1-1-2010

Achieving Congruence: Building a Case for Implementing a District-Wide Interim Benchmark Assessment that is Aligned with a Balanced Literacy Framework

Theodore Feller
Portland State University

Achieving Congruence: Building a Case for Implementing a District-Wide Interim Benchmark Assessment that is Aligned with a Balanced Literacy Framework

by

Theodore Feller

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

Doctor of Education
in
Educational Leadership:
Administration

Dissertation Committee:
Tom Chenoweth, Chair
Pat Burk
Steve Isaacson
Jason Ranker
Susan Conrad

Portland State University
2010

ABSTRACT

For generations—and certainly for the last 30 years—proponents of traditional and progressive philosophies have argued over how best to educate our children. Although this debate is often carried out in the political and academic spheres, the difficulties created by not being able to resolve the differences between the two belief systems become blatantly clear in the pedagogy of early literacy. On the one hand, traditionalists argue for a direct and explicit instructional methodology, and on the other hand, progressives advocate for Whole Language or Balanced Literacy instruction. The classroom often becomes a battlefield as advocates of these opposing schooling paradigms struggle with each other. Differences emerge about which skills and what knowledge are the most important for students to master. Conflicts arise over which methodology is most effective in ensuring that students gain access to bodies of knowledge. The result is that the real world of classroom instruction often becomes a mish-mash of content and strategies that derive from both philosophies. Student assessments frequently contribute to the confusion because they are not aligned with the knowledge and skills students are expected to acquire as well as with the strategies teachers use. Without assessments that are tightly coupled with the underlying philosophy of an instructional program, with classroom practice, and with high-stakes summative assessments, it is extremely difficult for both teachers and administrators to have confidence that they are offering their students the best possible learning opportunities.

Interim/benchmark assessments are vital tools for linking classroom instruction with year-end assessments and an essential element of any comprehensive assessment system. Currently, the Dynamic Indicator of Beginning Early Literacy Skills, commonly referred to as DIBELS, is a widely used interim/benchmark assessment. It serves many districts and schools quite well. However, many progressive educators believe that the DIBELS assessment is not well-aligned with a Balanced Literacy approach. In this dissertation the author examines the following essential question about early literacy interim/benchmark assessments: (a) Is the relationship between the assessed level on the Developmental Reading Assessment (DRA), which fits within a Balanced Literacy framework, and student's performance on high stakes accountability test as strong as the relationship of DIBELS to these same tests; and (b) does the DRA have a degree of predictive validity comparable to DIBELS?

The study demonstrated a strong relationship between the DRA and performance on OAKS and that the DRA has a degree of predictive validity that is comparable to DIBELS. The results from the study support the claim that a curriculum-based measure, such as the DRA, can be used as a literacy screening assessment to detect potential reading difficulties. These results give support to progressive educators who wish to have a viable alternative DIBELS.

ACKNOWLEDGMENTS

Gratitude and appreciation to my advisor Tom Chenoweth who guided me through this process. Tom has been advising me off and on through my academic life for more than 20 years; I am deeply thankful for that, Tom. I will always be indebted to you.

Deep appreciation to the instructional leadership of the Centennial School District—Amy Olson-Rocha, Andrea Sande, Carrie McGraw and Denene Holbrooke—who have taught me so much about literacy and have been so passionate about their work.

My two incredible children, Katie and David, brought me love, joy and incredible happiness every day. Their encouragement, sacrifice unconditional love and laughter were an enormous help all through this journey. Katie and David, you always believed in me and wouldn't let me give up. And to my colleague and best friend, Lori Silverman, your support, encouragement, and patience helped me to keep going.

Most of all for the hours of council, editing, supporting and believing in me when I didn't believe in myself—to my mothers, Mary Beth Van Cleave and Mary Kinnick, I couldn't have done this without you.

TABLE OF CONTENTS

Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Background

A few years ago, the district where I worked was awarded a joint grant from the federal government and the University of Oregon that centered on intensive interventions for primary students who were having difficulty learning to read. This phonics-based and skill-oriented program ran counter to my own beliefs about effective approaches to both teaching and assessing reading. When the grant coordinators met with our district administrators, I found myself asking them some very hard questions. After the meeting, I began to wonder why my hostility toward the coordinators had been so intense. Even though I disagreed with their pedagogical approach to reading, I was also puzzled and embarrassed by my own behavior.

A few years later and working in another district, I found myself in a similar situation. I attended a meeting with a professional development coordinator for the state of Oregon who had just left a position with the United States Department of Education as a regional coordinator for the Reading First initiative. She and I got into a heated conversation about the services that would be available for our school district as we undertook a new initiative. The only professional development the State was prepared to offer relied on the same skill-based and phonics-based pedagogy that had frustrated me in my former school district.

I begin this dissertation with these two personal stories in order to give a face to the philosophical debate that I and other educators around the country have been caught up in for generations and, also, to make my own beliefs transparent. As I reflect on my 23 years in education I realize that nearly all my differences with other professionals can be traced back to the tension between two very different and often polarizing points of view about schooling. This tension is rooted in competing views about the nature of education and the purpose of schooling—the traditional and the progressive:

> Traditional concepts of schooling emphasize the primacy of subject matter, the importance of passing an inherited body of knowledge on to the young, drill, memorization of facts, the authority of the teacher, and formal instructional methods.... Such "progressive" ideas and classroom instructional practices as the primacy of practical experience in learning (learning by doing), child-centered education, individual instruction, informality in the classroom, group discussion, team learning, and laboratory instruction had become defining characteristics of good teachers and good schools. (Owens, 2001, p. 9)

These differences appear to be no closer to resolution today than before. In fact, in the area of educational research, evaluation and assessment, they have become increasingly stark over the last 30 years (Murphy & Louis, 1999). They have strongly influenced the development of comprehensive accountability systems and specifically classroom assessment systems. I am particularly interested in the latter because they are an essential element of any school district's comprehensive assessment plan.

In thinking about how to design an effective assessment plan for my school district, I found that I could not really understand or evaluate assessment systems without locating them within a broader context because they take very different forms depending on whether they come from a traditional or progressive mind-set. It seems to me that the

traditional concept of schooling has a fundamentally behaviorist perspective on teaching and learning, while the progressive view takes a constructivist perspective.

From a traditionalist's perspective the relationship between the knower and the known is clear and direct. Knowledge can be defined through a set of understood skills, facts, and concepts as well as the cognitive and behavioral strategies that a learner uses. Because what needs to be known can be defined explicitly, this knowledge can be brought together in a unifying way. E. D. Hirsch's (cited in Owens, 2001) *4552 Facts All Students Should Know* is a recent example of this perspective. For the traditionalist, the issue is not how and why the learner acquired these facts or why they might be important to the learner. For traditionalists, learning is the process of expanding the behavioral repertoire (*i.e.* facts, skills and concepts), rather than building on the learner's ideas and thinking. For example, the traditionalist is interested not in how a student understands and learns the *4552 Facts* but in how a student can be led to behave in such a way he can master certain behaviors such correctly identifying these Facts when they asked on a test or reciting them when asked by the teacher (Phillips & Soltis, 1998). A student learns these behavioral and cognitive repertoires through practice. The teacher is seen as an expert who possesses the knowledge and transmits it to the learner through manipulation, while the learner is a passive receiver of information (Langer, 1997; Phillips & Soltis, 1998).

Because a constructivist lens views reality as subjective, what can be known is ever changing and individually or socially constructed. Whole Language and Applied Based Mathematics are recent examples of curricula based on this philosophy. Constructivism assumes that knowledge is best understood through its application.

