First World War (Nye, 2015). Later, he became director of a research group in Berlin (Mead, 2007) and was recognized as a renowned expert in the field of chemical kinetic research (Nye, 2015). He resigned his position in protest of the rise of Hitler and the Nazi party's anti-Semitism and hostility towards minorities (Mead, 2007). Polanyi moved to England where he continued his work at the University of Manchester in the physical chemistry laboratory. In 1948 the university created a position as chair in social studies specifically for Polanyi and the focus of his worked changed (Nye, 2015). During the Second World War, Polanyi assisted friends and family to escape the Nazis but one of his sisters and other friends and family members were victims of the

Holocaust (Mead, 2007). His work as a chemist was highly acclaimed, but his experiences during the war and his observations of the effect of ideology on scientific thinking and the treatment of scientists in the Soviet Union made him believe that he could make a more effective contribution by focusing on critical issues through the lens of philosophy rather than chemistry (Mead, 2007; Nye, 2015; Polanyi, 1966b). So Polanyi began writing and lecturing on the nature of knowledge, levels of knowing, intuition, skills, and performance (Nye, 2015). In 1958, after 10 years of writing, he published Personal Knowledge: Towards a Post-Critical Philosophy. Now considered a classic, at the time it was written it was highly criticized for its opposition to mainstream thinking about the philosophy of science (Nye, 2015). The following year he was elected as a Senior Research Fellow at Oxford and for more than a decade he published and lectured in the United States, Great Britain, and Europe (Mead 2007).

Michael Polanyi (1961, 1966a, 1966b) wrote about the nature of thought, meaning, personal knowledge, and tacit knowledge. He took aim at science based on rigid empiricism and logic, instead offering a model in which scientific questions and breakthroughs are derived from the tacit knowledge scientists gained from their experience and training (Polanyi, 1966b). His work has influenced thinkers in sociology, political science, psychology, economics, education, and even theology (Nye, 2015). For the purposes of this paper I will focus on Polanyi's thoughts on tacit knowledge, attempting to make connections to the ways teachers understand and make decisions about their students.

Polanyi (1966b) defines tacit knowing using the phrase "we know more than we can tell" (p. 4). By this he means that a person's knowledge is composed not only of the things he or she can describe explicitly, but also a tacit component that is difficult, if not impossible, to define. As an example, Polanyi uses the recognition of a person's face. When looking at a face, a person can tell immediately whether it is a friend or a stranger. However, the person would be unable to pinpoint precisely why this one face is that of a friend instead of a stranger. If asked, the person could perhaps describe certain aspects of the face, the size of features, maybe eye color or distinguishing marks, but these characteristics could as easily describe an entirely different person. But, if shown a photo of a different someone with those same characteristics just described, the person would immediately know it was a stranger rather than the friend. Only so much of the knowledge we have can be put into words (Polanyi, 1966b).

In the same way, teachers have layers of knowledge about their students, some aspects explicit and describable and others tacit. The totality of what a teacher knows about his/her students cannot be described. There is a component that is understood based on the teacher's experiences both with the particular student as well as all of the students who have come before (Coleman, 2014). A teacher could have two students in his/her class, both with identical test scores, demographic characteristics, attendance rates, yet the teacher recommends additional assessments or instruction for only one. Why? If asked, the teacher may be able to describe why the one student needs more assistance, but that explanation may not make sense to the outsider. Without the tacit

knowledge of the teaching/learning context and the students themselves the teacher's decision may seem capricious or nonsensical. Only so much can be put into words.

Polanyi (1966a) describes knowing as consisting of two aspects. The first, the distal, we can know and name. The second, the proximal, we know of only because of the effect it has, its impact. Polanyi (1966b) calls this the "phenomenal structure of tacit knowing" (p. 11); we are aware of the proximal only because we can recognize and name the distal. The proximal is that aspect we know and cannot tell. This knowledge is not necessarily subconscious, but it is subsidiary to the conscious, explicit aspect of knowledge (Polanyi, 1966b). An example is riding a bike. There are specific actions such as pedaling or maintaining balance, that are explainable (distal), but there are other aspects of riding a bike that are just as important to master, but are ineffable. How does one pedal and maintain balance at the same time? How does an expert bike rider know precisely when to begin braking in order to stop in time? Both aspects of knowledge are needed in order to proficiently ride a bike. Those ineffable aspects are the proximal and are recognized when the person successfully rides down the street (Polanyi, 1966a). Additionally, attention to a specific component part, focusing too much on pedaling without attention to balance for example, impedes the integration of the whole, causing a crash even though the rider was pedaling perfectly (Polanyi, 1962).

Another way to think about the dual aspect of knowledge is the integration of the distal and the proximal into a coherent whole (Polanyi, 1966a). This is different than thinking about something as composed of separate parts that are only understood

when they are put together, like a puzzle. Instead, Polanyi describes tacit knowing as the proximal and distal being separate aspects brought together as a whole and only understood completely when both are taken into consideration (1966a). The act of understanding a complete entity consists of alternating attempts to focus or concentrate back and forth between the parts and the whole (Polanyi, 1961). These efforts are complementary, and both are needed to some extent in order to develop a full understanding. However, focusing on the particular weakens the sense of coherence, and when attention moves outwards towards the whole "the particulars tend to become submerged" (Polanyi, 1961, p. 460).

This description of the dual aspect of knowledge and developing understanding by alternatively focusing on both the part and the whole can be helpful when attempting to understand how teachers make decisions about students. It is somewhat like practicing medicine. Both teachers and doctors identify symptoms and make professional judgments regarding the meaning of the symptoms and suggest treatment options based on their background, training, and tacit knowledge (Garcia & Ford, 2001). Just as two very different diseases might present with the same symptoms, two students might have the same test scores, but the treatment could be quite different. The skilled practitioner might decide student A would benefit from an intervention while student B simply needs more time to develop a skill or deepen understanding. These decisions are based on an accumulation of factors or particularities that, when seen as a whole create two very different images of the students' potentials (Wansart, 1995).

Teacher understanding or judgment of student risk consists of discerning the patterns inherent in the student's family life, previous educational experience, participation in class, performance in classroom activities, friendships with others, innate abilities, developmental levels, and recognizing that which cannot be understood based on the sum of the pieces (Gurm, 2013).

As Polanyi (1961) states, the understanding of a whole is based on the understanding of how the parts come together in a particular way, perhaps differently than how the same parts come together in a different way in a different context (or for a different students). When something is broken down and one focuses on its component parts the whole loses cohesion (Polanyi, 1961). This can be compared to making educational decisions based solely on test data and ignoring those variables that teachers can observe and factor into decision-making. Comprehension may be enhanced by understanding the parts; there may be a deeper appreciation of the whole based on knowing of the individual components. However, it is dangerous to assume that simply understanding the details equates to a complete understanding of the whole.

In order to be competent, to fulfill their professional responsibilities, teachers focus on the whole student. Although individual muscle movements are not perceived separately when dancing, when one knows a dance it comes naturally. However, if a dancer pulled a small muscle and could no longer waltz without pain, he or she would recognize something was wrong. In the same way, teachers might not be able to describe each individual component that allows them to know a student is on track, but when something is wrong the teacher intervenes to correct the issue. In contrast,

standardized tests focus on the particular or specific. Screening tools are designed to describe or measure individual components of knowledge and, based on the responses and trajectories of many students, predict future performance or identify students at risk. The standardized test does not know that a student performed poorly on a particular day because his/her dog died, but the teacher would recognize the context of the student's personal situation and allow for the possibility that the score is not necessarily reflective of the student's ability. Similarly, teachers' abilities to understand the motivational systems of their students, to understand how motivation is connected to their environmental opportunities, allows them to better understand each student's abilities and help them to enhance their creativity, productivity, and happiness (Kenrick, Griskevicius, Neuberg, & Schaller, 2010).

When a teacher recognizes risk or lack of risk by beginning with knowledge of the whole child, the teacher applies meaning to the part, the test score, differently than decision rules for a universal screening process which apply risk based on statistically derived predictions of certain outcomes. Teachers and tests start from different places in order to make sense or determinations about students. Teachers begin with an image of the whole child and then make meaning of the test score based on their knowledge of the child. Tests begin with a whole that consists of many children from many different backgrounds and make meaning of the test score based on the knowledge of the statistical probability of certain outcomes based on many students scoring in particular ways or how the student scores in relation to a standard on a criterion referenced test. The individual characteristics of the student are not relevant when the

purpose of the test is to determine how the student performs in relation to other students or a standard. In an education system designed around standardization and predictable outcomes for students, it is not surprising that the expectation is for students to conform to an ideal and individualization to be suppressed (Robinson & Aronica, 2015).

Ways of Knowing

The concept of tacit knowing was Polanyi's (1966b) attempt to describe how a person's knowledge has a component that is difficult, if not impossible, to define. Professionals, such as teachers, make decisions based partially on knowledge they cannot always express. Others have defined the ways of knowing differently. Teaching is described as a holistic (Szesztay, 2004), transformative (Yorks & Kasl, 2006), even artistic (Conklin, 1970) endeavor. Borko and Shavelson found that teachers make up to 1,500 decisions every day (as cited in Cuban, 2011), each of which affects their students. As professionals making these decisions about students, teachers draw from all the different types of knowledge they have in order to make the best possible choice among many options. Theorists have attempted to capture and explain these teachers' ways of knowing in multiple studies.

Gurm (2013) categorized the types of knowing teachers have and use to perceive and understand their environment and students in five ways: personal, aesthetic, emancipatory, ethical, and empirical. Personal knowing is based on the authentic relationship between teacher and learner, aesthetic is the art of knowing by doing, emancipatory is knowing the learner in the context of environment and history,

ethical knowing is the moral knowledge of teacher conduct in their roles, and empirical knowing consists of the things that can be seen, heard, or touched (Gurm, 2013). The empirical knowledge or scientifically determined explanation is only one piece of the whole range of knowing, teachers also access their tacit knowledge based on the other categories Gurm (2013) described in order to understand their students and their students' performance. These ways of knowing are described in more depth below.

Personal knowledge. Personal knowledge is gained from the relationship between teacher and student. Teaching is an active process that involves reflection and adjustment to the needs of the students. Teachers observe an action in the classroom, hear a student discussing a topic, or read a piece of writing and must make a decision about how to respond. The response could be very different depending on the relationship between the teacher and student as well as the teacher's knowledge of the student. The interpretation of the action, discussion, or writing is dependent not only on the item in question, but on the context and background of the student who produced the item and the teacher's relationship with the student (Barbour, 2004). The reflection can take place nearly simultaneously with the response and involves a component of intuition or implicit knowing about the relationship between the student, the classroom, the teacher, and the educational objectives (Szesztay, 2004).

The idea that knowledge is contextual (Barbour, 2004; Szesztay, 2004) is helpful in understanding how teachers make decisions about students. Barbour (2004) described this as an embodied way of knowing which explicitly recognizes that

individual differences impact how something is known or understood. Who is doing the knowing matters. Different teachers understand children in different ways depending on the teacher's background and experiences in relationship with those of the student and the school and community. Teachers are participatory, not objective in the decision-making process and the ways teachers know their students change depending on the teacher (Barbour, 2004).

Conklin (1970) argued that knowing and teaching are best appreciated or criticized in the same way as artistic endeavors, what he describes as aesthetics. However, his description of aesthetics fits better as a part of personal knowledge rather than aesthetic knowledge in this organizational structure. According to Conklin (1970), knowledge is more visible and measureable when it is on display for others to appreciate and understand than when it is in the mind of the knower, and the act of knowing is even less visible than the knowledge in the mind. When a teacher attempts to communicate an idea or piece of knowledge to students, he/she is trying to make the invisible knowledge visible and on display. Conklin (1970) described teaching as a form of communication, the teacher communicates ideas and knowledge to the student as a way to make the knowledge in his/her mind on display for another. While there are ways to measure this visible display of knowledge, knowing is intensely private, "only the knower can know whether he knows" (Conklin, 1970, p. 259). The relationship between the teacher and student affects the nature of the knowledge transfer as well as the knowing itself. The teacher is participatory in the learning process, and master teachers utilize methods appropriate to the students, subject matter,

and the personality of the teacher. The context and relationship affect the knowledge and the knowing.

An individual student's education takes place within a particular point in time. This history has an effect on the student, the school, and the society at large (Simpson, 1971). Changes in society mean changes in individuals. Context matters. Societal changes bring changes to students and teachers, which disrupts and modifies the school. Teachers understand this history and consider this context as part of their understanding of students, which informs their decisions (Simpson, 1971). Teachers are aware of the impact of social emotional factors on student learning because they inhabit the same space and time as their students. Teachers take account of the emotion and engage their feelings and those of their students to foster transformative learning (Yorks & Kasl, 2006).

Aesthetic knowledge. Teaching is an active profession, and teachers gain aesthetic knowledge through the act of teaching their students. Although an educated person may not have explicit recall of the facts and figures, or may not be able to pass a test years after completing a course of study, the value of an education is still present in the lens through which the educated individual views the world (Broudy, 1979). Teachers view their students, their professional world, through the lens of their experience as educators and their experience teaching the specific students in their class. Teachers have various levels of training and experience, professional education or development which have allowed them to develop the ability to implicitly know things about their students that they may not be able to logically defend. A teacher

may be explicitly aware of specific information about a student, but the teacher's intuition is distinct in that it conveys information separately from logic and the accumulation of information (Garcia & Ford, 2001). The teachers may not know why or how they know something, but their actions teaching students have given them aesthetic knowledge. They know more than they can say (Polanyi, 1966b).

Tacit knowledge is the implicit understanding gained from previous experience and learning that may not be able to be explicitly explained or even recalled (Broudy, 1979). The purpose of an education is to "fund the mind explicitly with contents that in time will become tacit resources for the building of contexts" (Broudy, 1979, p. 452) Teachers who have completed courses of study and who have perhaps spent years honing their professional expertise perceive the classroom and students through the lens of these tacit resources. While studying to become licensed, attending trainings, practicing and applying new skills, teachers explicitly learn how to assess student abilities. The intuition provides valuable information that can give teachers insight into how to proceed with instruction or an intervention for a student (Garcia & Ford, 2001). Accepting this description of tacit knowledge, it would follow that as their abilities as teachers develop, they begin to establish an aesthetic understanding of their students' abilities.

Engaging holistically with students in the learning environment allows teachers to understand their students at a deeper level. This intuitive knowledge is based on sensation and perception (Yorks & Kasl, 2006) and allows teachers to empathize with their students. The tacit (Broudy, 1979; Coleman, 2014) or ineffable (Conklin, 1970)

knowledge is applied subconsciously when teachers make decisions about students, but they are not always able to describe why they know what they know. This ineffability means the expert (the teacher) is unable to explain or communicate the knowledge to in a way that non-experts (the others) can understand (Broudy, 1979; Conklin, 1970). The holistic nature of teaching, teachers drawing on all of their skills, knowledge and intuition at once, leaves little time for reflection in the moment in order to make decisions. There is an immediate connection between noticing and doing, the immediacy of teaching (Szesztay, 2004), that is a thoughtful response to what is happening, but the reasoning cannot be readily described to the outside viewer, the non-expert.