Learning is not practicing the same skill over and over or applying explicitly taught concepts by rote memory; rather it is the application of strategies learned and practiced through interactions with peers. Constructivists view learning as the process of acquiring intellectual and social tools that enable the learner to determine what is to be known instead of relying on a required set of behavioral or cognitive repertoires. Teachers are most effective when they are facilitators of learning rather than purveyors of knowledge. In a constructivist classroom, the teacher becomes a guide or a facilitator of knowledge, and the learner constructs knowledge (Ernest, 1994; Langer, 1997; Phillips & Soltis, 1998).

These complex and contrasting views influence virtually every area of our educational system, including the field of assessment and accountability (Murphy & Louis, 1999). They have had a profound influence on the types of instruments that have been designed for both schools and students for more than 100 years. The debate over appropriate student assessment, evaluation and accountability systems is rooted in questions such as these: What knowledge, skills, or behavioral repertoires do we want students to master? How do we measure them? How does this debate play out in the quest for sound evaluation and accountability?

From a traditional or behaviorist perspective, scientific knowledge is acquired through efforts to understand objective reality in a manner that is free from bias. Reality can be positively defined and understood, and causal relationships can be absolutely established. Scientific knowledge is quantifiable and empirically validated. Because assessment is straightforward, consistent, and can be universally applied, it lends itself easily to accountability standards (Phillips & Soltis, 1998).

On the other hand, progressives contend that taking scientific knowledge and its causal relationships as natural and inevitable realities makes it possible for the privileged class to maintain its dominance. They view scientific knowledge as culturally and socially constructed through social relationships that are situationally unique. Reality is at least partly subjective, and scientific knowledge is not an absolute, inevitable reality. Scientific knowledge, however, does have its uses: it establishes causal relationships that help us understand our cultural constructions and social relationships, and it can expose the inequalities and social injustices of our society (Langer, 1997; Phillips & Soltis, 1998).

From a constructivist perspective the lives of the students and their families, students' relationships with each other and with adults in the school, and the teaching and learning that occurs in classrooms throughout the school and community define what knowledge is important. The quality of this living system is what needs to be assessed and evaluated (Ernest, 1994; Langer, 1997; Phillips & Soltis, 1998). Assessments emphasize how students make meaning of what they are learning rather than the behavioral repertoires they need to master (Calkins, 2001). These types of assessments do not fit easily within a context of high stakes testing (Popham, 2008).

Many educators cherish the false hope that evaluation and assessment can provide powerful tools to help resolve the question about which of these paradigms best facilitates student learning. We naturally look to researchers and test developers for definitive answers. However, over last 30 years, a tacit understanding has emerged within the field of research and evaluation and specifically student assessment that is slanted toward the traditional point of view. This understanding controls the ways in which we

frame our research, evaluation, or assessment questions. Because the designs we use to construct our studies, assessments or tests reinforce the traditional paradigm over the progressive paradigm, they inevitably serve the interests of some educators while marginalizing others. In the *Handbook of Research on Educational Administration* Donmore (Murphy & Louis, 1999) described the confusion and politicalization that results from not clearly understanding our underlying assumptions:

> The complex and contradictory conceptualizations assembled under the big tents we construct to accommodate our differences are more likely to confuse than enlighten administrators and policy makers who look to the academy for an alternative to brute politics as a way to make decisions. Indeed, when we recommend contradictory things, we almost guarantee that research will be used selectively as political weapon instead of as a tool to help resolve educational disputes intellectually rather than through the use of brute power. (p. 36)

Nowhere is this dispute more contentious than in the area of early literacy. Behaviorists such as Engelmann (2008), Simmons et al. (2000) and Torgesen (Kosanovich, Ladinsky, Nelson, & Torgesen, 2007; Mathes & Torgensen, 1998) have argued for a skills first *Direct Instruction* pedagogy. Progressives, such as Meier (2000, 2010), Calkins (2001), and R. L. Allington (personal communication, January 2010) have favored the more child centered Whole Language or its more centrist descendant Balanced Literacy. At this point the terms themselves have become highly charged. Direct Instruction is an anathema to the progressives, and the traditionalists talk about Whole Language as if it were educational malpractice. Although the middle ground is almost impossible to find, it is possible to argue that both pedagogies have their strengths. Proponents on each side can cite research to support their claims about their pedagogy's effectiveness in educating young readers. I actually believe that each approach has merit and, if practiced by knowledgeable, professional teachers, will have strong outcomes in

terms of student achievement (National Reading Panel [NRP], 2000; P. D. Pearson, 2004).

If this is the case, why are people so upset? Here it is probably important to note that educators who favor Balanced Literacy are more upset than those who are comfortable with Direct Instruction (P. D. Pearson, 2004). One of the primary concerns of Balanced Literacy advocates has to do with what has been happening nationally in the area of early literacy assessment practices (R. L. Allington, personal communication, January 2010).

Traditionalists have taken the lead in this area by developing interim/benchmark assessment tools that regularly measure student progress in achieving reading proficiency by the end of third grade as determined on standardized assessments (P. D. Pearson, 2004). The Dynamic Indicator of Beginning Early Literacy Skills (DIBELS) (University of Oregon, 2010) assessment tool has had success in identifying which students are on track to reach grade level standards and which are not. DIBELS appears to be a tool that all educators can use to help determine how effective their instructional program is and which students need support to reach benchmark. Because of numerous studies that demonstrate its strength, many school districts now mandate DIBELS, and state departments of education also support it.

Problem Statement

Even though it is widely used and has a body or research behind it, many progressive educators are uncomfortable with an interim/benchmark assessment such as DIBELS because it is based on a traditional pedagogical philosophy that relies on discrete skill development and prefers scripted reading programs (Allington, 2009; R. L.

Allington, personal communication January 2010). Teachers have very little latitude to use their judgment in assessing whether a student is learning to read; instead they administer prescribed tests (Howe, 2008; University of Oregon, 2010). Their role in making both curricular and assessment decisions is curtailed (Howe, 2008). From a traditionalist's standpoint, these limits ensure consistency both in instruction and in assessment and lead to better student outcomes (Engelmann, 2008; Howe, 2008; Kame'enui & Simmons, 1990). K. Howe, a former regional director of Reading First, while speaking at a 2008 conference sponsored by the Oregon Department of Education (ODE) stated, "We don't want teachers to think, we just want them to administer the prescribed curriculum."

Statements such as these run contrary to the progressives' fundamental beliefs. From their perspective, assessment tools, such as DIBELS, have significant shortcomings: they over analyze skills and behavioral repertoires while under examining how the learner makes meaning out of what is read, and they attempt to minimize rather than enhance and integrate the teacher's role in the assessment process and in the instructional decision making process. Progressive educators believe these shortcomings have serious implications for both students and practitioners (P. D. Pearson, 2004). In fact when progressive educator R. L. Allington (personal communication, January 2010) spoke at the Rethinking Literacy Conference in Portland, Oregon, he began his keynote by saying, "I was hesitant to come to Oregon because it is the home of DIBELS." Allington is not shy about claiming that assessments such as DIBELS have done significant damage to literacy instruction across the nation.

There are reasons, both political and philosophic, why Direct Instruction and DIBELS have become the coin of the assessment realm. Of all the claims, the most frequent and persuasive is that DIBELS is backed by scientific research. Indeed, numerous studies have been conducted that demonstrate a strong correlation between a student's performance on DIBELS and high stakes summative assessments (Buck & Torgesen, 2002; Hosp & Fuchs, 2005; Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008; Shaw & Shaw, 2002; Silberglitt, Burns, Madyun, & Lail, 2006; Wilson, 2005). The difficulty that progressive educators face is that, at this time, they are unable to cite an equally impressive body of empirical research to validate their use of meaning first interim assessments.

This dissertation examines one school district's attempt to develop and implement an interim literacy assessment that is aligned with its progressive pedagogical philosophy. This is a district of approximately 6,000 students located in East Multnomah County of Portland, Oregon. From this point on, it is referred to as *Emerald*.

Emerald School District's instructional program has a long history and culture of constructivism. Originally, its early literacy curriculum was based on a Whole Language approach and then later moved to a Balanced Literacy framework with Reading Recovery as its intervention program. Although the district leadership (i.e., Superintendent, Director of Curriculum, Principals and Literacy Coaches) was well-satisfied with its curriculum, it was not comfortable with its assessment practices and was eager to develop a comprehensive assessment system. As Emerald's Director of Assessment, it fell to me to facilitate the development of an assessment system in which they could be confident. As the Leadership Team analyzed their assessment practices, it became very apparent

that they knew it was important to have a reliable interim assessment tool, but DIBELS posed a real problem for them. They viewed DIBELS as relying too heavily on skills and as not giving enough attention to how students engaged with text. Furthermore, it was closely tied to the scripted Reading First program that the United States Department of Education promoted to the exclusion of the meaning first Balanced Literacy framework that Emerald was using. It was clear to the Leadership Team that DIBELS was an awkward, if not impossible, fit with its philosophy and curriculum and would therefore be virtually impossible to implement district wide. The tension on the Leadership Team was played out among teachers on the ground level. Naturally, their professional development was based on a Balanced Literacy framework, and, when their discussion turned to assessment, DIBELS bashing could be heard throughout the district's elementary schools.

As the newly appointed Director of Assessment, I was charged with helping to find a solution to this problem. The district was already using a progressive interim reading assessment, the Developmental Reading Assessment (DRA) (Pearson Education, 2009), but its implementation was inconsistent. My task was two-fold—to ensure that the DRA was being well-implemented and to figure out whether or not it could serve as a good alternative to DIBELS. Over a span of 2 years, I worked with the literacy coaches, reading specialist and the curriculum director to put an effective interim assessment in place. Our roles were well-defined: the curriculum specialists decided what assessment would be used and how it would be used; my job was to facilitate its implementation. I was excited about taking up the challenge because I, and my Emerald colleagues, shared the same progressive point-of-view, and we were all eager to see the DRA being used

effectively. My only reservation about embarking on this initiative was that I was unsure about whether or not the DRA was as strong an interim assessment as DIBELS. I felt that it was important to study the comparative effectiveness of these two interim assessments in order to reassure myself as well the Leadership Team and teachers of Emerald that we were on the right track. Without findings that demonstrate its efficacy, Emerald, and other progressive districts, are understandably hesitant to rely on it.

My study examines the effectiveness of the DRA as a screening tool for students who potentially have reading difficulties and its ability to regularly measure student progress in achieving reading proficiency in the intermediate grades (i.e., third grade through sixth grade). It addresses the claim that there is no empirical research analyzing and evaluating the predictive validity of the DRA on passing an end of year high stakes accountability test. This dissertation examines the following essential questions about early literacy interim/benchmark assessments: (a) Is the relationship between the assessed level on the DRA and students' performance on high stakes accountability tests as strong as the relationship of DIBELS to these same tests; and (b) does the DRA have a degree of predictive validity comparable to DIBELS?

Because I serve both as the Director of Assessment for the Emerald School District and as the researcher for this study, I further discuss any potential for researcher bias in chapter 3.

Significance of the Study

This research will be important to educators in the Emerald School District, and potentially, to other progressive educations, for several reasons. First, if there is an assessment tool that is equally valid and also compatible with their Balanced Literacy

curriculum, they can better integrate assessment within their instructional programs, and these assessments can inform their instructional decisions. Second, they will be able increase the instructional time during a school year because students will not be subjected to duplicate assessments, and more classroom time can be devoted to teaching. Finally, if teachers have access to assessments that are aligned with their pedagogical and epistemological beliefs, they will trust what the assessment is telling them. If they can trust the results, they can truly link their instruction to high stakes accountability results without having to narrow their instruction and to compromise what they believe are the best instructional practices.

Conclusion

This study investigated the correlation and predictive value of a Balanced Literacy interim assessment such as the DRA on the Oregon Assessment of Knowledge and Skills (OAKS) in order to establish its utility as a viable alternative to DIBELS. Before beginning this exploration, a Definition of Terms is provided that is intended to facilitate the discussion to follow. Chapter 2 explores the literature related to interim literacy assessments including their function, purpose and underlying philosophies. Chapter 3 describes the purpose and methodology of this study. Chapter 4 analyzes the results of the study, and finally chapter 5 summarizes the results of the research, suggests further lines of research, and makes recommendations to the policy makers of Emerald School District as well as to other districts wishing to adopt the DRA as an interim literacy assessment.

Definition of Terms

Accuracy of Reading: The number of words read minus the number of miscues not self-corrected divided by the number of words read.

At Risk Student Classification: This student is not on-track or predicted to be not on-track to meet grade level reading standards.

Balanced Literacy: A differentiated instructional approach to teaching reading based on a pedagogy which asserts that reading skills are best taught in the context of authentic reading activity. The typical balanced literacy framework consists of these components: Read Aloud, Shared Reading, Independent Reading, and Word Work.

Curricular and Instructional Congruence: Instruction that is carefully planned and mutually supported in both remedial and regular programs in order to provide at-risk students with the content and strategies needed for achieving success in the regular classroom. Instructional programs that have curricular congruence employ similar philosophies, instructional materials, instructional methods and student activities, reading strategies, and reading goals across all their programs.

Curriculum Based Measurements: A type of assessment used to monitor student performance throughout an entire year's curriculum.

Developmental Reading Assessment (DRA): A method for assessing and documenting primary students' development as readers over time. Its purpose is to identify students' reading level, defined as a text on which students meet specific criteria in terms of accuracy, fluency, and comprehension. Additional purposes include the following: identifying students' independent reading strengths and weaknesses, planning for instruction, monitoring reading growth, and, for the grades 3-5, preparing students to

meet classroom testing expectations, and, finally, providing information to teachers, schools, and region regarding reading achievement.

Direct Instruction: An explicit, carefully sequenced and scripted model of literacy instruction based on a pedagogy that asserts that all details of instruction must be standardized and controlled in order to minimize the chance of students' misinterpreting the information being taught and to maximize the reinforcing effect of instruction.

Dynamic Indicator of Beginning Early Literacy Skills (DIBELS) Assessment: A set of procedures and measures for assessing the acquisition of early literacy skills from kindergarten through sixth grade. It consists of short (1 minute) fluency measures used to regularly monitor the development of early literacy and early reading skills.

Emerald School District: The fictitious name given to a school district that is used in this study. It has approximately 6,000 students and is located in East Multnomah County of Portland, Oregon.

Formative Assessments: Assessments that teachers incorporate into their classroom instruction. They supply ongoing feedback that enables teachers to adjust their instructional practice and gives students timely information which allows them to refine their leaning strategies. Formative assessments are administered often and regularly during the school year and are embedded within the classroom instruction (Popham, 2008; Stiggins, Arter, Chappuis, & Chappuis, 2006).