Likewise, teacher knowing is recursive; the teacher continuously accesses the information he or she has about the students, their backgrounds, and the class structure or goals in order to engage in a continuous process of reflection and modification of action (Coleman, 2014). The act of teaching is a reflection of knowing. Observation and response during teaching are the tangible indications of the knowing teachers have about their students (Wansart, 1995). This aspect of knowing in action is experiential and is as valid an aspect of knowing as empirical or rational knowledge (Barbour, 2004).

Emacipatory and ethical knowledge. As society changes and individuals within the society gain or lose stability, as their basic needs are met or not met, the personalities of the individuals undergo a change, which then affects the social structures (Simpson, 1971). The school environment shifts depending on the

personalities of the individual teachers and students within the environment. As the demographics of students change in schools, as more and more students experience less and less stability, the social structures change. Recognizing the context students are living in, understanding the differences between students, is a critical form of knowledge that allows teachers to establish appropriate learning goals that are situational and unique to the learners (Simpson, 1971).

Wansart (1995) understood the teacher in action as a researcher, reflecting on and responding to the students, the school, and the classroom work. This knowing in action is a strengths-based approach, starting with an attempt to understand what a student can do and what he or she knows rather than what he or she cannot do and does not know. It is a reflection of emancipatory knowing because this focus on student ability is based on understanding learning from the students' point of view rather than the teachers' or the schools' cultures (Wansart, 1995). Additionally, it is more effective to build one's own knowledge through the reflection/response process of teaching rather than taking and applying the knowledge of experts (McConaghy, 1986).

The standardization movement has devalued the non-western, non-English language world-views of many students and pushed teachers into an education model based almost entirely on empirical knowledge (Diaz Soto & Tuinhof De Moed, 2011; Robinson & Aronica, 2015). This cognitive imperialism (Diaz Soto & Tuinhof De Moed, 2011, p. 329), or the imposition of a single, westernized, way of thinking, ignores the languages, cultures, and experiences students bring with them to the

classroom. Teacher knowledge of how or whether their students' cultures and languages match those of the majority culture inform their understanding of the students in the context of their school environment.

The application of empirical pedagogical and content knowledge in tandem, pedagogical content knowledge, is what distinguishes veteran teachers from beginners (Gudmundsdottir, 1991). This pedagogical content knowledge is a type of tacit knowledge that teachers apply in the classroom in order to effectively instruct their students. While it is based on empirical understandings, the application itself is more implicit than explicit. It is the lens through which teachers view and make decisions about their students (Gudmundsdottir, 1991).

The teacher in the learning environment with the students, interacting in a holistic way, has access to a more complete version of the student than an assessment result. Context matters. An individual cannot be taught (or assessed) without consideration of and understanding previous experiences and the effect of those experiences on the individual (Simpson, 1971). In the same way, pieces of information (data) must be considered as part of a whole when making decisions. The pieces can be examined individually but this tends to result in losing awareness of the whole (Polanyi, 1961). To be understood, the data must be considered as part of the entire context (Conklin, 1970). Scholarly disciplines, such as teaching, create a structure in the learner that is composed of the particular facts, concepts, and ways of thinking about the world that are distinct from other disciplines. This structure remains even if the specific facts or lessons have been forgotten (Broudy, 1979).

Empirical knowledge. Empirical knowledge or understanding is quite different from the implicit or tacit knowledge teachers have regarding their students. This type of knowledge is based on sensory information and suggests what something is rather than an impression of what it is like (Garcia & Ford, 2001). Aspects of empirical knowledge are the teacher's content and pedagogical knowledge (Gudmundsdottir, 1991). Standardized test results are other forms of empirical knowledge commonly used by teachers when making decisions about students. This type of empirical knowledge is what is often used to determine whether a teacher's judgments about his or her students' academic abilities or achievement are valid.

Risk Determination

Schools and teachers access different ways of knowing when determining which students are at risk. Begeny, Krouse, Brown, and Mann (2011) described that teachers' judgments of their students' academic achievement were critical to understand due to the fact that teachers make decisions about their students each day related to "instructional materials, teaching strategies, and student-learning groups" (p. 23) and that the judgments influenced teachers' expectations, the interactions between teachers and students, and even student outcomes. Issues such as eligibility for Special Education services or which students will receive additional support or participate in small group instruction or interventions are also highly influenced by teachers' determination of how much risk each student is at for academic failure or poor achievement (Frontera & Horowitz, 1995; Hoge & Coladarci, 1989; Martin & Shapiro, 2011). Given the importance of teacher judgments of student academic risk, it would

seem likely there would be a large body of research on the accuracy of the determinations, but that is not the case.

A 1989 literature review conducted by Hoge and Coladarci included studies in classrooms at many grade levels and content areas as an examination of the extent to which a teacher's judgment corresponded to the students' achievement. The researchers found only 16 published studies that were based in natural settings (no simulations) in which the teachers' judgments about their students were compared with student data collected concurrently. They found a median correlation of .66 between teacher judgment and the criterion measure when looking at the results of all of the studies suggesting teachers are able to accurately match the results determined by an assessment. However, the range for the correlations for all of the studies was .28-.92, which suggests variance in the ability of individual teachers or for different groups of students (Hoge & Coladarci, 1989). The authors encouraged further research of this type to determine if there were patterns to the differences they found between the abilities of certain teachers to judge student achievement and if there were differences based on grade levels or subject matter (Hoge & Coladarci, 1989).

Three studies compared teachers' ratings of their students' risk with reading assessments to see which was more accurate at predicting low reading achievement or identification of learning disabilities later (Fletcher & Satz, 1984; Frontera & Horowitz, 1995; Kenny & Chekaluk, 1993). The results, however, were mixed. One study of 312 children in Australia found teachers were better than psychometric tests

of language and reading at predicting students who were on track rather than at risk (Kenny & Chekaluk, 1993). A study of 571 students in Florida found the opposite in a longitudinal study: teachers were more accurate than the four criterion-based tests given in kindergarten at identifying students at risk in 2nd grade (Fletcher & Satz, 1984). The third study determined that teachers' assessments of 57 students' reading levels were almost equally predictive as reading achievement test scores (Frontera & Horowitz, 1995).

Two studies looked at the issue of teachers' reading risk determinations compared to standardized reading assessments in a slightly different way. Rather than looking at the accuracy of the predictions based on identification of a learning disability or reading achievement difficulty, these studies compared teachers' ratings with other measures to determine if there was predictive validity of year-end performance on summative assessments (Kapelis, 1975; Payne & Payne, 1991). In the first study by Kapelis, after six weeks of instruction, 1st grade teachers were asked to predict end-of-year reading achievement for their students. The forecasts were compared with two standardized tests administered after two weeks, the Meeting Street School Screening Tests and Slingerland's Pre-reading Screening Procedures, and found the tests had slightly better abilities (.62 and .68 correlations) than teachers (.48) to predict year-end reading achievement (Kapelis, 1975). The second study asked 36 K-5th grade teachers after several weeks of school to determine which students were academically at risk then compared the teacher predictions with free/reduced lunch rates, grade retentions, and standardized test scores to see if there were

correlations (Payne & Payne, 1991). The researchers found moderate correlations with teacher judgments (.40 for the Iowa Test of Basic Skills Reading) for all students, but quite different correlations when the results were disaggregated by student race; there was a .51 correlation for White students and .28 for Black students (Payne & Payne, 1991). The authors suggested test bias or the fact that there were high correlations between teacher identification of risk with free/reduced lunch rates (and a much higher percentage of Black students qualified) could account for the discrepancy, but it is also possible teacher bias played a role as 27 of the 36 teachers included in the study were White.

An alternate model for calculating student risk takes into consideration individual student contextual factors (VanDerHeyden, 2013). Due to the innate error rate in the screeners, students who are closest to the cut scores are the most difficult to correctly identify and in some systems the students are at such higher risk altogether (for example, large numbers of English learners, students in poverty, highly mobile populations) that the errors associated with the screener itself make the universal screening process mathematically inefficient for identifying individual student risk (VanDerHeyden, 2013).

Martin and Shapiro (2011) examined the accuracy of teacher judgment of early literacy skills in kindergarten and 1st grade students. The authors asked teachers to consider their classroom assessment of literacy and use that information to sort their students into low- and at or above typically-achieving. They then compared these results to the students' scores on the Nonsense Word Fluency (NWF) and Phoneme

Segmentation Fluency (PSF) measures of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). They found a moderately strong ($p < .01$) correlation between student scores and teacher predictions for each of the measures. In other words, teachers were able to predict whether or not their students would be considered at-risk in the area of reading as measured DIBELS. Teachers were also asked to choose one low-achieving and one typical- or higher-achieving student and to attempt to predict the NWF and PSF scores. While there were moderately strong correlations between students' scores and teachers' predictions, Martin and Shapiro (2011) examined the differences between the predictions and the actual scores and found consistent overestimation by the teachers of the actual performance of these targeted students, particularly for the NWF.

Begeny, Krouse, Brown, and Mann (2011) interviewed 27 teachers in 1st-5th grades and asked them each to estimate the reading performance for eight of their students on Word Correct Per Minute (WCPM) on two grade-level reading passages and Language Arts scores on the Palmetto Achievement Challenge Test (PACT). The researchers found moderate relationships between PACT (.58, $p < .01$) and WCPM (.51, $p < .01$) estimates and the students' actual performance, but in general teachers were more accurate at judging their high performing students than low or average performing students (Begeny, Krouse, Brown, & Mann, 2011).

The accurate identification and assessment of academic ability has always been the goal of test developers. Fixed form tests generally do well at discriminating the differences between students in the middle of the range of possible scores (Finnerty,

2015). Issues such as the floor effect (student results clustered at the absolute bottom of the scale) or ceiling effect (student results clustered at the top of the scale) make it difficult to differentiate students at the extremes (Finnerty, 2015) leading researchers and educators to look for new ways, such as using computer adaptive tests, to accurately assess students.

Computer Adaptive Tests

One of the issues with the universal screening process for young students is that they are developing the skills they are being assessed for at the same time as the assessment. It can be difficult to accurately predict student risk because it is difficult to discriminate between a low performance that is based on a student at risk and low performance based on a student who has not yet been introduced to a concept or skill. In a sense the assessment is trying to hit a moving target (Speece, 2005). One consequence of this concurrent development and screening, particularly for the youngest students, is the presence of a floor effect. When large numbers of students score at the lowest level it can affect the predictive validity of screening tools meant to identify students at risk (Catts, Petscher, Schatschneider, Sittner Bridges, & Mendoza, 2009). Each test used to assess student skills or to predict future performance contains certain error rates. False negatives (students rated as on track when they actually are at risk) or false positives (students rated as at risk when they actually are not) are ongoing issues in any universal screening process in place in schools. Attempts to develop tests that allow greater diagnostic accuracy have led to adaptive tests that more closely assess each individual student's abilities.

Computer Adaptive Tests (CAT) are those in which the ability or competence of the test taker is assessed after each response and the difficulty of the proceeding item changes depending on whether the previous response was correct or incorrect (Gershon, 2005). Computer algorithms are designed to identify appropriate items from a large pool and offer them based on the current estimate of the test-taker's ability (Finnerty, 2015; Gershon, 2005). CAT assessments have been increasingly found in education following the development of desktop computing in the 1980s which gave entities smaller than governments or military the computing power necessary to develop and offer the assessments (Finnerty, 2015).

CAT assessments are based on targeted or adaptive tests such as the Stanford-Binet IQ Scale in which the human test administrator adjusts the items that are delivered based on the subject's responses (Gershon, 2005). Linear paper-based or computer-based assessments develop reliability by selecting questions of average difficulty and have high precision for the middle-range while CAT assessments have higher precision at the extremes (in other words, they are less likely to suffer from a floor effect than non-adaptive assessments) because only appropriately difficult items are administered (Finnerty, 2015; Gershon, 2015). Research studies have repeatedly found no difference between the performance or comprehension of test subjects on paper-based versus CAT versions of the same assessments (Finnerty, 2015; Gershon, 2005; Noyes, Garland, & Robbins, 2004). One study did find performance differences when reading passages were longer, particularly for lower-scoring test takers,

suggesting there may be increased cognitive demand based on the perceived difficulty of reading on the computer (Noyes, Garland, & Robbins, 2004).

While curriculum-based measures (CBM) have been extensively researched (e.g., Deno, 1985; Good & Kaminski, 1996; Kilgus, Methe, Maggin, & Tomasul, 2014; McGlinchey & Hixon, 2004; Parker et al., 2015), computer-adaptive assessments are newer and there is less research on their use in a Response to Intervention framework or for universal screening. The National Center on Response to Intervention (2010b) found convincing evidence for the accuracy of the classifications provided by the Measures of Academic Progress (MAP) for Primary Grades test based on a study of 4,659 K-2nd grade students comparing their MAP predictions with results of the Wisconsin Knowledge and Concepts Test. A study on kindergarten students found stronger relationships between the STAR Early Literacy CAT and performance on a summative state assessment when compared to a CBM (Clemens et al., 2015). A similar study found similar results in math with the STAR Math CAT for 3rd and 4th grade students (Shapiro & Gebhardt, 2012).

STAR Reading Assessment

According to the Technical Manual provided by the publisher, Renaissance Learning (2015), the STAR Reading assessment is a computer-adaptive, group-administered measure of reading comprehension. STAR Reading serves three purposes: providing a quick estimate of reading comprehension, assessing reading relative to national norms, and providing a means for consistently tracking growth over time. The assessment has changed and developed over time, the second

generation of STAR Reading was based on Item Response Theory and fixed the length at 25 items (Renaissance, 2015, p. 4). In 3rd grade and above, the test consists of 20 vocabulary-in-context questions and 5 authentic text passages followed by literal or inferential multiple-choice questions. The vocabulary-in-context questions require students to read and interpret the meaning of cloze sentences and to choose the most appropriate of four vocabulary words that best completes the sentence based on the context. The student's performance on the vocabulary-in-context section is used to determine the initial level for the authentic text passage items. The authentic text passages are drawn from children's and young adult literature, nonfiction books, newspapers, magazines, and encyclopedias and are leveled for each grade. The assessment is untimed but it is estimated to take approximately 15 minutes. Students complete the assessment online and their scores are automatically generated and recorded in the data system (Renaissance Learning, 2015).

Due to the adaptive nature of the test, the content of the STAR Reading assessment varies from one administration to another and with each student's performance. The technical manual (Renaissance, 2015) reports a generic reliability of .93 and a test-retest reliability of .85 for both 3rd and 4th grades overall. The assessment is based on Item Response Theory (IRT), which allows the degree of measurement error to be determined for each individual test (Renaissance, 2015). STAR Reading provides what are called "conditional standard errors of measurement (CSEM)" (Renaissance, 2015, p. 50) for each individual test score. These CSEMs are estimates of the reliability of individual scores. The CSEM will vary, potentially

substantially, from one student's score to another. The average CSEM for 3rd grade is 41 with a standard deviation of 15.4 and for 4th grade, an average of 50 with a standard deviation of 19.4 (Renaissance, 2015, p. 57).

The Center on Response to Intervention at American Institutes for Research (2014) found convincing evidence of the validity of the STAR Reading assessment. Reviews of several studies found concurrent validity between STAR reading and DIBELS or other CBMs in the range of .72-.82, and predictive validity to the Stanford Achievement Test 9 (SAT9) as well as a number of state end or year assessments ranging from .68-.82. When the results were disaggregated by race, the predictive validity for Hispanic students ranged from .55-.74 and for White students from .69-.75. The disaggregated generic reliability for all students ranged from .87-.94. Based on reviews of a number of independent studies as well as two large-scale studies incorporating over 100,000 students in seven states, the STAR Reading assessment received the Center on Response to Intervention's highest possible ratings in every category including classification accuracy, generalizability, reliability, and disaggregated reliability and generalizability (2014).