Interim/Benchmark/Predictive Assessment: A standardized medium scale, medium cycle assessment conducted three to four times a year. It has two functions: first, it evaluates students' knowledge and skills relative to predetermined curricular outcomes on summative standardized tests; second, it informs instructional program decisions at the

classroom, school and district level making it possible to monitor and adjust instruction at regular intervals (Christman et al., 2009; Marshall, 2008; Perie, Marion, & Gong, 2007; Popham, 2008; Stiggins et al., 2006).

Literacy Screening Tool/Probe (also know as a universal screening tool): A quick assessment that can accurately classify students as *at risk* or *low risk* for not meeting expected performance.

Low Risk Student: This student is on-track or predicted to be on-track to meet grade level reading standards.

Monitoring of Reading: The number of miscues self-corrected divided by the number of miscues self-corrected plus miscues not self-corrected.

Negative Predictive Value: An indicator of the proportion of students who were correctly identified as not at risk out of all students identified as not at risk on the screening instrument (Glover & Albers, 2007; Gredler, 1997, 2000; Jenkins, Hudson, & Johnson, 2007; Severson & Walker, 2002; Stage & Jacobsen, 2001; Wood, 2006).

Oral Reading Fluency (ORF) or DIBELS ORF: An assessment designed to measure a student's ability to accurately decode connected text at a predetermined rate of speed.

Oregon Assessment of Knowledge and Skills (OAKS): The state of Oregon's on-line statewide assessment system. As part of its work to improve the OAKS, Oregon has partnered with the American Institute for Research to create an online testing system that assesses students' mastery of Oregon content standards.

Positive Predictive Value: An indicator of the proportion of students who were correctly identified as at risk out of all students who were identified as at risk on the

screening instrument (Glover & Albers, 2007; Gredler, 1997, 2000; Jenkins et al., 2007; Severson & Walker, 2002; Stage & Jacobsen, 2001; Wood, 2006).

Progress Monitoring: A repeated measure of academic performance used to inform instruction of individual students in general education as well as students in special education in grades K-8. It is conducted at least monthly for the following purposes: to estimate rates of improvement, to identify students who are not demonstrating adequate progress and/or to compare the efficacy of different forms of instruction in order to design more effective, individualized instruction.

R-CBM: Reading curriculum-based measure. An R-CBM is a method teachers use to find out how students are progressing in basic reading skills. An R-CBM measures the rate at which a student can accurately decoded connected text.

Response To Intervention (RTI): The integration of assessment and intervention strategies within a multi-level prevention system designed to maximize student achievement and to reduce behavior problems. With RTI, schools identify students at risk for poor learning outcomes, monitor student progress, provide evidence-based interventions, and adjust the intensity and nature of those interventions depending on a student's responsiveness, and also identify students with learning disabilities.

Sensitivity: The ability of a screening tool to identify as at risk students who in fact perform unsatisfactorily on a future criterion measure (i.e., *true positives*) (Jenkins et al., 2007).

Some Risk Student Classification: This student is potentially at risk for not meeting grade level reading standards by the end of the year.

Specificity: The accuracy of a screening tool in identifying as low risk individuals who later perform satisfactorily on a future criterion measure (i.e., *true negatives*) (Jenkins et al., 2007).

Summative Assessments: One-time assessments designed to evaluate each student's performance in relation to a defined set of content standards. They are primarily used as accountability tools for districts, schools, teachers and students (Christman et al., 2009; Perie et al., 2007).

CHAPTER II

REVIEW OF THE LITERATURE

Introduction

An obstacle that Emerald School Districts needed to overcome was the claim that progressive early literacy interim assessments such as DRA lack statistical analysis to support their use, while traditional skills-first assessments such as DIBELS have a considerable body of statistical analysis supporting their use. This chapter reviews the literature that examines these claims by (a) discussing the conflicting philosophical and pedagogical perspectives on early literacy acquisition and the role assessment plays in this dispute, (b) discussing the function and purpose of early literacy interim assessments in a school or district's comprehensive assessment system and looking at how assessment systems promote or inhibit instructional congruency, (c) examining two frequently used interim literacy assessments and the research related to their validity of predicting performance on high stakes standardized tests, and finally, (d) exploring common uses of early literacy interim assessments by districts and schools.

The Reading Wars

As the enduring controversy over the nature and purpose of schooling finds expression in the methodology of literacy, the specifics of the discussion become especially complex (P. D. Pearson, 2004). A range of perspectives inevitably emerges about what knowledge and which skills we want students to acquire and which

methodology is most effective in ensuring that they gain access to bodies of knowledge. Although the challenges are many as educators try to implement a robust, coherent curriculum, developing assessments that align with the knowledge and skills students are expected to acquire as well as the strategies teachers use in their classrooms is particularly difficult.

A classroom based on a behaviorist approach looks very different than one based on a constructivist approach. Phillips and Soltis (1998) described the workings of a behaviorist classroom. It is—

> …a process of expanding the behavioral repertoire, not a matter of expanding the ideas in the learner's mind.…In terms of a classroom example, the behaviorist is interested not how a pupil understands and learns Einstein's theory, but in how a pupil can be led to behave in such a way that he or she can do certain things (such as get the correct answer to problems, perform experiments, write down certain equations when asked by a teacher, and so on. (p. 23)

This emphasis on getting the right answer in the teaching and assessment of reading is apparent in many classrooms on a daily basis. All too often schools and districts measure students' success in reading, not by examining how well they understand and interact with what is being read, but by how proficiently they can repeat the behavioral tasks of accurately decoding connected text at a predefined rate. For a progressive school and district such as Emerald, these types of assessments have limited usefulness.

The constructivists have very different expectations. Langer (1997) used a tennis analogy to explain this approach to learning:

> Learning the basics in a rote, unthinking manner almost ensures mediocrity … Consider tennis. At tennis camp I was taught exactly how to hold my racket and toss the ball when serving. We were all taught the same way. When I later watched the U.S. open, I noticed that none of the top players served the way I was

> taught, and more important, each of them served slightly differently. Most of us are not taught our skills, whether academic, athletic, or artistic, by the real experts. The rules we are given to practice are based on generally accepted truths about how to perform the task and not on our individual abilities. If we mindlessly practice these skills, we are not likely to surpass our teachers. Even if we are fortunate enough to be shown how to do something by a true expert, mindless practice keeps the activity from becoming our own. (p. 14)

A Balanced Literacy theory of teaching reading grows from this perspective. Teachers, who have adopted this approach, do not expect all students to acquire an exact, predefined set of behavioral repertoires to be successful readers; rather, progressives see reading as "thinking guided by print" (Calkins, 2001, p. 13). Interestingly, despite all their differences, early literacy instructional programs based on both Direct Instruction and Balanced Literacy approaches have shown mixed results in terms of student outcomes (Calkins, 2001; Kame'enui, Carnine, Dixon, Simmons, & Coyne, 2002; Kame'enui & Simmons, 1990; NRP, 2000; P. D. Pearson, 2004). Advocates from both persuasions can point to studies that validate their approaches and can mount passionate arguments against studies that call them into question (P. D. Pearson, 2004).