Summary

This chapter introduced the theoretical framework, tacit knowing, for the study described in chapters 3-5. Research on the various ways of knowing: personal, aesthetic, emancipatory and ethical, and empirical, were applied to the work of teachers. Next was a discussion of the literature base on risk determination in schools. The chapter concluded with research on computer adaptive tests and a detailed

description of the STAR Reading assessment, which was used as the criterion measure in this study.

In the area of reading there have been very few studies comparing teachers' judgment of students' ability with a criterion measure and, as Hoge and Coladarci (1989) discovered, even fewer that concurrently compare teachers' determination of students' reading risk with a standard reading risk assessment measure. There have been some that examined teachers' abilities to predict student outcomes. Certain studies comparing teacher predictions of reading achievement with year-end summative evaluations included only particular groups of students such as White, male kindergarten (Fletcher & Satz, 1984), White 1st grade, (Kapelis, 1975), Hispanic 4th grade (Frontera & Horowitz, 1995), or only English speaking kindergarten, 1st, or 2nd grade students (Kenny & Chekaluk, 1993). Others were based on a limited number of students (Frontera & Horowitz, 1995; Martin & Shapiro, 2011) or from students in a single school (Frontera & Horowitz, 1995; Martin & Shapiro, 2011; Payne & Payne, 1991). These types of limitations, coupled with the small number of studies actually available, make it difficult to generalize the results. This research study addresses this gap in the literature, attempting to determine whether or not teachers have the ability to identify which of their students are at risk in the area of reading utilizing their tacit professional knowledge and the different ways they know their students.

Chapter Three: Methods

This chapter includes information on the purpose of the study, the researcher, study subjects and setting, instrumentation, procedures, data analysis, limitations of the study, and ethical considerations.

The Purpose

A review of the literature revealed a gap in the research related to teachers' abilities to identify individual student risk in reading. Many districts across the United States utilize a Response to Intervention (RTI) framework of instruction including universal screening for individual student risk. In fact, 17 states require utilization of RTI data in some form for identification of students with Specific Learning Disability for special education services, six states require districts to submit an RTI plan as part of the special education process, four states have established timelines for adoption of RTI, and an additional 12 states provide information within state regulations for districts that choose to utilize RTI (Hauerwas, Brown, & Scott, 2013). Universal screening is becoming more and more common and districts are investing time, effort, and funding in purchasing tools and establishing screening protocols. And yet, there does not seem to be research on whether or not teachers are able to identify their students' risk without the assessment tools or processes.

The purpose of this research was to determine if, in the context of a Response to Intervention framework, teachers' professional judgments were equally predictive at determining risk level as the data provided by the screening tools in common use in

school districts to identify students who would benefit from a reading intervention. The study examined two research questions: (a) what is the relationship between teacher judgment of student reading risk levels and the screening tool risk levels, and (b) are there variations in the relationships related to student characteristics including identification as having limited English proficiency, receiving special education services, race/ethnicity, gender, or being economically disadvantaged?

The Researcher

The researcher works in a suburban school district in the Pacific Northwest with approximately 17,500 students in K-12th grades. She is the Assessment and Accountability Coordinator responsible for state testing, state and federal reporting, data analysis, and accountability measures. She has three years of experience in this role, has served on several state and regional committees related to assessment and accountability, and is currently Chair of a regional assessment consortium. Previously the researcher was Coordinator of Services for English Learners for the same district, principal of a Pre-K–8th grade Catholic School, and taught elementary English learners for a neighboring district. She is a licensed elementary classroom teacher with K-12 endorsements in English to Speakers of Other Languages and a Continuing Administrative License. The researcher has a B.A. in History from Oregon State University, M.A.T. in Elementary Education from the University of Portland, M.A.I.S. in Museum Studies, History & Anthropology from Oregon State University, and is a doctoral candidate at the University of Portland working towards an Ed.D. in Leadership & Learning.

Subjects and Setting

Teachers of 3rd and 4th grade students were chosen as subjects for this study because of the changes in reading instruction and behaviors that occur during these grade levels. Instruction begins to focus more on comprehension and vocabulary and less on skills such as phonics or phonemic awareness. By the end of 3rd grade most students will have developed the skills necessary for independent reading, and continued growth is related more to refining and utilizing comprehension strategies, interpreting what has been read, and gaining information from the text. In addition, in most states 3rd grade is the first year students begin taking summative standardized state tests.

The subjects of this study were 31 teachers from 16 elementary schools in a suburban school district in the Pacific Northwest. The analysis utilized test scores from their 3rd and 4th grade students. More of the participants were 3rd grade teachers (61%) than 4th grade (35%) and one teacher was a Special Education specialist who taught both 3rd and 4th graders. The majority of the teachers were White (87%) and Female (90%). The total number of years teaching ranged from two to 26 years (M = 14.39, SD = 6.12) and the number of years teaching the current grade ranged from one to 12 years (M = 5.45, SD = 3.80). Table 3.1 shows the breakdown of demographic information of all participants.

Table 3.1

Teacher Demographics

	n	%
Female	28	90
Male	3	10
Asian	1	3
Hispanic	1	3
Multi-Racial	2	7
White	27	87
Teaching 3rd Graders	19	61
Teaching 4th Graders	11	35
Both	1	3
Teaching 1-5 Years	2	7
Teaching 6-10 Years	9	29
Teaching >10 Years	20	65
Teaching Assigned Grade 1-3 Years	13	42
Teaching Assigned Grade 4-6 Years	7	23
Teaching Assigned Grade >6 Years	11	35

Note. One teacher taught Special Education and had both 3rd and 4th grade students.

The total enrollment of each school ranged from 228 to 559 students (M = 430.38, SD = 86.05). The student sample included 42% of the 3rd grade students and 24% of the 4th grade students in the district.

Table 3.2 below shows the numbers and percentages of students in each grade level in the study, in the district, and in the state. The district numbers and percentages were from October 1, 2015 (J. McGloghlon, personal communication, October 20, 2015). The state numbers for male/female and ethnicities were from the state 2014-15 Fall Membership Report (Oregon Department of Education, 2015). The state numbers for the percent of students considered Economically Disadvantaged and Students with Disabilities are taken from the 2015 Spring Membership Collection used for publishing the school, district, and state report cards (J. Wiens, personal communication, November 23, 2015). The percentage of students in the state who are English Learners is the unduplicated count of students, meaning each student was only

counted once even if he/she attended more than one school, reported for the 2014-15 school year in the LEP Collection (K. Miller, personal communication, November 23, 2015).

The number of students considered Economically Disadvantaged in the state is higher than in previous years due to the implementation of the Community Eligibility Provision, or CEP, program (J. Wiens, personal communication, November 23, 2015). In 2013-2014, in the state, 56% of students in 3rd grade and 55% in 4th grade were considered economically disadvantaged. The CEP program allows a district to count all students in a school or district as economically disadvantaged without requiring parents to complete applications for free/reduced price meals as long as the school has at least 40% of all students identified, and they provide free breakfast and lunch to all students (Dupuis & Hall, 2015). Only one school in the study district participates in this program. The percentages of students in each particular group in the sample closely match the percentages of students in 3rd and 4th grades in the district as a whole. Likewise, the district percentages, with the exception of Economically Disadvantaged, closely match the percentages of the state. The sample appears to be representative of students across the state.

Table 3.2

Student Demographics

	Sample						District		State	
	All		3rd Grade		4th Grade		3rd Gr.	4th Gr.	3rd Gr.	4th Gr.
	n	%	n	%	n	%	%	%	%	%
Female	411	52	261	51	150	54	52	52	49	49
Male	380	48	253	49	127	46	52	48	51	51
Economically Disadvantaged	394	50	269	52	125	45	47	46	60	60
English Learners	131	17	95	18	36	13	15	13	17	15
Students with Disabilities	122	15	77	15	45	16	15	16	15	16
American Indian/ Alaskan Native	7	1	6	1	1	<1	1	<1	1	1
Asian	45	6	27	5	18	6	6	7	4	4
Black/African American	11	1	10	2	1	<1	1	3	2	2
Hispanic	154	19	117	23	37	13	21	17	24	24
Native Hawaiian/ Pacific Islander	8	1	7	1	1	<1	1	1	1	1
Multi-Racial	52	7	28	5	24	9	6	7	6	6
White	514	65	319	62	195	70	65	66	62	62

Instrumentation

Reading risk was determined in two ways: the Reading Risk Evaluation Teacher Tracking Tool and STAR Reading.

Reading risk evaluation teacher tracking tool. Teachers completed a spreadsheet containing each of their student's names with their judgment of reading risk according to the Response to Intervention framework in operation in the district. Each student's current reading performance was rated as At/Above Benchmark, On Watch, Intervention, or Urgent Intervention based on individual teacher's classroom

observations, teacher administered and determined assessments, and general knowledge of the students' abilities. Teachers were also given an option to list or describe the information or assessments they used to generate their judgments of student risk and performance predictions. An example of this tool is provided in Appendix A.

This tool was piloted before administering it to the subject teachers. First, three expert teachers from outside the district reviewed the tool and provided feedback on readability and comprehension. Next, the tool was presented for review to the district's Director of Early Literacy and the Associate Director of Teaching and Learning to ensure the questions were appropriate for the teachers in the study and to receive input on its alignment with the vocabulary, timelines, and terminology utilized by the district.

STAR reading. According to the Technical Manual provided by the publisher, Renaissance Learning (2015), the STAR Reading assessment is a computer-adaptive, group-administered measure of reading comprehension. STAR Reading serves three purposes: providing a quick estimate of reading comprehension, assessing reading relative to national norms, and providing a means for consistently tracking reading growth. As discussed in Chapter 2, in 3rd grade and above, the assessment consists of 20 vocabulary-in-context questions and 5 authentic text passages followed by literal or inferential multiple-choice questions. (Renaissance Learning, 2015).

In their review of screening tools, the Center on Response to Intervention at American Institutes for Research (2010) found convincing evidence that the STAR

Reading assessment was a reliable and valid tool and that the classifications for predicting proficiency on state achievement tests were accurate. The evidence for reliability, validity, and classification was also convincing when disaggregated. The generalizability was determined to be broad, indicating studies were based on a large representative national sample with cross-validation (National Center on Response to Intervention, 2010b). Specific information about the reliability and validity of the assessment can be found in Chapter 2 of this paper.

Based on the results of the STAR Reading assessments and the norm-referenced percentiles adopted by the district, students were categorized as At/Above Benchmark, On Watch, Intervention, or Urgent Intervention. Students were considered At/Above Benchmark if their score placed them in at least the 40th percentile among other students at their grade-level, On Watch if they were between the 25th and 39th percentiles, Intervention if they were between the 10th and 24th percentiles and Urgent Intervention if they were below the 10th percentile.

Research Design and Procedures

Data for this study were collected during a four-week period between September 8 and October 2, 2015. The week the fall universal screening window opened, teachers were sent a link to a Google spreadsheet that included the names and district identification numbers of each of the students in their class along with a column to rate their judgment of each student's current reading risk. In addition, teachers were given an open field to optionally list or describe the information or assessments they used to generate their judgments of student risk. Teachers were

instructed to complete the spreadsheet before administering the STAR Reading assessment. Demographic information was also collected on each teacher including: race/ethnicity, number of years teaching, number of years teaching the current grade level, and number of years at the school.

Following district protocols, each student completed the STAR Reading assessment during the fall universal screening window (September 8-October 2). Students completed the assessment in computer labs or in their classrooms on mobile laptop labs. The total testing time was approximately 15 minutes for each class. Students who were absent or who for some other reason were not tested with their classes completed the assessment at a later date, within the testing window, following school procedures for completing make-up assessments. Among the students whose teachers participated in the study 11 (1.37%) did not complete the STAR Reading assessment due to excessive absences or other reasons. When there were multiple scores for the same student within the universal screening window only one was retained. If the assessments were completed within two days of each other it was assumed that there was an issue with the original administration and the last score was used. If there were more than two days between the assessments the original score was used. Thirty-two students (4.05%) had more than one STAR Reading result from the fall screening window.

The results from the STAR Reading assessment were exported from the system database and were added to a spreadsheet with the results of the teacher risk assessments. Teachers were assigned identification numbers that were then associated

with each of their students. At this point the names of the teachers and students were removed from the spreadsheet maintaining their confidentiality

Data Analysis

Descriptive statistics were run to determine the percentage of students at each level of risk using the two methods. For the purposes of analysis, students rated as At/Above Benchmark or On Watch were considered not at-risk and those rated as Intervention or Urgent Intervention were considered at-risk. This follows the district practice of intervention when students assess into at-risk categories. Students identified as On Watch during universal screening as part of the Response to Intervention framework do not participate in reading interventions unless additional information indicates a need for intervention. Thus, for all analyses, the risk assessments were dichotomized as either not at-risk or at-risk and comparisons were made between the dichotomized categorizations and the original grouping into four categories of risk. STAR Reading and teacher identification for risk were used to create four groups: 1) both identified, 2) STAR Reading only identified, 3) teacher only identified, or 4) neither identified. Descriptive statistics were run to determine the percentage of students in each category. Descriptive statistics were also run to determine whether or not there were differences related to student characteristics including identification as having limited English proficiency, receiving special education services, race/ethnicity, gender, or being economically disadvantaged

Percent exact agreement tests were used to determine the concurrent validity of the two measures. Suen (1988) described the need to clarify appropriate statistical

measures to use when establishing agreement, reliability, accuracy, and validity between measures and/or human raters in observations. He determined that percent agreement was an appropriate index to use with a criterion-referenced measure (Suen, 1988). In this study, percent exact agreement was calculated as the percent of students with a matching risk categorization from both STAR Reading and the teacher (i.e., both STAR Reading and the teacher rated the student as at-risk or not at-risk). Differences in the proportions of students identified based on participation in an English language services program, receiving special education services, race/ethnicity, gender, or being economically disadvantaged were also analyzed using the chi-square goodness-of-fit test to determine whether the two methods differed significantly from one another in the proportion of each group of students identified as at-risk.

For all analyses, students' race/ethnicity designations were combined into historically underserved race/ethnicity, which included American Indian/Alaskan Native, Black/African American, Hispanic, Native Hawaiian/Pacific Islander, and Multi-Racial and not historically underserved race/ethnicity, which included Asian and White. This was done to allow for the small numbers of students in some race/ethnicity categories and as an attempt to meaningfully determine if there were differences due to race/ethnicity.

On the Reading Risk Evaluation Teacher Tracking Tool, teachers were asked to describe what information they used when forming their risk determination. This was voluntary, it was not required that teachers complete this section in order to be

included in the study. This data was analyzed as a supplement to the quantitative analysis of the test scores. All but one teacher ($n = 30$) provided information but the level of detail varied. Some teachers listed information for individual students. Others described in general the types of information they considered to form their determinations for all students; still others provided a combination of a general description of the evidence considered with more specific pieces of information for certain individual students.