Philosophical and Methodological Congruence

Although apologists on both sides resist this claim, methodology may not be the determining factor in the success of an instructional program. So, if it is not the methodology that determines the success of a particular approach, then what is? Many researchers argue that a critically important characteristic of a successful reading program is that its methodology is congruent with its philosophy and that both the philosophy and methodology are congruent throughout the school or district (Allington, 1990, 2009; Allington & Johnston, 1986; Engelmann, 2008; Lyon, Fletcher, Torgesen, Shaywitz, & Chhabra, 2004; Wilson-Bridgman, 2003). Emerald, for example, had adopted a Balanced

Literacy theory of teaching reading but was without a common assessment tool that could create a shared language around a student's progress. Because classroom and Title I assessments were constructivist in nature while their special education assessments derived from a behaviorist philosophy, they ended-up with a confusing assessment system. They realized that it did not work to have a behaviorist philosophy and constructivist methodology or to have constructivist methodology in general curriculum while having a behaviorist methodology in special education reading programs. Neither did it work to have a mixture of assessment instruments that were not necessarily aligned with each other or with the overall assessment program (Allington, 1990, 2009; Allington & Johnston, 1986; Engelmann, 2008; Lyon et al., 2004; Wilson-Bridgman, 2003). Allington (2009) wrote at length about the effects of this kind of misalignment in early literacy instruction:

> As a result of our findings, we hypothesized that the lack of coordination was one of the reasons few interventions produced very many struggling readers who made accelerated reading progress. Instead, we argued, the very design of the two sets of reading lessons seemed more likely to develop confused readers than thoughtful readers. The dominant design we found seemed more likely to develop struggling readers who obtained partial knowledge of multiple strategies and full and consistent use of none. (p. 91)

In Allington's (Allington & Johnston, 1986) work the core instructional program that was used was a Balanced Literacy methodology, while the interventions were based on Direct Instruction methodology. Even though Allington is a constructivist, his research emphasizes congruence over methodology. Wilson-Bridgeman (2003) also conducted research in the area of curricular congruence and reached similar conclusions. He pointed out that various supplemental programs designed to help lagging readers have

been largely ineffective because of a lack of articulation and lack of congruency between regular and supplemental programs (Wilson-Bridgman, 2003).

Wilson-Bridgman's (2003) research demonstrates the problem that arises when a school's core instructional methodology, including classroom assessment, is not aligned with its intervention or supplemental programs such as Title I and Special Education. Emerald found that without congruent goals, common assessments, and good communication among all its staff—including the classroom teachers, special education learning specialists, and Title I reading specialists—it became extremely difficult, if not impossible, to work effectively with struggling readers. Emerald is certainly not alone in having to confront the confusion that arises from lack of congruency.

Several scholars argue that this persistent, but largely unacknowledged incongruence, is at least partly responsible for the failure of district and school improvement efforts in the area of early literacy (Allington, 2009; Cuban, 2007a, 2007b; Wilson-Bridgman, 2003). Wilson-Bridgman summarized the problem this way:

> Finally, with regard to coordinating efforts to present a common message to both remedial and classroom teachers, a common message should be communicated to both teachers of the at-risk student about assessment for the purpose of informing instruction. The use of common assessments to inform instruction, such as running records, can be tools used to bring about congruent instruction for a student at risk....When both teachers have a clear understanding of how to observe and assess students to determine their levels of development…they are both more likely to respond to the student's developmental needs with the appropriate instruction. Also in terms of assessment, this study indicated that the systematic collection of data to monitor growth and improvement over time and to demonstrate the efficacy of a program can influence changes in teacher practice and provide motivation for continued efforts for improvement in instruction. (p. 66)

Over the last decade research, such as Wilson-Bridgman's (2003), has pushed school reformers to embrace data-driven decision-making as the central strategy in their

efforts to improve schools. The assumption is that using reliable and valid common student assessment data will shine a light on which competencies students are acquiring and which they are not. Given this information, educators can make the path to improvement more clear and specific. It is commonly believed that this clarity will translate into more focused decisions about use of resources as well as decisions about which core instructional methodology and intervention or supplemental programs are the most effective (Christman et al., 2009). Common assessments that are administered at various points in the school year are essential to an ongoing, data-driven approach to instructional improvement. These tools are referred to as interim benchmark or predictive assessments.

Interim Assessments

Interim benchmark or predictive assessments link formative classroom assessments to a district or statewide summative assessment. In order to define more clearly what an interim assessment is and its role within a comprehensive assessment system, it is necessary first briefly to define each of these assessments:

Summative Assessments: Summative assessments (also known as end of year assessments, exit exams or accountability tests) are usually one-time assessments designed to evaluate each student's performance in relation to a defined set of content standards. They are primarily used as accountability tools for districts, schools, teachers and students (Christman et al., 2009; Perie et al., 2007).

Formative Assessments: Formative assessments are assessments that teachers incorporate into their classroom instruction. They supply ongoing feedback that enables teachers to adjust their instructional practice and gives students timely information that

allows them to refine their leaning strategies. Formative assessments are administered often and regularly during the school year and are imbedded within the instruction of the classroom (Popham, 2008; Stiggins et al., 2006).

Educators are drawn to the use of formative assessments as a vehicle both to improve student learning and to improve students' scores on critical accountability tests. However, the latter goal of using formative assessment to improve student test scores on assessments is controversial. Many assessment experts contend that the use of formative assessments will not improve scores on most accountability tests or summative assessments (Blume, 2009; Clune & White, 2008; Popham, 2008; Stiggins et al., 2006). They believe that the formative assessment process can have significant impact on student performance only if the tests are instructionally sensitive, and summative assessments are not instructionally sensitive (Christman et al., 2009; Perie et al., 2007; Popham, 2008; Sirotnik & Kimball, 1999; Stiggins et al., 2006).

Interim Assessment: The Interim Assessment (also known as benchmark or predictive) within a comprehensive system provides a bridge between formative and summative assessments. An interim assessment is a standardized medium scale, medium cycle assessment conducted three to four times a year. It has two functions: first, it can evaluate students' knowledge and skills relative to predetermined curricular outcomes on summative standardized tests; second, it can inform instructional program decisions at the classroom, school and district level making it possible to monitor and adjust instruction at regular intervals. Given these functions, the interim assessment has instructional as well as evaluative or predictive utility (Christman et al., 2009; Marshall, 2008; Perie et al., 2007; Popham, 2008; Stiggins et al., 2006). Because a valid and reliable interim

assessment has the capacity to evaluate instructional practice, it can be used as part of a formative assessment process to give feedback to both teachers and students. It also can be used as one part of an instructional program evaluation system that attempts to empirically validate effective practices that are enabling students to have access to curriculum. Finally, since a valid and reliable interim assessment can evaluate student knowledge and skills relative to predetermined curricular outcomes, it can predict success on a summative assessment that is measuring student performance on the same curriculum.

The interim assessment is not only a valuable tool within a comprehensive assessment program; it is a critical tool because it looks in two directions linking both to classroom instruction and to high-stakes summative assessments. It is important to note, however, that, within this potentially powerful system, formative assessments must be very carefully constructed so that they provide ongoing data that are well-aligned with the information that the interim assessment is designed to gather and that the interim assessment has also been calibrated to ensure that it is well-aligned with the summative assessment.

In a comprehensive system, the three types of assessment are closely related. The interim assessment is the bridge between the formative and the summative assessments. A strong assessment system uses ongoing classroom based formative assessments that are conducted regularly (at least once every 2 weeks) to inform and evaluate the instructional practice in the classroom and track the progress of student learning. The interim assessment, which takes more time and is standardized, not only evaluates instructional practice and student learning but also gives feedback to classroom teachers about the

validity and reliability of their formative classroom assessments. And finally, the summative assessment measures student achievement in relation to grade level standards and also evaluates the reliability and validity of the interim assessment. The summative assessment should be conducted only once a year because of its development costs and the amount of time needed to administer it. The interim assessment is a critical component in connecting classroom practice to high stakes accountability tests.