This information was taken from each teacher's individual spreadsheet and combined into a single spreadsheet containing the teacher identifier, student identifier (if applicable), and the teacher response. Using the constant comparative method of data analysis described by Merriam (2009), the responses were reviewed and codes were assigned to each response. In situations where a teacher included multiple responses for individual students, multiple codes may have been assigned (i.e., if the teacher noted consideration of test scores and a conference with the student, two different codes would have been assigned to the response). When the teacher provided a list of responses and indicated the types of information were used when making decisions for all students, multiple codes were assigned. Once the initial codes were assigned to all of the responses, the researcher grouped the codes into preliminary categories. These categories were combined and refined, ensuring that each response had a category and that the categories were mutually exclusive. In the end there were five categories established for the information provided by teachers.

Ethical Considerations

Institute Review Board (IRB) permission was granted on 8/19/2015 and district approval was granted by the superintendent before beginning the study. All IRB guidelines were followed. No consent forms granting permission for children to participate in the study were required due to the fact that no additional information was collected outside of the regular assessments given to every student in the district. The researcher had access to the demographic information and assessment results for required work activities. Consent was gained from teacher participants based on their voluntary completion of a data spreadsheet as described in the instructions (Appendix B) and the consent form that was completed by each participant.

Summary

The purpose of this research was to determine if, in the context of a Response to Intervention framework, teachers' professional judgments were equally predictive at determining risk level as the data provided by the screening tools in common use in school districts to identify students who would benefit from a reading intervention. The study examined two research questions: (a) what is the relationship between teacher judgment of student reading risk levels and the screening tool risk levels, and (b) are there variations in the relationships related to student characteristics including identification as having limited English proficiency, receiving special education services, race/ethnicity, gender, or being economically disadvantaged? Student reading risk was categorized as At/Above Benchmark, On Watch, Intervention, or Urgent Intervention using the norm-referenced percentile cut scores from STAR Reading

adopted by the district and teacher evaluation based on self-determined criteria. These two methods of risk determination were compared in order to answer the research questions and the results are found in Chapter 4 of this paper.

Chapter Four: Results

This chapter includes the results of the analysis of the data gathered in the study. First I will discuss the STAR Reading results including descriptive statistics on the scale scores and risk ratings for all students disaggregated by grade, gender, demographic category, and historically underserved race/ethnicity. Second, I will describe the results of the Reading Risk Evaluation Teacher Tracking Tool including the evaluations for all students, disaggregated by grade, gender, demographic category, and historically underserved race/ethnicity. Third, I will share the results of the analysis on the interaction between the two methods of risk determination including the overall risk match and the percent exact agreement risk determination for all students, disaggregated by grade, gender, demographic category, and historically underserved race/ethnicity. Fourth, I will describe the results of the chi-square test on the percent exact agreement risk determination of student groups with less than 80% agreement risk determination, and the percent exact agreement risk determination by teacher. Finally, I will discuss the survey results from teachers describing the different types of information individuals used to inform their decisions of their students' risk evaluations.

STAR Reading Results

All students in the study were assessed with STAR Reading during the district universal screening window (September 8 to October 2, 2015). Table 4.1 shows the total number of students along with the range, mean, and standard deviation of the

scale score for all students, disaggregated by gender, demographic category, and historically underserved race/ethnicity. The scores ranged from 61-1299, with a mean of 378 (*SD* = 205) for all students. The means for different groups of students ranged from 203 (*SD* = 149) for students with disabilities to 434 (*SD* = 213) for not economically disadvantaged.

Table 4.1

STAR Reading Scale Scores for All Students

	n	Min.	Max.	*M*	*SD*
All Students	791	61	1299	378	205
Female	411	61	1299	392	200
Male	380	66	1228	363	209
Economically Disadvantaged	394	61	1059	322	180
Not Economically Disadvantaged	397	68	1299	434	213
English Learners	131	61	472	206	107
Not English Learners	660	67	1299	412	203
Students with Disabilities	122	62	674	203	149
Students without Disabilities	669	61	1299	410	198
Historically Underserved Race/Ethnicity	232	61	105	302	181
Not Historically Underserved Race/Ethnicity	559	67	1299	410	206

Table 4.2 shows the total number of students along with the range, mean, and standard deviation of the STAR Reading scale score for 3rd grade students, disaggregated by gender, demographic category, and historically underserved race/ethnicity. For 3rd grade students the scores ranged from 61-968, with a mean of 328 (*SD* = 175) for all students. The means for different groups of 3rd grade students ranged from 156 (*SD* = 108) for students with disabilities to 374 (*SD* = 174) for not economically disadvantaged.

Table 4.2

STAR Reading Scale Scores for 3rd Grade Students

	n	Min.	Max.	Mean	SD
All 3rd Grade Students	514	61	968	328	175
Female	261	61	968	345	179
Male	253	66	935	310	171
Economically Disadvantaged	269	61	968	285	165
Not Economically Disadvantaged	245	68	935	374	174
English Learners	95	61	438	188	96
Not English Learners	419	67	968	359	174
Students with Disabilities	77	66	530	156	108
Students without Disabilities	437	61	968	358	167
Historically Underserved Race/Ethnicity	168	61	650	257	144
Not Historically Underserved Race/Ethnicity	346	67	968	361	179

Table 4.3 shows the total number of students along with the range, mean, and standard deviation of the STAR Reading scale score for 4th grade students, disaggregated by gender, demographic category, and historically underserved race/ethnicity. For 4th grade students the scores ranged from 62-1299, with a mean of 472 (SD = 223) for all students. The means for different groups of 4th grade students ranged from 255 (SD = 119) for English learners to 530 (SD = 234) for not economically disadvantaged.

Table 4.3

STAR Scale Scores for 4th Grade Students

	n	Min.	Max.	Mean	SD
All 4th Grade Students	277	62	1299	472	223
Female	150	62	1299	474	210
Male	127	72	1228	469	238
Economically Disadvantaged	125	62	1059	401	187
Not Economically Disadvantaged	152	70	1299	530	234
English Learners	36	62	472	255	119
Not English Learners	241	70	1299	504	217
Students with Disabilities	45	62	674	283	174
Students without Disabilities	232	78	1299	508	213
Historically Underserved Race/Ethnicity	64	62	1105	417	216
Not Historically Underserved Race/Ethnicity	213	70	1299	488	223

Based on the results of the STAR Reading assessments and the percentiles (Table 4.4) established by the district, students were categorized as At/Above Benchmark, On Watch, Intervention, or Urgent Intervention in relation to reading risk.

Table 4.4

STAR Reading Cut Scores

	Percentile	Risk Status	Scale Score
3rd Grade	≤ 9th Percentile	Urgent Intervention	< 176
	10th Percentile	Intervention	177-258
	25th Percentile	On Watch	259 - 318
	≥ 40th Percentile	At/Above Benchmark	≥ 319
4th Grade	≤ 9th Percentile	Urgent Intervention	< 264
	10th Percentile	Intervention	265 – 349
	25th Percentile	On Watch	350 – 414
	≥ 40th Percentile	At/Above Benchmark	≥ 415

Students were considered At/Above Benchmark if their score placed them in at least the 40th percentile among other students at their grade-level, On Watch if they were between the 25th and 39th percentiles, Intervention if they were between the 10th and 24th percentiles and Urgent Intervention if they were below the 10th percentile. For all students, a slightly higher percentage of female students, 54%, than male students, 50%, were considered At/Above Benchmark based on their STAR Reading results. This was also true for 3rd grade students, 53% of females and 46% of males), but in 4th grade both female and male students had the same percentage, 57%, considered At/Above Benchmark. For all students, 20% were determined to be in the highest risk category, Urgent Intervention. The group with the largest percentage of students considered Urgent Intervention was male 3rd grade students at 26%. Table 4.5 shows the number and percent of students in each risk category overall, by grade, and by gender.

Table 4.5

STAR Reading Risk Ratings

	At/Above Benchmark		On Watch		Intervention		Urgent Intervention	
	n	%	n	%	n	%	n	%
All Students	411	52	112	14	106	13	162	20
Female	223	54	61	15	54	13	73	18
Male	188	50	51	13	52	14	89	23
3rd Grade Students	254	49	72	14	72	14	116	23
Female	138	53	37	14	35	13	51	20
Male	116	46	35	14	37	15	65	26
4th Grade Students	157	57	40	14	34	12	46	17
Female	85	57	24	16	19	13	22	15
Male	72	57	16	13	15	12	24	19

Table 4.6 shows the number and percent of students in each risk category disaggregated by economically disadvantaged, English learner, students with disabilities, and historically underserved race/ethnicity. There were variations on the percentages of students in different risk categories when grouped by demographic category. Students who were economically disadvantaged were less likely to be considered At/Above Benchmark than students not economically disadvantaged, 40% to 64% respectively, and more likely to be considered Urgent Intervention at 28% versus 13% than students not economically disadvantaged. Some of the largest discrepancies were between English learners and non-English learners and students with disabilities compared to those without disabilities. Only 10% of all English learners were rated as At/Above Benchmark compared to 60% of all non-English learners. A similar pattern was found for Urgent Intervention with 52% of English learners and 14% of non-English learners at this highest risk level. For students with disabilities with 15% At/Above Benchmark and 60% in Urgent Intervention as compared to students without disabilities with 59% At/Above Benchmark and only

13% Urgent Intervention, the pattern repeated. There was less discrepancy between students considered historically underserved race/ethnicity or not, but a difference was revealed: 60% of students not considered historically underserved race/ethnicity were At/Above Benchmark and only 34% of historically underserved race/ethnicity. Similarly, in the Urgent Intervention category, 31% of students considered historically underserved race/ethnicity and 16% of students not were in this highest risk category.

Table 4.6

STAR Risk Ratings by Demographic Category

	At/Above Benchmark		On Watch		Intervention		Urgent Intervention	
	n	%	n	%	n	%	n	%
Economically Disadvantaged (ED)								
All ED Students	159	40	60	15	64	16	111	28
All not ED Students	252	64	52	13	42	11	51	13
3rd Grade ED Students	103	38	41	15	45	17	80	30
3rd Grade not ED Students	151	62	31	13	27	11	36	15
4th Grade ED Students	56	45	19	15	19	15	31	25
4th Grade not ED Students	101	66	21	14	15	10	15	10
English Learners (EL)								
All EL Students	13	10	16	12	34	26	68	52
All not EL Students	398	60	96	15	72	11	94	14
3rd Grade EL Students	9	10	12	13	28	30	46	48
3rd Grade not Students	245	59	60	14	44	11	70	17
4th Grade EL Students	4	11	4	11	6	17	22	61
4th Grade not EL Students	153	64	36	15	28	12	24	10
Students with Disabilities (w/Dis)								
All Students w/Dis	18	15	10	8	21	17	73	60
All Students w/o Dis	392	59	102	15	85	13	89	13
3rd Grade Students w/Dis	7	9	6	8	12	16	52	68
3rd Grade Students w/o Dis	247	57	66	15	60	14	64	15
4th Grade Students w/Dis	11	24	4	9	9	20	21	47
4th Grade Students w/o Dis	146	63	36	16	25	11	25	11
Historically Underserved Race/Ethnicity (URE)								
All URE Students	78	34	38	17	45	19	71	31
All not URE Students	333	60	74	13	61	11	91	16
3rd Grade URE Students	52	31	28	17	32	19	56	33
3rd Grade not URE Students	202	58	44	13	40	12	60	17
4th Grade URE Students	26	41	10	16	13	20	15	23
4th Grade not URE Students	131	62	30	14	21	10	31	15

Reading Risk Evaluation Teacher Tracking Tool Results

Teachers evaluated each student's current reading performance as At/Above Benchmark, On Watch, Intervention, or Urgent Intervention based on each individual teacher's classroom observations, teacher administered and determined assessments, and general knowledge of the students' abilities. Table 4.7 below shows the number and percent of students in each risk category overall, by grade, and by gender. Overall, 52% of students were At/Above Benchmark and 15% at Urgent Intervention based on teacher risk evaluation. For all students, a higher percentage of female students, 56%, than male students, 49%, were considered At/Above Benchmark. This was also true for 3rd grade students with 56% of females and 47% of males and in 4th grade, 53% of females and 51% of males, considered At/Above Benchmark. For all students, 15% were placed in the highest risk category, Urgent Intervention. The group with the largest percentage of students considered Urgent Intervention were male 3rd grade students, at 19%.

Table 4.7

Teacher Risk Ratings by Grade and Gender

	At/Above Benchmark		On Watch		Intervention		Urgent Intervention	
	n	%	n	%	n	%	n	%
All Students	410	52	133	17	130	16	118	15
Female	225	56	74	18	52	13	60	15
Male	185	49	59	16	78	21	58	15
3rd Grade Students	266	52	91	18	67	13	90	18
Female	146	56	49	19	23	9	43	17
Male	120	47	42	17	44	17	47	19
4th Grade Students	144	52	42	15	63	23	28	10
Female	79	53	25	17	29	19	17	11
Male	65	51	17	13	34	27	11	9

Table 4.8 shows the number and percent of students in each risk category disaggregated by economically disadvantaged, English learner, student with disabilities, and historically underserved race/ethnicity. Similar to the pattern seen with STAR Reading, there were variations between the percentages of students in different risk categories when grouped by demographics. Students who were economically disadvantaged were less likely to be considered At/Above Benchmark than students not economically disadvantaged, 41% to 63% respectively, and more likely considered Urgent Intervention at 20% versus 10% of students not economically disadvantaged. Some of the largest discrepancies were between English learners and non-English learners and students with disabilities compared to those without disabilities. Only 16% of all English learners were rated as At/Above Benchmark compared to 59% of all non-English learners. The same pattern was found for Urgent Intervention with 31% of English learners and 12% of non-English learners at this highest risk level, but these differences were less than those found with STAR Reading. A similar pattern was identified for students with disabilities with 12% At/Above Benchmark and 59% in Urgent Intervention as compared to students without disabilities with 59% At/Above Benchmark and only 7% Urgent Intervention. There was less discrepancy between students considered historically underserved race/ethnicity or not. 59% of students not considered historically underserved race/ethnicity were At/Above Benchmark and only 35% of historically underserved race/ethnicity. Similarly, in the Urgent Intervention category, 19% of students

considered historically underserved race/ethnicity and 13% of students not historically underserved race/ethnicity were in this highest risk category.