Instruction and Assessment Incongruence

Little debate exists within the research and literature concerning the three essential components of a robust comprehensive assessment system. Why, then, do districts, such as Emerald, find it a challenge to develop such a system specifically for literacy? The challenge does not lie with summative assessments because they are usually predetermined and supported at the state level as is the case in Oregon. Implementation is not only straightforward but also mandated. Nor does it lie with formative assessments. With high quality training and professional development, effective formative assessment processes can be implemented within a classroom and throughout a school (Popham, 2008; Stiggins et al., 2006). The interim assessment is where the challenge lies, and the challenge is far greater for the Balanced Literacy model than for the Direct Instruction model (Clune & White, 2008; Perie et al., 2007).

Emerald has struggled to improve student performance in the area of early literacy. As they analyzed their instructional program they uncovered a lack of communication, common goals and instructional alignment between the core reading program and the supplemental instructional programs such as Special Education services and Title I reading programs. Based on the research that they were reading, the Emerald

School District decided that having a common reading assessment would be an important step in achieving instructional congruency.

The problem that they faced was that the available common literacy assessment that had been proven to be aligned with the OAKS was DIBELS. Using this skills-based assessment would not resolve the instructional congruency issue that we were facing. On the other hand, Emerald had been using the DRA albeit very inconsistently and without any staff training. With better staff development Emerald believed that they could vastly improve its implementation and were convinced that it was aligned with their philosophy and curriculum. Unfortunately, Emerald could find no studies on whether it predicted success on the OAKS assessment and/or what benchmark scores they could use in identifying which students were on track for reading success and which were not.

The following section of this chapter examines the interim literacy assessment choices available to Emerald.

Two Types of Interim Assessments

Two types of assessment are commonly used in Oregon: the first is the DIBELS; the second is the DRA. This literature review will briefly describe these two interim assessments, define their underlying pedagogical premise, and then analyze the limitations of each in order to establish the necessity of this study.

DIBELS

Curriculum based measurements assess the acquisition of early literacy skills from kindergarten through sixth grade. Subtests of these assessments, such as a 1-minute ORF test, are regularly used to monitor the development of early literacy and early reading skills. School districts have used the ORF to assess students' early literacy

development, predict whether a student might or might not be on track to pass a summative reading/literature assessment, and to determine student progress (Buck & Torgesen, 2002; Madelaine & Wheldall, 2005; Shaw & Shaw, 2002; Wilson, 2005). These studies focus on the predictive utility of ORF (i.e., the rate and accuracy of decoding connected text), which is a subtest of a DIBELS interim assessment, toward passing a summative assessment. Based on the review of the literature concerning risk screening and interim assessment, DIBELS is the most used and studied tool. The subtest within a DIBELS that is used to predict future performance on a summative assessment is the ORF assessment. The DIBELS ORF (also referred to as a CBM or R-CBM) assesses a student's ability to accurately decode connected text at a predetermined rate of speed.

Nationwide DIBELS ORF has been determined to have strong correlation to performing well on a third grade summative assessment. In Oregon, Good, Simmons, and Kame'enui (2001) conducted a series of linked, short-term, longitudinal studies. These studies examined the utility and predictive validity of the ORF subtest in DIBELS measures from kindergarten through third grade with the Oregon Assessment-Reading/Literature as a high-stakes reading outcome. The authors concluded that—

> The results of this study support accuracy and fluency with connected text as an important foundation for reading competence. Students who read grade-level material at a rate of 110 words correct per minute or better were likely to meet or exceed expectations on the OSA. Students who were able to read less than 70 words correct per minute on grade-level material were not likely to meet expectations on the OSA. (p. 283)

Studies in at least four other states, including Florida, Colorado, Arizona, and Michigan, supported the study (Good et al., 2001).

Florida. Buck and Torgesen (2002) attempted to determine whether measures of ORF are valid and reliable predictors of important reading outcomes and performance on high stakes tests such as the Florida Comprehensive Assessment Test. Specifically, the researchers wanted to determine whether performance on brief, 1-minute measures of ORF are predictive of achievement in reading as measured by the reading portion of the Florida Comprehensive Assessment Test-Sunshine State Standards (FCAT-SSS). If their hypothesis was correct, these 1-minute assessments could provide early warning signs that students may, or may not, succeed on the FCAT-SSS. They conclude the following:

> This initial study demonstrates that, for a large heterogeneous group of third graders, performance on brief oral reading fluency measures can quite accurately predict whether or not a given students will attain a score at level 3 or above on the FCAT reading test. (p. 9)

Colorado. Shaw and Shaw (2002) examined the utility of the DIBELS ORF assessment in predicting the performance level on the third grade (English) reading CSAP, the standards-based reading comprehension assessment that is administered each year. The researchers arrived at these conclusions:

> For this group of third-grade students, 39 of 43 (91%) of the students who scored 90 or above on the DIBELS ORF in the spring scored proficient or advanced on the CSAP, and 11 of 15 (73%) of the students who scored below 90 on the DIBELS ORF scored unsatisfactory or partially proficient. …If future scores for additional students bear out this high percentage of correct CSAP score placements, the DIBELS' utility to predict proficient/advanced and unsatisfactory/partially proficient on the CSAP will be excellent. (p. 9)

Arizona. Wilson (2005) conducted a study for the Tempe School District to determine whether third grade students who reach a benchmark level of ORF are likely to meet the standard on the Arizona Instrument to Measure Standards (AIMS) Reading test

and, conversely, whether students with poorly developed reading fluency are unlikely to meet the standard. Wilson concluded the following:

> ORF can identify those students who are likely to meet the proficiency standard on AIMS with good accuracy (those in low risk category). Further, ORF can identify those who are quite unlikely to reach proficiency (those in the at risk category). Accuracy is somewhat better for identifying students who are not on track to meet the AIMS standard. (p. 4)

Ohio. Vander Meer, Lentz, and Stollar (2005) examined the end of third grade and beginning and end of fourth grade ORF goals set forth by Good and Kaminski (2002) in comparison to Ohio expectations for fourth grade reading proficiency and explored the correlations between ORF and the reading portion of the Ohio Proficiency Test (OPT). Again, this, like the other studies cited, sought to confirm the connection between achieving the DIBELS benchmark fluency goals and passing Ohio's Fourth Grade Reading Proficiency Test. Vander Meer et al. concluded the following:

> The relationship between DIBELS/CBM and the Ohio Fourth Grade Reading Proficiency Test (OPT) were examined in two ways, correlations and the adequacy of DIBELS/CBM criteria as year-end goals or indicators of need for reading intervention. In general, with this sample DIBELS/CBM performance has an adequate relationship with a standardized test of reading, and benchmark goals and "at-risk" criteria would appear valid for setting goals and deciding which students need interventions. (p. 12)

Michigan. Schilling, Carlisle, Scott, and Zeng (2007) looked at the predictive validity of DIBELS. Their study gathered data from first through third graders who made up the first Reading First cohort in Michigan. The authors of the study found that—

> ...DIBELS subtests significantly predicted year-end reading achievement on the ITBS, Reading Total subtest. They also showed that DIBELS at-risk benchmarks for oral reading fluency (ORF) were reasonably accurate at identifying second

and third graders who were reading below the twenty-fifth percentile at the end of the year (80% and 76% for second and third graders, respectively). (p. 429)

Based on the findings of these studies, DIBELS ORF appears to be a useful tool to help educators determine the effectiveness of their instructional program as well as identify students who need support to reach benchmark. Table 1 displays a summary of the correlation between DIBELS ORF and performance state high stakes accountability assessments.