Table 4.8

Teacher Risk Ratings by Demographic Category

	At/Above Benchmark		On Watch		Intervention		Urgent Intervention	
	n	%	n	%	n	%	n	%
Economically Disadvantaged (ED)								
All ED Students	160	41	68	17	86	22	80	20
All not ED Students	250	63	65	16	44	11	38	10
3rd Grade ED Students	114	42	49	18	46	17	60	22
3rd Grade not ED Students	152	62	42	17	21	9	30	12
4th Grade ED Students	46	37	19	15	40	32	20	16
4th Grade not ED Students	98	65	23	15	23	15	8	5
English Learners (EL)								
All EL Students	21	16	20	15	49	37	41	31
All not EL Students	389	59	113	17	81	12	77	12
3rd Grade EL Students	21	22	15	16	31	33	28	30
3rd Grade not Students	245	59	76	18	36	9	62	15
4th Grade EL Students			5	14	18	50	13	36
4th Grade not EL Students	144	60	37	15	45	19	15	6
Students with Disabilities (w/Dis)								
All Students w/Dis	15	12	14	12	21	17	72	59
All Students w/o Dis	395	59	119	18	109	16	46	7
3rd Grade Students w/Dis	8	10	6	8	6	8	57	74
3rd Grade Students w/o Dis	258	59	85	20	61	14	33	8
4th Grade Students w/Dis	7	16	8	18	15	33	15	33
4th Grade Students w/o Dis	137	59	34	15	48	21	13	6
Historically Underserved Race/Ethnicity (URE)								
All URE Students	81	35	43	19	65	28	43	19
All not URE Students	329	59	90	16	65	12	75	13
3rd Grade URE Students	59	35	33	20	43	26	33	20
3rd Grade not URE Students	207	60	58	17	24	7	57	17
4th Grade URE Students	22	34	10	16	22	34	10	16
4th Grade not URE Students	122	57	32	15	41	19	18	9

Match Between STAR Reading and Teacher Risk Evaluations

In the universal screening process in place in the district, students who fall in the category of At/Above Benchmark and On Watch, are not considered at risk while those in the Intervention or Urgent Intervention category are considered at risk for not

meeting grade level benchmarks in reading. These categories were therefore collapsed into At Risk (i.e., Intervention and Urgent Intervention categories) and Not at Risk (i.e., At/Above Benchmark and On Watch). The STAR Reading and teacher identified risk categories were analyzed through a 2x2 table with four groups: 1) both STAR and teachers identified as at risk, 2) STAR Reading only identified as at risk, 3) teacher only identified as at risk, or 4) neither identified as at risk. Table 4.9 shows the number and percent of students in each group for the students overall and in 3rd and 4th grade. When these results are combined, it appears that 18% (89) of 3rd grade students, 18% (49) of 4th grade students, and 17% (138) of all students have some sort of mismatch between the risk determination provided by the teacher and that established by STAR Reading.

Table 4.9

Overall Risk Match

	STAR – Not Risk		STAR – Risk	
	n	%	n	%
Teacher – Not Risk				
All Students	464	59	79	10
3rd Grade Students	297	58	60	12
4th Grade Students	167	60	19	7
Teacher – Risk				
All Students	59	8	189	24
3rd Grade Students	29	6	128	25
4th Grade Students	30	11	61	22

Table 4.10 includes the number and percent of students in each group disaggregated by grade and gender. For all students, 20% ($n = 75$) of male students and 16% ($n = 63$) of female students have a mismatch between the risk determination provided by the teacher and that established by STAR Reading. For 3rd grade students,

19% (n = 47) of male students and 16% (n = 42) of female students have a mismatch between the risk determination provided by the teacher and that established by STAR Reading. For 4th grade students, 22% (n = 28) of male students and 14% (n = 21) of female students have a mismatch between the risk determination provided by the teacher and that established by STAR Reading.

Table 4.10

Overall Risk Match by Grade and Gender

	STAR – Not Risk		STAR – Risk	
	n	%	n	%
Teacher – Not Risk				
All Male	204	54	40	11
All Female	260	63	39	10
3rd Grade Male	133	53	29	12
3rd Grade Female	164	63	31	12
4th Grade Male	71	56	11	9
4th Grade Female	96	64	8	5
Teacher – Risk				
All Male	35	9	101	27
All Female	24	6	88	21
3rd Grade Male	18	7	73	29
3rd Grade Female	11	4	55	21
4th Grade Male	17	13	28	22
4th Grade Female	13	9	33	22

For students overall, the largest differences between the risk determination given by the teacher and that established by STAR Reading were found for English learners at 27% (n = 36), historically underserved race/ethnicity at 23% (n = 54), and students who are economically disadvantaged with 20% (n = 79) risk determination mismatched. Table 4.11 includes the number and percent of all students in each group disaggregated by economically disadvantaged, English learners, students with disabilities, and historically underserved race/ethnicity.

Table 4.11

Overall Risk Match by Demographic Category

	STAR – Not Risk		STAR – Risk	
	n	%	n	%
Teacher – Not Risk				
Economically Disadvantaged	184	47	44	11
Not Economically Disadvantaged	280	71	35	9
English Learners	17	13	24	18
Not English Learners	447	68	55	8
Students with Disabilities	22	18	7	6
Students without Disabilities	442	66	72	11
Historically Underserved Race/Ethnicity	93	40	31	13
Not Historically Underserved Race/Ethnicity	371	66	48	9
Teacher – Risk				
Economically Disadvantaged	35	9	131	33
Not Economically Disadvantaged	24	6	58	15
English Learners	12	9	78	60
Not English Learners	47	7	111	17
Students with Disabilities	6	5	87	71
Students without Disabilities	53	8	102	15
Historically Underserved Race/ethnicity	23	10	85	37
Not Historically Underserved Race/Ethnicity	36	6	104	19

For 3rd grade students, the largest differences between the risk determination given by the teacher and that established by STAR Reading were found for English learners at 30% ($n = 32$), historically underserved race/ethnicity at 24% ($n = 40$), and students who are economically disadvantaged with 21% ($n = 57$) risk determination mismatched. Table 4.12 includes the number and percent of 3rd grade students in each group disaggregated by economically disadvantaged, English learners, students with disabilities, and historically underserved race/ethnicity.

Table 4.12

3rd Grade Risk Match by Demographic Category

	STAR – Not Risk		STAR – Risk	
	n	%	n	%
Teacher – Not Risk				
Economically Disadvantaged	125	47	38	14
Not Economically Disadvantaged	172	70	22	9
English Learners	14	15	22	23
Not English Learners	283	68	38	9
Students with Disabilities	10	13	4	5
Students without Disabilities	287	66	56	13
Historically Underserved Race/Ethnicity	66	39	26	16
Not Historically Underserved Race/Ethnicity	231	67	34	10
Teacher – Risk				
Economically Disadvantaged	19	7	87	32
Not Economically Disadvantaged	10	4	41	17
English Learners	7	7	52	55
Not English Learners	22	5	76	18
Students with Disabilities	3	4	60	78
Students without Disabilities	26	6	68	16
Historically Underserved Race/Ethnicity	14	8	62	37
Not Historically Underserved Race/Ethnicity	15	4	66	19

For 4th grade students, almost all groups had above 80% match. The only group with a risk determination mismatch percentage higher than 20% was historically underserved race/ethnicity at 22% ($n = 14$) mismatch between the risk determination given by the teacher and that established by STAR Reading. Table 4.13 includes the number and percent of 4th grade students in each group disaggregated by economically disadvantaged, English learners, students with disabilities, and historically underserved race/ethnicity.

Table 4.13

4th Grade Risk Match by Demographic Category

	STAR – Not Risk		STAR – Risk	
	n	%	n	%
Teacher – Not Risk				
Economically Disadvantaged	59	47	6	5
Not Economically Disadvantaged	108	71	13	9
English Learners	3	8	2	6
Not English Learners	164	68	17	7
Students with Disabilities	12	27	3	7
Students without Disabilities	155	67	16	7
Historically Underserved Race/Ethnicity	27	42	5	8
Not Historically Underserved Race/Ethnicity	140	67	14	7
Teacher – Risk				
Economically Disadvantaged	16	13	44	35
Not Economically Disadvantaged	14	9	17	11
English Learners	5	14	26	72
Not English Learners	25	10	35	15
Students with Disabilities	3	7	27	60
Students without Disabilities	27	12	34	15
Historically Underserved Race/Ethnicity	9	14	23	36
Not Historically Underserved Race/Ethnicity	21	10	38	18

The percent exact agreement between STAR Reading and teacher evaluation shows some variation based on groups of students. Overall, for all students, there was a match between STAR Reading and the teacher evaluation 83% of the time. Of the 33 different groupings of students, all but seven showed over 80% agreement on risk/not risk between STAR Reading and teacher evaluation. The only group in which all grades had less than 80% match were historically underserved races/ethnicities, which had a 77% match for all grades, 76% for 3rd grade students, and 78% for 4th grade students. Table 4.14 below shows the percent exact agreement for risk/not risk between STAR Reading and teacher evaluation, organized from smallest to largest, and includes the number and percent of students in each group with matching risk

evaluations as well as the percent of each group determined to be at risk only by STAR Reading, or teacher evaluation.

Table 4.14

Percent Exact Agreement Risk Determination

	Percent Exact Agreement Risk/No Risk		STAR Only Risk	Teacher Only Risk
	n	%	%	%
3rd Grade English Learners	66	70	23	7
All English Learners	95	73	18	9
3rd Grade Historically Underserved Race/Ethnicity	128	76	16	8
All Historically Underserved Race/Ethnicity	178	77	13	10
4th Grade Male	99	78	9	13
4th Grade Historically Underserved Race/Ethnicity	50	78	8	14
3rd Grade Economically Disadvantaged	212	79	14	7
All Economically Disadvantaged	315	80	11	9
4th Grade English Learners	29	80	6	14
All Male	305	81	11	9
All Students without Disabilities	544	81	11	8
3rd Grade Male	206	82	7	12
All 4th Grade	228	82	7	11
4th Grade Economically Disadvantaged	103	82	5	13
4th Grade Not Economically Disadvantaged	125	82	9	9
3rd Grade Students without Disabilities	355	82	13	6
All Students	653	83	10	7
All 3rd Grade	425	83	12	6
4th Grade Not English Learners	199	83	7	10
4th Grade Students without Disabilities	189	83	7	12
All Female	348	84	10	6
3rd Grade Female	219	84	4	12
4th Grade Not Historically Underserved Race/Ethnicity	178	84	7	10
All Not English Learners	558	85	8	7
All Not Historically Underserved	475	85	9	6
4th Grade Female	129	86	5	9
All Not Economically Disadvantaged	338	86	9	6
3rd Grade Not English Learners	359	86	9	5
3rd Grade Not Historically Underserved Race/Ethnicity	297	86	10	4
3rd Grade Not Economically Disadvantaged	213	87	9	4
4th Grade Students with Disabilities	39	87	7	7
All Students with Disabilities	109	89	6	5
3rd Grade Students with Disabilities	70	91	5	4

A chi-square goodness-of-fit analysis of the seven student groups with less than 80% STAR Reading and teacher evaluation risk/not risk match revealed that most

groups showed significant differences between the observed versus expected frequencies of match between STAR Reading and teacher evaluation. The groups with significant differences included all English learners ($p = .001$), 3rd grade English learners ($p < .001$), all historically underserved race/ethnicity ($p = .005$), 3rd grade historically underserved race/ethnicity ($p = .007$), and 3rd grade economically disadvantaged ($p = .015$). Table 4.15 below includes the sample sizes, observed and expected frequencies of the matched students, and percent exact agreement for each group, along with the chi-square results.

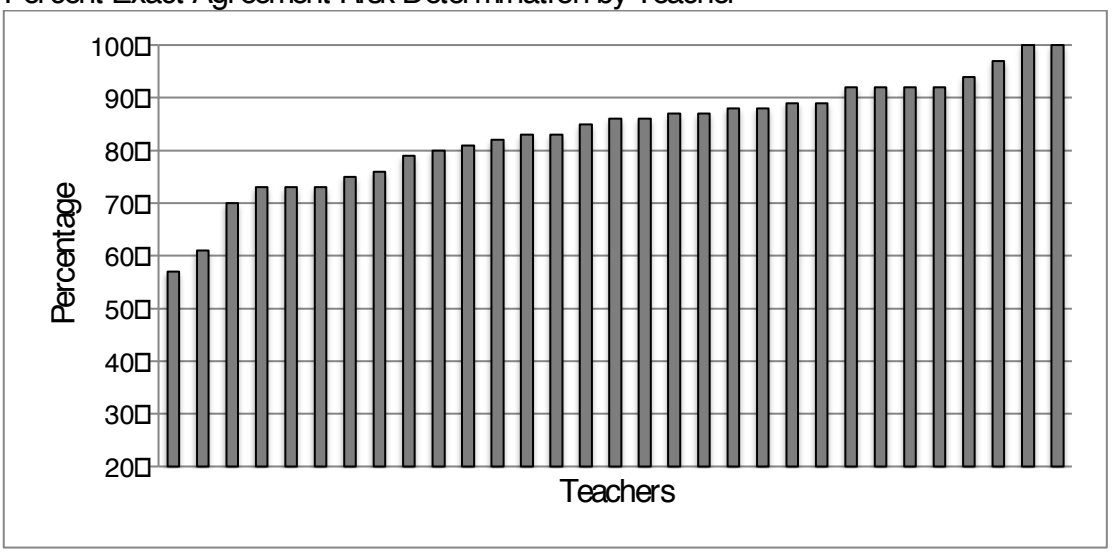
Table 4.15

Chi-Square of Categories with Less than 80% Percent Exact Agreement Risk Determination

	Sample Size	Observed Frequencies	Expected Frequencies	Percent Exact Agreement	Chi	p
4th Grade Gender					3.06	.080
Male	127	99	105	78%		
Female	150	129	124	86%		
All English Learners (EL)					10.98	.001
All EL	131	95	108	73%		
All Not EL	660	558	545	85%		
3rd Grade English Learners (EL)					14.21	< .001
3rd Grade EL	95	66	79	70%		
3rd Grade Not EL	419	359	346	86%		
4th Grade English Learners (EL)					.088	.767
4th Grade EL	36	29	30	80%		
4th Grade Not EL	241	199	198	83%		
All Historically Underserved Race/Ethnicity (URE)					7.75	.005
All URE	232	178	192	77%		
All Not URE	559	475	462	85%		
3rd Grade Historically Underserved Race/Ethnicity (URE)					7.35	.007
3rd Grade URE	168	128	139	76%		
3rd Grade Not URE	346	297	286	86%		
4th Grade Historically Underserved Race/Ethnicity (URE)					1.00	.317
4th Grade URE	64	50	53	78%		
4th Grade Not URE	213	178	175	84%		
3rd Grade Economically Disadvantaged (ED)					5.92	.015
3rd Grade ED	269	212	222	79%		
3rd Grade Not ED	245	213	203	87%		

When examined by teacher, the percent exact agreement ranged from 57% to 100% (Figure 4.1). Most, 71% ($n = 22$) had above 80% percent exact agreement with STAR Reading. While there did not appear to be patterns based on teacher demographics, two teachers did stand out as having lower rates of agreement, 57% and 61% respectively, with STAR Reading. There did not appear to be a correlation ($r = -.07$, $p > .05$) between the number of years of teaching and the percent exact agreement nor a correlation ($r = -.10$, $p > .05$) between the number of years of teaching the current grade and the percent exact agreement.

Figure 4.1



Information Used by Teachers When Making Decisions

On the Reading Risk Evaluation Teacher Tracking Tool, teachers were asked to describe what information they used when forming their risk determination. This was voluntary, not required to be included in the study, as explained by the researcher during recruitment meetings and again on the consent form and the tool itself. This

data source is supplementary to the quantitative analysis of the test scores. All but one teacher ($n = 30$) provided information but the level of detail varied. Some teachers listed information for individual students. Others described in general the types of information they considered to form their determinations for all students; still others provided a combination of a general description of the evidence considered with more specific pieces of information for certain individual students. When the information was summarized, five categories of information teachers used emerged: information from previous teachers, classroom observation or class work, conferencing, formal assessments, and personal knowledge.

Information from previous teachers. When making their evaluations, 17 (57%) teachers reported utilizing information from previous teachers to make their decisions, such as a review of cumulative folders, discussions with the previous classroom teacher or specialists who had worked with the student, notes from team meetings from the previous year, or knowledge of whether or not the student participated in a reading intervention the previous year. One teacher described utilizing "teacher feedback from last year," when making decisions about a student and another mentioned that a particular student was "at/above last year" but that he/she had observed the student struggling during the first weeks of the current year and rated the student as On Watch. Two teachers included formalized processes of reviewing "placement cards" or other methods of conveying information about students from one year to the next between teachers. Another met with a student's

previous case manager who "had no concerns with reading" and so rated this student as At/Above Benchmark.

Classroom observations and class work. Most teachers, 21 (70%), described that they used information gained through classroom observations or reviewing class work. The most common were listening to students read (including in math or other subject areas), observing independent reading, or attention to the books students chose for independent reading. One teacher described noticing that a student did not appear to enjoy reading and rated the student's risk as Intervention based on the fact that "he picks easy books and has been refusing to read." This same teacher described another student as At/Above Benchmark partially based on the observation that "he reads every spare minute." Other observations teachers reported making were attention to reading informal passages, reviewing classroom writing, behavior in class, general interactions with the student, and perception of reading fluency or comprehension. One teacher even went so far as to respond, "I did not use ANY prior assessments to gauge their levels, I only used in-class work for the past two weeks to assess where they most likely are for 4th grade work" (emphasis in original).

Conferencing. Another common response was conferencing. Eleven (37%) teachers included meeting with students one on one to discuss reading preferences, their summer reading, the way they chose books, or general academic issues. One teacher described the conference process, "I met with students and had them share their book bag choices. They put them as too easy, too hard, or just right. Then I had them read about one page of each book to predict their possible reading ability."

Teachers valued the information gained through conferencing with students. One described learning about a new student through a conference, “I get the feeling there are some focus issues which may affect her reading.”

Formal assessments. A total of 20 (67%) teachers reported drawing on formal assessments, both of reading and other academic content or non-academic topics, to inform their decisions about students. Some teachers used a student’s English language proficiency level or designation as an English learner in the decision-making process. Most mentioned knowing whether or not a student had an identified disability, and one teacher noted a Talented and Gifted designation. Common reading assessments included Informal Reading Inventory (IRI), STAR Reading or STAR Early Literacy results from the previous spring, Dynamic Indicators of Basic Early Literacy Skills (DIBELS) oral reading fluency (ORF) or phonics screenings, the Read Naturally placement test, sight word assessments, placement tests from the reading adoption, or running records. One teacher in a two-way bilingual classroom utilized the Spanish reading screener, Indicadores Dinámicos del Éxito en la Lectura (IDEL), for risk determinations. Some teachers expressed surprise at the results of formal reading assessments, others that the results validated their beliefs. At times the results of different tools gave conflicting information. One teacher mentioned a student rated as Urgent Intervention in the previous year-end STAR Reading assessment whose current DIBELS-ORF and running record indicated the student was at grade level. Ultimately the teacher rated this student as On Watch, in part based on “home life issues” in combination with the formal assessment results.

Personal knowledge. The last category of information utilized by teachers was their knowledge about students. Eleven (37%) teachers mentioned their perception of a student's personality using phrases such as, "doesn't seem confident," "learns quickly," "reminds me of my son," or "enthusiastic" as considerations when rating students. A student's general academic ability, reading habits (including summer reading), or work habits such as "ability to complete tasks without assistance," or "ability to explain ideas to peers" were often considered for individual student risk determinations. Family life was also sometimes used. Student mobility, knowledge of a student's family through siblings or meeting with parents, and knowledge of personal traumas all impacted teachers' ratings. One teacher rated a student as Intervention despite assessment results indicating the student was At/Above Benchmark because the student had, "no confidence and a traumatic event that affects every moment of his life." Most teachers considered their general perceptions of student, "For all students I used a gut feeling based on the first few days of school, along with having them read passages and do a quick comprehension check."

Table 4.16 includes each teacher's percent exact agreement between STAR Reading and the Reading Risk Evaluation Teacher Tracking Tool along with which categories of information each teacher self-reported using to form their decisions when completing the tracking tool.

Table 4.16

Percent Exact Agreement and Categories of Information by Teacher

Teacher	% Exact Agreement Risk Evaluation	Information from Previous Teachers	Classroom Observations and Class Work	Conferencing	Formal Assessments	Personal Knowledge
107	57				X	
117	61		X			
105	70		X			
112	73	X	X		X	X
124	73	X	X			
130	73	X		X	X	X
108	75					
114	76		X		X	
127	79	X	X		X	X
119	80	X			X	X
125	81	X	X			
121	82		X		X	X
103	83				X	X
120	85				X	
109	86	X	X	X	X	X
110	86	X	X	X	X	
123	86	X	X		X	
101	87	X	X	X		X
113	87		X	X		
126	88	X	X	X	X	X
129	88		X			
100	89	X	X	X	X	
118	89			X		
102	92		X		X	
106	92	X	X	X		
115	92	X	X		X	
128	92	X	X	X	X	
111	94				X	
104	97	X			X	X
116	100			X		
122	100	X	X		X	X

Summary

This study investigated if, in the context of a Response to Intervention framework, teachers' professional judgments were equally predictive at determining risk level as the data provided by the screening tools in common use in school districts to identify students who would benefit from a reading intervention and found that

there was 83% agreement between the STAR Reading and teacher evaluation of each student's reading risk. This study also found there was little variation in the percent agreement based on student characteristics. An exception to this pattern was for historically underserved races/ethnicities, which showed consistently lower percent agreement (77% overall, 76% 3rd grade, and 78% 4th grade). A chi-square goodness-of-fit analysis of the 7 groups with less than 80% STAR Reading and teacher evaluation match between risk/no risk revealed that five showed differences between the observed versus expected frequencies of match between STAR Reading and teacher evaluation. Analysis of the responses from teachers identified five categories of information used when making decisions about students: information from previous teachers, classroom observations, conferences, formal assessments, and knowledge about students. These important take-ways will be discussed further in Chapter 5.

Chapter Five: Discussion and Conclusions

This chapter begins with a summary of the study followed by a discussion of the findings. Next is a description of the limitations of this study. The chapter ends with implications for practice and suggestions for future research in this area.

Summary of Study

This study was conducted in response to a need to examine the efficacy of teacher judgment about students' reading risk. In the national context of an increasing emphasis on accountability, the use of data, and standardized testing, teachers' judgments are not always valued. Instead, systems have been designed that encourage relying on the results of screening tools or other standardized tests, sometimes without referencing teachers' professional judgment.

A review of the literature revealed a gap in the research related to teachers' abilities to identify individual student risk in reading. Many districts across the United States utilize a Response to Intervention (RTI) framework of instruction including universal screening for individual student risk. In fact, 17 states require utilization of RTI data in some form for identification of students with Specific Learning Disability for special education services, six states require districts to submit an RTI plan as part of the special education process, four states have established timelines for adoption of RTI, and an additional 12 states provide information within state regulations for districts that choose to utilize RTI (Hauerwas, Brown, & Scott, 2013). Universal screening is becoming more common and districts are investing time, effort, and

funding in purchasing tools and establishing screening protocols. And yet, there does not seem to be research on whether or not teachers are able to identify their students' risk without the assessment tools or processes.

The purpose of this research was to determine if, in the context of a Response to Intervention framework, teachers' professional judgments were equally predictive at determining risk level as the data provided by the screening tools in common use in school districts to identify students who would benefit from a reading intervention. The study examined two research questions: (a) what is the relationship between teacher judgment of student reading risk levels and the screening tool risk levels, and (b) are there variations in the relationships related to student characteristics including identification as having limited English proficiency, receiving special education services, race/ethnicity, gender, or being economically disadvantaged?

The subjects of this study were 31 teachers from 16 elementary schools in a suburban district in the Pacific Northwest. The analysis utilized test scores from their 3rd and 4th grade students. The demographics of the school district and sample appear representative of students across the state. Reading risk was determined in two ways, a Reading Risk Evaluation Teacher Tracking Tool and STAR Reading. Using the Reading Risk Evaluation Teacher Tracking Tool, teachers rated each student's current reading performance as At/Above Benchmark, On Watch, Intervention, or Urgent Intervention based on individual teacher's classroom observations, teacher administered and determined assessments, and general knowledge of the students' abilities. Teachers were also given an option to list or describe the information they

used to generate their evaluations of student risk and performance predictions. Following district protocols, each student completed the STAR Reading assessment during the fall universal screening window (September 8-October 2, 2015). Based on the results of the STAR Reading assessments and the norm-referenced percentiles established by the district, students were categorized as At/Above Benchmark, On Watch, Intervention, or Urgent Intervention. Students were considered At/Above Benchmark if their score placed them in at least the 40th percentile among other students at their grade-level, On Watch if they were between the 25th and 39th percentiles, Intervention if they were between the 10th and 24th percentiles, and Urgent Intervention if they were below the 10th percentile.

Discussion of Findings

The first research question explored whether or not there was a relationship between teacher judgment of student reading risk levels and the STAR Reading risk levels. The results of each assessment were dichotomized into two categories: at-risk, including those rated as Intervention or Urgent Intervention, or not at risk, those rated as At/Above Benchmark or On Watch. Analysis of the results found 83% agreement between the STAR Reading and teacher evaluation of each student's reading risk. There was little variation based on gender, grade, or other student characteristics. For most students, the reading risk determined by teachers in the first few weeks of school matched that determined by a commercially published assessment. Students were more likely to be found at-risk by STAR Reading and not at-risk by their teachers, 10%, than the opposite, 7%. Possibly teachers were applying different knowledge they

had about their students, knowledge which made them believe students were at more or less risk than their test scores might indicate. It is important to note that teachers responded within the first few weeks of school and thus had little information to use when making their determinations about student risk so it is possible they were mistaken in their risk determinations based on limited information about their students. There were also teachers who reviewed the STAR Reading results from the previous year which would have influenced their determinations and may have contributed to the high percent exact agreement.

The second research question asked if there were variations in the relationships related to student characteristics including identification as having limited English proficiency, receiving special education services, race/ethnicity, gender, or being economically disadvantaged. When disaggregated by various student groupings, the percent exact agreement was above 80% for all but 7 groups. For students overall, the largest percentages of disagreement between the risk determination given by the teacher and that established by STAR Reading were found for English learners at 27%, historically underserved race/ethnicity at 23%, and students who are economically disadvantaged with 20%. For 3rd grade students, the largest percentages of disagreement between the risk determination given by the teacher and that established by STAR Reading were found for English learners at 30%, historically underserved race/ethnicity at 24%, and students who were economically disadvantaged with 21%. For 4th grade students, the only group with a risk determination disagreement percentage higher than 20% was historically underserved race/ethnicity at 22%

disagreement between the risk determination given by the teacher and that established by STAR Reading. The higher percentage disagreement could be due to the fact that English learners are more likely to be considered historically underserved race/ethnicity. Another consideration is that the teacher with the lowest percent exact agreement taught in a dual language Spanish-English program. Roughly half of this teacher's students were Hispanic English learners. This teacher rated the students based on their Spanish literacy while STAR Reading rated their risk in English reading. This could account for this teacher's low percent exact agreement and may have contributed to the lower percent exact agreement for English learners and historically underserved race/ethnicity, particularly at the 3rd grade. Another possibility is that STAR Reading is a more accurate identifier for these groups of students than teachers that might be affected by biases.

In its review of various assessments, the National Center on Response to Intervention (2010b) reported differences for the predictive validity for STAR Reading for Hispanic (.55-.74) versus White (.69-.74) students. The lower predictive validity reported by the NCRTI for Hispanic students matches what was found for historically underserved races/ethnicities in this study. There were fewer matches between the teacher evaluations of risk and STAR Reading for students considered part of a historically underserved race/ethnicity, which could mean teachers are more accurate predictors, but maybe not. If the risk determinations are compared to the year-end summative assessment results, it might be possible to determine if STAR Reading or teachers were more accurate at identifying student risk early in the school

year. Studies on STAR Reading show predictive validity as low as .55 for Hispanic students, a population that is growing in size and percentage of total, which is concerning. If teachers are better able to identify which students require intervention their judgments should be considered as part of the risk determination process.

Based on this study, there appears to be a relationship between a teacher's judgment of reading risk and the risk determined by STAR Reading. In the context of a Response to Intervention framework, the purpose of screening tools such as STAR Reading are to identify students that might be at risk for not meeting grade-level benchmarks at year end. Teachers are then able to provide an intervention that will allow students to make faster progress and catch up. Research on Response to Intervention determined that using universal screening tools allowed schools to identify which students would benefit from interventions, increasing the number of students who were at benchmark at the end of the year (e.g., Hosp, Hosp, & Cole, 2011; Hughes and Dexter, 2011) This study suggests that a teacher's professional judgment might be used as a screening tool, eliminating the need to purchase and maintain a commercially published assessment for the purposes of universal screening. A school or district could establish a protocol, utilizing the Reading Risk Evaluation Teacher Tracking Tool, or another similar document, of systematically collecting and considering a teacher's evaluation of student risk, which could then be used to determine which students would benefit from an intervention. The information that teachers access when making their decisions is broader than that available from STAR

Reading (or other universal screeners) and could potentially allow for a more accurate risk determination.

When districts implement a Response to Intervention framework, there is the potential for students to be incorrectly identified as at-risk or not at-risk based on universal screening. Reliance on a single assessment to determine risk makes it more likely students who need an intervention would not get one, and resources (educator time, intervention materials, etc.) would be mistakenly used for students who do not need the additional assistance. It is important to have a better understanding of which students are incorrectly identified as at-risk or not at-risk by STAR Reading or by their teachers in order to more accurately intervene with students who actually need the additional support. Another issue is the range of agreement amongst teachers. Although most had above 80% match with STAR, there were a few with much lower percentages. This study did not examine the consistency of correct risk determinations of teachers over time, which would help determine if the lower rates were aberrations or if some teachers are better at identifying risk than others.

When teachers make decisions about their students they draw on a wide range of knowledge about the students, triangulating the empirical knowledge from the test score with other types of information they have about the students' academic strengths and weaknesses as well as personal characteristics. In this study, the responses provided by teachers indicated they utilized information from previous teachers, classroom observations, conferences with students, and/or formal assessment information (both of reading as well as other areas such as language proficiency) when

rating the risk their students had in the area of reading. It should be noted that teachers were told responding to this portion of the Reading Risk Evaluation Teacher Tracking Tool was optional and there were no specific directions on how to respond. Each teacher determined how and what to report.

A review of the literature revealed four categories of ways of knowing: personal, the knowledge based on the authentic relationship between teacher and learner; aesthetic, the art of knowing by doing; ethical/emancipatory, the moral and contextual knowledge teachers and students have in their roles, environment, and history; and empirical, the things that can be seen, heard, or touched (Gurm, 2013). These categories make distinctions about the different types of knowledge, but the study revealed that the ways teachers utilize this information when making decisions about students is anything but distinct. Rather, the various pieces of information are interwoven, as in a tapestry, to form the teacher's evaluation of each student's risk. Teachers use their empirical (test scores) knowledge in combination with personal and ethical/emancipatory, all in the context of the aesthetic knowledge gained from their everyday interactions with students to arrive at a judgment about reading risk. The accuracy of a teacher's evaluation of his/her students risk appear to be stronger (as measured by agreement with the STAR Reading risk determination) when accessing a range of ways of knowing rather than one or two categories.