Table 1

Correlations between DIBELS ORF and State High Stakes Tests

Correlation Coefficient Grade	*ORE ORF 2001*	*FLA ORF 2002*	*COL ORF 2002*	*AZ ORF 2005*	*OHIO ORF 2005*	*MINN ORF 2005*
Window	Spring	Spring	Fall/Sp	Spring	Sp/Fall	Spring
Third	.67	.70	.70	.74		.68
Fourth			.67		.61	
Fifth			.75			.65

(Buck & Torgesen, 2002; Good et al., 2001; Shaw & Shaw, 2002; Stage & Jacobsen, 2001; Vander Meer et al., 2005; Wilson, 2005; Wood, 2006)

Clearly the relationship between the DIBELS ORF and state wide assessments is strong throughout the country and has been replicated many times. Table 2 represents the predictive analysis of the DIBELS benchmarks and performance on high stakes accountability assessments. It shows the percentage of students who were classified as at risk of not reaching benchmark based on their DIBELS ORF score who did not also reach benchmark. Conversely it also shows the percentage of students who were classified as at low risk of not reaching benchmark that actually did meet the state grade level standard. Table 2 also shows the overall percentage of students that were accurately assessed using

the DIBELS ORF screening tool. Finally Table 2 shows the percentage of students that fell in between at risk and low risk. These students are defined as Some Risk.

Table 2

Predictive Analysis of DIBELS

State & Year	*Grade & ORF Window*	At Risk & Did Not Meet Benchmark	Low Risk & Did Meet Benchmark	Accurate Prediction
Oregon 2001	3rd Spring	72%	96%	88%
Florida 2002	3rd Spring	81%	91%	88%
Colorado 2002	3rd Spring	73%	91%	86%
Arizona 2005	3rd Spring	82%	91%	88%
Ohio 2005	4th Fall	74%	89%	87%
	4th Spring	46%	90%	83%

(Buck & Torgesen, 2002; Good et al., 2001; Shaw & Shaw, 2002; Stage & Jacobsen, 2001; Vander Meer et al., 2005; Wilson, 2005; Wood, 2006)

Even though DIBELS ORF has a specific utility when it comes to predicting which second and third grade students will pass third grade summative assessments, it is less reliable in predicting success or failure on summative assessments beyond grade 3 (Jenkins et al., 2007; Silberglitt et al., 2006; Wood, 2006). Silberglitt et al. (2006) conducted a study in which they concluded that—

> The magnitude of the relationship between R-CBM and the state accountability test declined significantly with advancing grade levels. Correlations dropped from strong to only moderate at increasing grade levels, which is not surprising given the previous research on factors influencing growth rates for reading fluency. (p. 533)

Silberglitt et al. further concluded that—

> This diminishing relationship could have implications for educators. Using R-CBM to predict state accountability test performance is an increasingly popular practice in schools, especially given the current climate of NCLB. Based on past

> research demonstrating the strong relationship between R-CBM and state tests in the early grades, educators may be assuming that this relationship will remain strong in the later grades as well. The current study suggests the possibility that such an extrapolation is erroneous, and that R-CBM may not have as great a value for the purpose of predicting achievement test scores in the later grades. (p. 533)

The final problem is that—even if the DIBELS ORF assessment has strong predictive utility—many progressive educators believe that it runs contrary to their fundamental beliefs about teaching and learning. These differences with DIBELS ORF include everything from what we assess to how we assess students reading. With regards to how we assess, many progressive educators, believe that assessment is an interactive process between the teacher and learner. The core of the DRA is a one-on-one interactive conference between the student and the teacher. Using the DRA Four-Step Plan, teachers can assess a student's independent reading level, identify strengths and needs, and recommend individual instructional strategies. These steps include Reading Engagement, Oral Reading, Comprehension, and Teacher Analysis. As students develop their reading and comprehension skills, the assessment increases in difficulty to reflect this progress. The DRA assessment is an interactive and adaptive process of teaching students reading ability.

To insure reliability of the assessment tool, the DIBELS ORF assessment attempts to minimize and control teacher judgment and interaction with the student. The administration of the DIBELS ORF is a highly scripted process and has standardized the level of connected text based on grade level at which each student must read for the assessment.

In terms of what we assess, many progressive educators also have concerns about what DIBELS ORF measures. Educators such as R. L. Allington (personal

communication, January 2010) takes issue with DIBEL ORF and its emphasis on phonics-based skills which, from his perspective. gives it little instructional utility, especially within a Balanced Literacy framework. For example, DIBELS ORF assesses the student's ability to read connected text at predetermined rate. By contrast, when progressive educators assess reading fluency, they are interested in both how well the reader can decode text and whether the reader is reading with meaning. They take into consideration whether the student is using correct phrasing, intonation, and expression (R. L. Allington, personal communication, January 2010; Manzo, 2005; Shelton, Altwerger, & Jordan, 2009). The difficulty for Emerald teachers, and others like them, is that they fear that the DIBELS assessment's narrow premise about which behaviors predict reading success causes it to over analyze some skills and under examine other very important ones (R. L. Allington, personal communication, January 2010; Manzo, 2005; Shelton et al., 2009).

The Emerald's Leadership Team's belief that DIBELS emphasizes phonics and oral reading fluency and under emphasizes comprehension skills raised significant concerns. One concern was that the teachers, reading specialists and administrators would invariably feel pressure to improve performance on the DIBELS assessments which would result in staff measuring students' reading ability by how fast and accurately they could decode connected text rather than by how well they understood what they had read. Hammer (2003) studied effects of ORF on reading comprehension and concluded that "while the repeated reading of text does lead to an increase in student's accuracy, there were too many variables that influenced the results to show a significant correlation between oral reading fluency and comprehension" (p. 4). It seemed clear that it would be

unwise to give comprehension short shrift if the intended purpose of a literacy program was to develop lifetime readers (Hammer, 2003; Manzo, 2005; Shelton et al., 2009). Another concern the Leadership Team had was that, even though DIBELS might be a good screening tool, it would not point the way to appropriate instructional strategies. For teachers to know what to do about their students' reading difficulties, they would still need to administer the DRA assessment so that they could decide on the focus of instruction for students who were deemed to be at risk of not making grade level standard.

Hosp and Fuchs (2005) raised both of these concerns in their study examining whether the relation between CBMs and specific reading skills changed as a function of grade. They concluded the following:

> Although screening, diagnostic, and progress monitoring assessments all provide useful information to help students achieve success in reading, they differ in the amount of information, cost, and time it takes to obtain the information. Screening assessments are meant to be short and efficient. However, this format makes them inappropriate for diagnostic assessment because they do not assess all skills and/or any one particular skill in depth. And because screening assessments lack depth and breadth, they do not provide educators with meaningful information that can be linked directly to instructional strategies. Diagnostic assessments, on the other hand, are designed to identify which skills the student has or has not mastered so that an appropriate course of instruction can be determined. (p. 9)

Clearly DIBELS assessments have significant limitations for Emerald and other progressive schools and districts because they do not effectively assess critical skills of comprehension, and they do not have instructional utility. Finally, they contribute to a problem that educators and the public alike find troublesome, if not alarming: we spend too much time assessing and not enough time teaching our students. Emerald teachers were strongly against using DIBELS assessments because it only informed them of what

they already believed they knew about their students—who was at risk of not making grade level standards. DIBELS would not inform them why a student was at risk nor how they should their focus their instruction. The Emerald staff could build a case for not using DIBELS, but could they build an equally strong case for using the DRA?