The teachers with the three lowest percent exact agreement with STAR Reading reported using limited forms of information about their students. The teacher with the lowest agreement, 57%, reported using only the results of a reading

assessment in order to rate the students. This particular teacher teaches in a dual language immersion program and the literacy instruction is in Spanish while STAR Reading assesses English reading. The teacher responded based on risk in Spanish reading. This could account for the lower percent exact agreement, but it is worth noting that the teacher reported making the evaluations based solely on the reading assessment, the empirical way of knowing. The two teachers with the next lowest percent exact agreement, 61% and 70%, reported that they did not use any formal assessment data. Rather, they relied on a "gut feeling" (ethical/emancipatory) or "only class work" (aesthetic) when making their risk evaluations. Teachers with higher percent exact agreement ratings, with a few exceptions, reported a wider range of types of knowledge that were used for making their evaluations.

Two teachers had 100% agreement with the risk determinations from STAR Reading. One teacher reported one-on-one conferences with students as the only method for determining student risk. It is hard to know specifically what was discussed in the conferences, but it is reasonable to assume there was a range of information from academic history to test scores to personal information about each student's family life. The second teacher with 100% exact agreement with STAR Reading listed a wide range of information from formal assessment results to personal characteristics, information from the previous year's teacher to knowledge of performance in the classroom. The variety of ways of knowledge reported by this teacher could account for the high percent exact agreement with the risk determinations from STAR Reading.

Limitations

This study has several major limitations that must be discussed. First, the results were limited by the fact that the sample was drawn from a single school district in a single area of the country. This sample was one of convenience as the researcher works for the school district and has convenient access to the student assessment results. This, combined with the inclusion of only 3rd and 4th grade students, makes it uncertain if the results are generalizable. Second, teacher participation was optional (there was a 33% response rate), and it is possible this influenced the results. Teachers who were less confident in their ability to identify student risk may have chosen not to participate and the sample size of only 31 teachers also makes it difficult to generalize. Third, this study was limited to an initial investigation into the abilities of teachers to identify student risk. While a longitudinal study of teachers' abilities to predict student risk, particularly in comparison with universal screening tools, would be useful, the limitations for the project did not allow the necessary time. The study was based on the assumption that STAR Reading correctly identifies which students are at-risk as a way to determine whether or not teachers could also correctly identify student reading risk. The researcher chose to focus on a comparison of teachers' perceptions at the beginning of the school year with assessment results from the same time period in an attempt to capture the tacit knowledge teachers have about their students' abilities before they are influenced by reviewing the universal screening results. Fourth, it is not possible to ensure teachers completely disregarded STAR Reading results. Teachers self-reported that they completed the Reading Risk Teacher Tracking Tool

before reviewing their students' STAR Reading results, but it is not possible to verify this. Also, a number of teachers reported that they reviewed the STAR Reading results from the previous spring, which may have influenced the percent of agreement with reading risk. Fifth, future research should utilize more complicated methods, such as multi-level models, that take into consideration the nestedness of the data.

Implications for Practice

School districts have invested vast amounts of time and money into developing Response to Intervention frameworks including universal screening of all students for academic risk. The district subject to this study spends approximately $100,000 annually ($10/student) on the STAR assessments in addition to the first-year costs, which included approximately $10,000 per school for set-up (K. Rush, personal communication, February 8, 2016). In addition to universal screening for all students in reading and math in K-8th grades, STAR Reading is widely used (and STAR Math to a lesser extent) for monitoring the progress of students who are receiving interventions. It could be a tremendous cost savings if the STAR assessments were able to be used only for students requiring more diagnostic assessment or progress monitoring and teacher judgments of their students' reading risk were used as the universal screening tool. There were some students in the study identified as at-risk by STAR Reading and not at-risk by the teacher and vice-versa. Without further studies there is no way to determine which method is more accurate. Districts looking for a cost savings could consider utilizing teacher professional judgment as a universal screening tool, but caution should be taken to ensure students are correctly identified.

Besides the cost savings, the use of teachers' risk determinations by a district is an indication of trust in those teachers' professional judgments. Schools and districts are under increasing pressure to prove their effectiveness through the use of student test scores. Ratings systems and focus on school and district accountability have had negative consequences for teachers. Increases in levels of stress and perceived decreases in efficacy (Berryhill, Linney, & Fromewick, 2009) or agency (Robinson, 2012) are results of these systems of accountability and the constraints of education policies which are outside of the teachers' control. A district or school that shows it values its teachers' contributions through the systematic consideration of each teacher's professional judgment could mitigate the negative consequences of the systems of accountability.

Future Research

The purpose of universal screening is to determine which students are at risk for not meeting grade-level benchmarks at year-end. This study compared the risk determinations of teachers to that of a commercially published screening tool, STAR Reading. Studies found the predictive validity of STAR Reading for year-end state assessments ranged from .68-.82 (National Center on Response to Intervention, 2010b). While this is highly predictive it is not perfect and there are students that are incorrectly identified as at risk or not at risk based on STAR Reading alone. Could it be that teachers are better predictors? In the end, when there is disagreement, is the teacher correct or STAR Reading? Further studies could attempt to answer this question by comparing the risk designations and the year-end summative assessments

to determine if there are patterns to which students, by which teachers, may be better predicted by STAR Reading or teacher evaluation. A comparison of the predictive validity of teachers versus STAR Reading could help determine when teachers are better at predicting which students might be at risk.

Further research into how teachers determine which assessment or observational information is most appropriate for each student might provide insight into a holistic approach to risk determination. In this study, some teachers indicated they used different assessments, personal information, or knowledge of work habits depending on the student. It could be helpful to determine if there are patterns to what information is most valuable when teachers are making their risk determinations.

This study focused on elementary school teachers, all of whom have at least some training in reading instruction. It is unclear whether the same results would be found for teachers of younger or older students, in the area of math, or behavior, or for those teachers without training in the subject area. A broader study including a variety of subject areas and teachers with different background training would help determine if these findings are unique to elementary school teachers and students.

Conclusion

Predicting which students are at risk and which do not require supplementary services or supports is a key function of any school system. In addition, schools exist in a system of accountability in which data and student achievement results are a primary focus of attention. School leaders are pressured to accurately predict which students are on track and which are not at multiple points throughout the school year,

and this pressure is often applied to teachers as well. The implementation of a Response to Intervention framework including the utilization of universal screening tools has been a common way schools have approached predicting student risk for groups of students in a system. However, the accurate prediction of risk for large populations of students is different from making individual decisions about individual students. Teachers access and apply many ways of knowing about their students that extend beyond the empirical data available from a universal screening tool such as STAR Reading. STAR Reading is highly predictive of future performance on summative assessments when applied to large groups of students, but less accurate for individual students due to their differences. Differences whose impact are best evaluated by a trained teacher aware of confounding factors such as family tragedies, developmental levels, and personality variances. Prediction of student achievement, at the individual and aggregate levels, could be more accurate when teacher professional judgment is considered as a valid method of determining students' achievement and risk.

References

Adams, J. E. (2016) Education reform – Overview. Retrieved from http://education.stateuniversity.com/pages/1944/Education-Reform.html

Ardoin, S. P., & Christ, T. J. (2008). Evaluating curriculum-based measurement slope estimates using data from triannual universal screenings. School Psychology Review, 37(1), 109-125.

Barbour, K. (2004). Embodied ways of knowing. Waikato Journal of Education, 10, 227-238.

Begeny, J. C., Krouse, H. E., Brown, K. G., & Mann, C. M. (2011). Teacher judgments of students' reading abilities across a continuum of rating methods and achievement measures. School Psychology Review, 40(1), 23-38.

Berryhill, J., Linney, J. A., & Fromewick, J. (2009). The effects of education accountability on teachers: Are policies too stress provoking for their own good? International Journal of Education Policy and Leadership, 4(5), 1-14.

Broudy, H. S. (1979). Tacit knowing as a rationale for liberal education. *Teachers College Record*, *80*(3), 446-462.

Catts, H. W., Petscher, Y., Schatschneider, C., Sittner Bridges, M., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities*, *42*(2), 163-176.

Center on Response to Intervention at American Institutes for Research (2014). *Screening tools chart*. Retrieved from http://www.rti4success.org/star-reading

Coleman, L. J. (2014). The invisible world of professional practical knowledge of a teacher of the gifted. *Journal for the Education of the Gifted*, *37*(1), 18-29.

Conklin, K. R. (1970). The aesthetics of knowing and teaching. *Teachers College Record*, *72*(2), 257-265.

Cowan, R. (2014). Ways of knowing, outcomes and 'comparative education': Be careful what you pray for. *Comparative Education*, *50*(3), 282-301.

Cuban, L. (2011, June 16). *Jazz, basketball, and teacher decision-making*. (Web log). Retrieved from https://larrycuban.wordpress.com/2011/06/16/jazz-basketball-and-teacher-decision-making/

Cubberley, E. P. (1916). *Public school administration: A statement of the fundamental principles underlying the organization and administration of public education*. (pp. 325-340). New York: Houghton Mifflin.

Curtis, V. (2012). *Teachers' Perceptions of Reading Achievement for Kindergarten-3rd Grade Students of Low Socioeconomic Status* (Doctoral dissertation). Retrieved from ProQuest. (3527827)

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52(3)*, 219-232.

Diaz Soto, L., & Tuinhof De Moed, S. (2011). Toward 'our ways of knowing' in the age of standardization. *Contemporary Issues in Early Childhood, 12(4)*, 327-331.

Dowdy, E., Doane, K., Eklund, K., & Dever, B. V. (2011). A comparison of teacher nomination and screening to identify behavioral and emotional risk within a sample of underrepresented students. *Journal of Emotional and Behavioral Disorders, 21(2)*, 127-137.

Dupuis, H., & Hall, K. Community Eligibility Provision (CEP) [PDF document]. Retrieved from http://www.ode.state.or.us/opportunities/grants/nclb/cep-for-title-i-a.pdf

Eklund, K., Renshaw, T. L., Dowdy, E., Jimerson, S. R., Hart, S. R., Jones, C. N., & Earhart, J. (2009). Early identification of behavioral and emotional problems in youth: Universal screening versus teacher-referral identification. The California School Psychologist 14, 89-95.

Fletcher, J. M., & Satz, P. (1984). Test-based versus teacher-based predictions of academic achievement: a three-year longitudinal follow-up. Journal of Pediatric Psychology, 9(2), 193-203.

Frontera, L. S. & Horowitz, R. (1995). Reading and study behaviors of fourth-grade Hispanics: Can teachers assess risk? Hispanic Journal of Behavioral Sciences, 17(1), 100-120.

Fuchs, L.S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. Learning Disabilities Research and Practice, 13, 204-219.

Garcia, J. L., & Ford, M. (2001). Intuition: The other way of knowing. Journal of Professional Counseling: Practice, Theory & Research, 29(1), 80-87).

Good, R., & Kaminski, R. (1996). Assessment for instructional decisions; Toward a proactive/prevention model of decision-making for early literacy skills. *School Psychology Quarterly, 11(4)*, 326-336.

Greer, F. W., Wilson, B. S., DiStefano, C., & Liu, J. (2012). Considering social validity in the context of emotional and behavioral screening. *School Psychology Forum: Research in Practice, 6(4)*, 148-159.

Gudmundsdottir, S. (1991). Ways of seeing are ways of knowing. The pedagogical content knowledge of an expert English teacher. *Journal of Curriculum Studies, 23(5)*, 409-421.

Gurm, B. K. (2013). Multiple ways of knowing in teaching and learning. *International Journal for the Scholarship of Teaching and Learning, 7(1)*, 1-7.

Hauerwas, L., Brown, R., & Scott, A. N., 2013. Specific learning disability and response to intervention: State-level guidance. *Exceptional Children, 80(1)*, 101-120.

Hayes, W. (2008). *No child left behind: Past, present, and future.* New York: Rowman & Littlefield Education.

Hernández, J. C. (2009, February 19). *New education secretary visits Brooklyn school*. The New York Times. Retrieved from http://cityroom.blogs.nytimes.com/2009/02/19/new-education-secretary-visits-brooklyn-school/?_php=true&_type=blogs&_r=1

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59(3)*, 297-313.

Hosp, J. L., Hosp, M. A., & Cole, J. K. (2011). Potential bias in predictive validity of universal screening measures across disaggregation subgroups. *School Psychology Review, 40(1)*, 108-131.

Hughes, C. A. & Dexter, D. D. (2011). Response to intervention: A research-based summary. *Theory Into Practice, 50*, 4-11.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36(4)*, 582-600.

Kapelis, L. (1975). Early identification of reading failure: A comparison of two screening tests and teacher forecasts. *Journal of Learning Disabilities, 8(10)*, 638-641.

Kenny, D. T., & Chekaluk, E. (1993). Early reading performance: A comparison of teacher-based and test-based assessments. *Journal of Learning Disabilities, 26(4)*, 227-236.

Kenrick, D. T., Griskevicius, V., Neuberg, S. L., & Schaller, M. (2010). Renovating the pyramid of needs: Contemporary extensions built upon ancient foundations. *Perspectives on Psychological Science, 5(3)*, 292-314.

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of School Psychology, 52*, 377-405.

Labaree, D. (1997). Public goods, private goods: The American struggle over educational goals. *American Educational Research Journal, 34(1)*, 39-81.

Labaree, D. (2010). *Someone has to fail: The zero-sum game of public schooling*. (pp. 80-105). Cambridge, Ma: Harvard University Press.

Lembke, E. S., McMaster, K. L., & Stecker, P. M. (2009). The prevention science of reading research within a response-to-intervention model. *Psychology in the Schools, 47(1)*, 22-35.

Martin, S. D. & Shapiro, E. S. (2011). Examining the accuracy of teachers' judgments of DIBELS performance. *Psychology in the Schools, 48(4)*, 343-356.

Mausethagen, S. (2013). A research review of the impact of accountability policies on teachers' workplace relations. *Educational Research Review, 9*, 13-33.

McConaghy, J. (1986). On becoming teacher experts: Research as a way of knowing. *Language Arts, 63(7)*, 724-728.

McGlinchey, M. T. & Hixson, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33(2)*, 193-203.

McGuinn, P. J. (2006). *No child left behind and the transformation of federal education policy, 1965-2005*. Lawrence, Kansas: University of Kansas Press.

Mead, W. B. (2007). "I know more than I can tell": The insights of Michael Polanyi. *Modern Age, 49(3)*, 298-307.

Mellard, D. F., McKnight, M., & Woods, K. (2009). Response to intervention screening and progress-monitoring practices in 41 local schools. *Learning Disabilities Research & Practice, 24(4)*, 186-195.

Merriam, S. B. (2009). **Qualitative research: A guide to design and implementation.** San Francisco: Jossey-Bass.

Messick, S. (1980). Test validity and the ethics of assessment. **American Psychologist, 35(11)**, 1012-1027.

Mintrop, H. (2012). Bridging accountability obligations, professional values and (perceived) student needs with integrity. **Journal of Educational Administration, 50(5)**, 695-726.

National Center on Response to Intervention (March, 2010a). **Essential components of RTI – A closer look at response to intervention.** Washington, DC: U.S. Department of Education, Office of Special Education Programs, National Center on Response to Intervention.

National Center on Response to Intervention (March, 2010b). **Users guide to universal screening tools chart.** Washington, D.C. U.S. Department of Education, Office of Special Education Programs, National Center on Response to Intervention.

National Commission on Excellence in Education (1983). **A nation at risk: The imperative for educational reform.** Washington, D.C.U.S. Government Printing Office.

Nye, M. J. (2015). Foreward. Polanyi, M. (1958). *Personal Knowledge: Towards a Post-Critical Philosophy (Enlarged Edition with a new Forward by Mary Jo Nye)*. Chicago: The University of Chicago Press.

Oregon Department of Education. (2015). *Fall membership report 2014-2015* [Excel Spreadsheet]. Retrieved from http://www.ode.state.or.us/search/page/?=3225

Parker, D. C., Zaslofsky, A. F., Burns, M. K., Kanive, R., Hodgson, J., Scholin, S. E., & Klingbeil, D. A. (2015). A brief report of the diagnostic accuracy of oral reading fluency and reading inventory levels for reading failure risk among second- and third-grade students. *Reading and Writing Quarterly, 31*, 56-67.

Payne, B. D., & Payne, D. A. (1991). The ability of teachers to identify academically at-risk elementary students. *Journal of Research in Childhood Education, 5(2)*, 116-126.

Polanyi, M. (1961). Knowing and being. *Mind, 70(280)*, 458-470.

Polanyi, M. (1962). Tacit knowing: Its bearing on some problems of philosophy. *Reviews of Modern Physics, 34(4)*, 601-616.

Polanyi, M. (1966a). The logic of tacit inference. *Philosophy, 41(155)*, 1-18.

Polanyi, M. (1966b). *The Tacit Dimension.* Chicago: The University of Chicago Press.

Renaissance Learning. (2015). *STAR reading™ technical manual.* Wisconsin Rapids, WI: Author.

Renaissance Learning, (2016). Numbers you can count on. In *About us*. Retrieved from http://www.renaissance.com/about-us

Robinson, K., & Aronica, L. (2015). *Creative schools: The grassroots revolution that's transforming education.* New York: Viking.

Robinson, S. (2012). Constructing teacher agency in response to the constraints of education policy: Adoption and adaptation. *The Curriculum Journal*, *23(2)*, 231-245.

Rose, M. (2009). *Why school: Reclaiming education for all of us.* (pp. 25-41). New York: The New Press.

Rowe, S., Witmer, S., Cook, E., & daCruz, K. (2014). Teachers' attitudes about using curriculum-based measurement in reading (CBM-R) for universal screening and progress monitoring. *Journal of Applied School Psychology, 30(4)*, 305-337.

Senechal, D. (2013). Measure against measure: Responsibility versus accountability in education. *Arts Education Policy Review, 114(2)*, 47-53. doi: 10.1080/10632913.2013.769828

Shapiro, E. S., & Gebhardt, S. N. (2012). Comparing computer-adaptive and curriculum-based measurement methods of assessment. *School Psychology Review, 41(3)*, 295-205

Simpson, E. L. (1971). Other ways of knowing: Educational goals. *Teachers College Record, 72(4)*, 559-565.

Speece, D. L., Schatschneider, C., Silverman, R. Pericola Case, L., Cooper, D. H. & Jacobs, D. M. (2011). Identification of reading problems in first grade within a response-to-intervention framework. *The Elementary School Journal, 111(4)*, 585-607).

Suen, H. K. (1988). Agreement, reliability, accuracy, and validity: Toward a clarification. *Behavioral Assessment, 10*, 343-366.

Szesztay, M. (2004). Teachers' ways of knowing. *ELT Journal, 58(2)*, 129-136.

Taubman, P. M. (2009). *Teaching by numbers: Deconstructing the discourse of standards and accountability in education*. New York: Routledge.

VanDerHeyden, A. M. (2013). Universal screening may not be for everyone: Using a threshold model as a smarter way to determine risk. *School Psychology Review, 42(4)*, 402-414.

Vinovskis, M. A. (2009). *From a nation at risk to no child left behind: National education goals and the creation of federal education policy.* New York: Teachers College Press.

Wansart, W. L. (1995). Teaching as a way of knowing: Observing and responding to students' abilities. *Remedial and Special Education, 16(3)*, 166-177.

Yorks, L. & Kasl, E. (2006). I know more than I can say: A taxonomy for using expressive ways of knowing to foster transformative learning. *Journal of Transformative Education, 4(1)*, 43-64.

Appendix A

Reading Risk Evaluation Teacher Tracking Tool

Teacher: Suzanne Teacher			**School: Northwest Elementary**		
Student First	**Student Last**	**Pupil Number**	**Grade**	**Reading Risk Now**	**What information did you use to decide on the student's reading risk? Why did you choose this particular level of risk?**
Donald	Smith	123457	3		
Sharon	Stone	123458	3		
Guillaume	Busch	123459	3		
Carly	Mabbot	123460	3		
Marsha	Kirschman	123461	3		
Keith	Shireman	123462	3		
Anders	Pickle	123463	3		
Iris	Harris	123464	3		
Sophia	McGloghlon	123465	3		
Edme	Brown	123466	3		
Kael	Weins	123467	3		
Laura	Peterson	123468	3		
Stephanie	Carter	123469	3		

Appendix B

Teacher Consent Form

Teacher Consent Form

You are invited to participate in a research study conducted by Leigh Anne Scherer, a doctoral student at the University of Portland. I hope to learn more about how teachers' perceptions of their students reading risk compares with the reading risk determined by the STAR Reading assessment. You were selected as a possible participant in this student because you are a third or fourth grade teacher in the [redacted] School District. I appreciate your consideration.

If you decide to participate you will be asked to complete a survey with your perception of each of your students' reading risk (At/Above Benchmark, On Watch, Intervention, or Urgent Intervention), as well as some demographic information. This survey should be completed BEFORE students complete the fall STAR Reading screening and should be based on the best information you have available. I anticipate this will take no more than 1/2 hour to complete.

It is possible you may be uncomfortable due to the fact that this is the beginning of the school year and you may not believe you have a lot of information regarding your students. The intent of the study is to better understand teachers' abilities to perceive their students risk and there is no correct or expected performance. Although I cannot guarantee you personally will receive any benefits from this research, it will provide valuable information regarding teachers' abilities to recognize reading risk in their students.

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law. Your identity will be kept confidential by removing your name and inserting an identification number in its place.

Your participation is completely voluntary. Your decision whether or not to participate will not affect your relationship with the [redacted] School District. If you decide to participate, you are free to withdraw your consent and discontinue participation at any time without penalty.

If you have any questions about the study, please feel free to contact Leigh Anne Scherer ([redacted], [redacted] or scherer11@up.edu) or Dr. Phyllis Egby, Assistant Professor, School of Education, University of Portland (503-943-7259, egby@up.edu). If you have questions regarding your rights as a research subject, please contact the Institutional Review Board (IRB@up.edu). A copy of this form will be emailed to you.

There are spaces below the signature field for demographic information about you as a teacher including: number of years teaching, number of years teaching this grade, gender, and race/ethnicity. When completing the number of years teaching and number of years teaching this grade please include the current year in your calculation. Include as a full year any school year spent teaching at least 1/2 time. Please do not include student teaching or substitute teaching (unless as a long-term sub in a single position for more than 1/2 the school year) in your calculation. For the race/ethnicity field please include as many races/ethnicities as you feel represent your background. For the gender field please indicate your gender identity. This information is optional and will only be used for analyzing the results. If you decline to respond please simply leave the fields blank.

Your signature below indicates that you have read and understand the information provided above, that you willingly agree to participate, that you may withdraw consent at any time and discontinue participation without penalty, that you will receive a copy of this form, and that you are not waiving any legal claims.

Name (Please Print):	**Date:**
Signature:	**Grade:**

Number of Years Teaching:	**Number of Years Teaching This Grade:**
Gender:	**Race/Ethnicity**

Appendix C

Statistical Analyses by Race/Ethnicity

STAR Reading Scale Scores for All Students Disaggregated by Race/Ethnicity

	n	Min.	Max	M	SD
American Indian/Alaskan Native	7	71	527	283	163
Asian	45	81	906	440	219
Black/African American	11	84	631	254	164
Hispanic	154	61	930	261	149
Native Hawaiian/Pacific Islander	8	88	501	341	161
Multi-Racial	52	77	1105	428	218
White	514	67	1299	407	205

STAR Reading Scale Scores for 3rd Grade Students Disaggregated by Race/Ethnicity

	n	Min.	Max	M	SD
American Indian/Alaskan Native	6	71	527	272	176
Asian	27	81	852	402	223
Black/African American	10	84	391	216	112
Hispanic	117	61	645	239	139
Native Hawaiian/Pacific Islander	7	88	501	332	172
Multi-Racial	28	77	650	330	140
White	319	67	968	358	175

STAR Reading Scale Scores for 4th Grade Students Disaggregated by Race/Ethnicity

	n	Min.	Max	M	SD
American Indian/Alaskan Native	1	347	347		
Asian	18	214	906	498	206
Black/African American	1	631	631		
Hispanic	37	62	930	332	169
Native Hawaiian/Pacific Islander	1	403	403		
Multi-Racial	24	232	1105	542	240
White	195	70	1299	487	225

STAR Risk Ratings All Students Disaggregated by Race/Ethnicity

	At/Above Benchmark		On Watch		Intervention		Urgent Intervention	
	n	%	n	%	n	%	n	%
American Indian/Alaskan Native	2	29	2	29	1	14	2	29
Asian	28	62	5	11	4	9	8	18
Black/African American	3	27	2	18	2	18	4	36
Hispanic	40	26	22	14	35	23	57	37
Native Hawaiian/Pacific Islander	4	50	2	25			2	25
Multi-Racial	29	56	10	19	7	14	6	12
White	305	59	69	13	57	11	83	16

STAR Risk Ratings 3rd Grade Students Disaggregated by Race/Ethnicity

	At/Above Benchmark		On Watch		Intervention		Urgent Intervention	
	n	%	n	%	n	%	n	%
American Indian/Alaskan Native	2	33	2	33			2	33
Asian	16	59	11	3	4	15	4	15
Black/African American	2	20	2	20	2	20	4	40
Hispanic	29	25	18	15	26	22	44	38
Native Hawaiian/Pacific Islander	4	57	1	14			2	29
Multi-Racial	15	54	5	18	4	14	4	14
White	186	58	41	13	36	11	56	18

STAR Risk Ratings 4th Grade Students Disaggregated by Race/Ethnicity

	At/Above Benchmark		On Watch		Intervention		Urgent Intervention	
	n	%	n	%	n	%	n	%
American Indian/Alaskan Native					1	100		
Asian	12	67	2	11			4	22
Black/African American	1	100						
Hispanic	11	30	4	11	9	24	13	35
Native Hawaiian/Pacific Islander			1	100				
Multi-Racial	14	58	5	21	3	13	2	8
White	119	61	28	14	21	11	27	14

Teacher Risk Ratings All Students Disaggregated by Race/Ethnicity

	At/Above Benchmark		On Watch		Intervention		Urgent Intervention	
	n	%	n	%	n	%	n	%
American Indian/Alaskan Native	2	29	2	29	1	14	2	29
Asian	30	67	4	9	5	11	6	13
Black/African American	3	27	2	18	3	27	3	27
Hispanic	41	27	30	20	53	34	30	20
Native Hawaiian/Pacific Islander	5	63			1	13	2	25
Multi-Racial	30	58	9	17	7	14	6	12
White	299	58	86	17	60	12	69	13

Teacher Risk Ratings 3rd Grade Students Disaggregated by Race/Ethnicity

	At/Above Benchmark		On Watch		Intervention		Urgent Intervention	
	n	%	n	%	n	%	n	%
American Indian/Alaskan Native	2	33	2	33	1	17	1	17
Asian	18	67	3	11	1	4	5	19
Black/African American	2	20	2	29	3	30	3	30
Hispanic	36	31	23	20	36	31	22	19
Native Hawaiian/Pacific Islander	5	71					2	29
Multi-Racial	14	50	6	21	3	11	5	18
White	189	59	55	17	23	7	52	16

Teacher Risk Ratings 4th Grade Students Disaggregated by Race/Ethnicity

	At/Above Benchmark		On Watch		Intervention		Urgent Intervention	
	n	%	n	%	n	%	n	%
American Indian/Alaskan Native							1	100
Asian	12	67	1	6	4	22	1	6
Black/African American	1	100						
Hispanic	5	14	7	19	17	46	8	22
Native Hawaiian/Pacific Islander					1	100		
Multi-Racial	16	67	3	13	4	17	1	4
White	110	56	31	16	37	19	17	9

Overall Risk Match by Race/Ethnicity

	STAR – Not Risk		STAR – Risk	
	n	%	n	%
Teacher Not Risk				
American Indian/ Alaskan Native	3	43	1	14
Asian	31	69	3	7
Black/African American	4	36	1	9
Hispanic	48	31	23	15
Native Hawaiian/Pacific Islander	5	63		
Multi-Racial	33	64	6	12
White	340	66	45	9
Teacher - Risk				
American Indian/ Alaskan Native	1	14	2	29
Asian	2	4	9	20
Black/African American	1	9	5	46
Hispanic	14	9	69	45
Native Hawaiian/Pacific Islander	1	13	2	25
Multi-Racial	6	12	7	14
White	34	7	95	19

3rd Grade Risk Match by Race/Ethnicity

	STAR – Not Risk		STAR – Risk	
	n	%	n	%
Teacher Not Risk				
American Indian/ Alaskan Native	3	50	1	17
Asian	18	67	3	11
Black/African American	3	30	1	10
Hispanic	38	33	21	18
Native Hawaiian/Pacific Islander	5	71		
Multi-Racial	17	61	3	11
White	213	67	31	10
Teacher - Risk				
American Indian/ Alaskan Native	1	17	1	17
Asian	1	4	5	19
Black/African American	1	10	5	50
Hispanic	9	8	49	42
Native Hawaiian/Pacific Islander			2	29
Multi-Racial	3	11	5	18
White	14	4	61	19

4th Grade Risk Match by Race/Ethnicity

	STAR – Not Risk		STAR – Risk	
	n	%	n	%
Teacher Not Risk				
American Indian/ Alaskan Native				
Asian	13	72		
Black/African American	1	100		
Hispanic	10	27	2	5
Native Hawaiian/Pacific Islander				
Multi-Racial	16	67	3	13
White	127	65	14	7
Teacher - Risk				
American Indian/ Alaskan Native			1	100
Asian	1	6	4	22
Black/African American				
Hispanic	5	14	20	54
Native Hawaiian/Pacific Islander	1	100		
Multi-Racial	3	13	2	8
White	20	10	34	17

Percent Exact Agreement Risk Determination

	Percent Exact Agreement Risk/No Risk		STAR Only Risk	Teacher Only Risk
	n	%	%	%
All American Indian/Alaskan Native	5	72	14	14
3rd Grade American Indian/Alaskan Native	4	67	17	17
4th Grade American Indian/Alaskan Native	1	100		
All Asian	40	89	7	4
3rd Grade Asian	23	86	11	4
4th Grade Asian	17	94		6
All Black/African American	9	82	9	9
3rd Grade Black/African American	8	80	10	10
4th Grade Black/African American	1	100		
All Hispanic	117	76	15	9
3rd Grade Hispanic	87	75	18	8
4th Grade Hispanic	30	81	5	14
All Native Hawaiian/Pacific Islander	7	88		13
3rd Grade Native Hawaiian/Pacific Islander	7	100		
4th Grade Native Hawaiian/Pacific Islander				100
All Multi-Racial	40	78	12	12
3rd Grade Multi-Racial	22	79	11	11
4th Grade Multi-Racial	18	85	13	13
All White	435	85	9	7
3rd Grade White	274	86	10	4
4th Grade White	161	82	7	10