DRA

The DRA (Pearson Education, 2009) is another interim/benchmark assessment that is widely used. It differs from DIBELS in that its emphasis is in comprehension and has the diagnostic capability of identifying both the skills students have mastered and those which they have not. It is a set of individually administered, criterion-referenced reading assessments that is designed to help teachers determine each student's independent reading level, identify instructional needs, and monitor progress. It is made up of Benchmark Assessment Books Leveled A-80 and measures reading behaviors including reading engagement, fluency and comprehension. School districts use the DRA to assess students' early literacy development and determine student progress.

Unlike the DIBELS assessment, the DRA is an adaptive tool that assesses students at their independent reading level rather than at a benchmark level based on the grade of the student. Districts that adopt the DRA district-wide generally favor a Balanced Literacy approach, and this interim assessment is used primarily to inform teachers about the instructional needs of each student. Even though the DRA appears to be a very useful in informing a teacher's instructional decisions, questions and concerns persist about whether it can be equally valuable as an evaluative and predictive tool.

The hesitation schools or districts such as Emerald have about relying on DRA and foregoing DIBELS is that very little, if any, empirical research has evaluated the

predictive validity of the DRA or any of its subtests on passing an end of year summative assessment such as OAKS. Because of this lack of research, some districts in Oregon administer both DIBELS and DRA interim assessments to students in order to fulfill the assessment needs of all stakeholders within a comprehensive system. This practice raised obvious concerns for the Emerald's Leadership Team about how much time would be used in testing at the expense of instructional time.

Another concern about using a DRA as an evaluative and predictive tool is that teacher judgment determines the level of the reading passage and which passage a student would be asked to read during the assessment. Giving teachers this much latitude was especially problematic in Emerald because the staff had received little or no training in how to administer the DRA. Emerald was not alone in dealing with these concerns. For example, Madelaine and Wheldall (2005) contended that "…over-reliance on teacher judgment for selecting low-progress readers for appropriate instruction, or for instructional decision-making, may be misplaced and that it may be preferable to employ a more objective, quick alternative based on CBM" (p. 33). Forbes, Poparad, and McBride (2004) have feared that variance in content or vocabulary used in a reading assessment passage could produce variance in the individual test results. Given these concerns as well as well as the lack of specific studies to empirically validate the utility of the DRA subtest on its predictive and evaluation capabilities, it became necessary to investigate further the utility of the DRA within the assessment system that a school district might intend to implement.

A third concern related to the DRA was the length of time required to complete it. One of the advantages of a DIBELS ORF is that it can be administered within just a

couple of minutes per student whereas the full DRA takes 5-10 minutes (Pearson Education, 2009). The DRA can help determine the focus of instruction; however, if a district is also going to use it as an interim screening tool, the time required to administer it to all students will take away from instructional time.

These concerns became very troubling for the Emerald Leadership Team. They recognized the importance of having a robust interim assessment that was capable of screening for students at risk and capable of predicting future success on summative assessments. They already knew that they did not like DIBELS ORF because it was not a good match for their Balanced Literacy curriculum, but they also had a huge unknown about the DRA. They needed to understand Interim Assessments and their uses before they could decide about the utility of the DRA, and they were especially interested in learning whether selected DRA subtests could be used just as effectively the subtest ORF functioned for DIBELS.

Uses of Interim Assessments

Because interim reading assessments play an increasingly important role in districts' efforts to ensure that students are making progress toward meeting benchmarks, it is important to know that they are effective and appropriate. One of their most widespread uses is to predict future performance on high-stakes accountability tests. By using specific subtests of these assessments, probes can be quickly administered to all students several times a year. These interim assessments are frequently included in the Response to Intervention (RTI) framework, a model that is currently used widely. RTI requires common assessments and screening tools throughout a school and/or district so that educators can ensure instructional congruency between classroom instruction and

interventions (Allington, 2009). Emerald was beginning to implement the RTI framework, and having their interim assessment be able to screen for students at risk of reading difficulties was essential.

In Oregon DIBELS is the most popular interim assessment. Two of its subtests—accuracy (the ability of a student to accurately decode connected text) and rate (the speed at which a student reads aloud connected text)—function as a screening tool. This tool, commonly referred to as ORF, is useful because it accurately predicts later reading success (Buck & Torgesen, 2002; Good et al., 2001; Hammer, 2003; Hosp & Fuchs, 2005; Shaw & Shaw, 2002; Silberglitt et al., 2006; Wood, 2006). Teachers use the results of assessments to determine which students need targeted interventions designed to prevent future reading failure.

Not only are brief probes, such as the ORF, efficient and effective in identifying children who need additional support, they also provide a means for monitoring progress of groups of students (e.g., a class or grade) to ensure that students who are at risk of academic failure receive appropriate reading interventions. Because it is a universal screening tool—meaning that it is used on all students, it must be easy and simple to administer. The ORF probe meets that criterion: it is a 1-minute timed reading of connected text designed to measure accuracy and rate of the student's reading. A teacher can screen an entire classroom using this subtest in a relatively short time.

The Emerald School District, however, much preferred to use two DRA subtests as their screening tool. The difference between the DRA subtests and ORF represents the difference in the underlying philosophy of the educators. In a traditional paradigm, students need to establish solid skills before comprehension can be facilitated, while in a

progressive paradigm students need to be engaged in the text before skills can be taught. It comes down to a question of skills first or meaning first. The problem that the Emerald Leadership Team and teachers faced, however is that the DRA is not backed by research demonstrating its efficacy in predicting outcomes on summative performance measures.

Screening measures can be evaluated on several dimensions, including classification accuracy (i.e., predictive utility), efficiency, and consequential validity (Glover & Albers, 2007; Jenkins et al., 2007). Two types of examinations relate to a screening measure's validity: criterion validity and classification accuracy or predictive utility (Glover & Albers, 2007; Jenkins et al., 2007). Studies of criterion validity examine correlations between performance on a screening measure and an established measure of reading (e.g., statewide reading exams), the latter administered either concurrently or at a future time (Buck & Torgesen, 2002; Good et al., 2001; Hammer, 2003; Hosp & Fuchs, 2005; Shaw & Shaw, 2002; Silberglitt et al., 2006; Wood, 2006). The strength of the correlation between the screening and criterion measure provides evidence of the screening measure's criterion validity.

The central function of a screening measure is its ability to accurately classify students (i.e., predictive utility) as at risk or low risk for not meeting expected performance. Simply documenting a relationship between a screening measure and a student's later reading ability does not establish its utility or its ability to predict concurrent and future success on established measures. If a screening tool is going to be subjected to a comprehensive analysis, it will be necessary to include an evaluation of the tool's accuracy and sensitivity in classifying low, middle and high performing students in relation to reading tests such as OAKS.

Whereas criterion validity analysis is helpful in identifying whether the measure holds potential as a screen, classification analysis is essential to assure its utility in predicting concurrent and future success on high-stakes accountability assessments. A screening tool that can do this will be extremely useful in determining which students are on track and which students need a full interim assessment to determine what instruction is necessary to prevent academic failure.

A screening measure's predictive utility is evaluated by its accuracy and degree of *sensitivity* and *specificity*. Jenkins et al. (2007) described these two measures:

> Sensitivity refers to the accuracy of a screening mechanism in identifying as "at risk" individuals who in fact perform unsatisfactorily on a future criterion measure (i.e. "true positives"). Specificity refers to the accuracy of a screening mechanism in identifying as "not at risk" individuals who later perform satisfactorily on a future criterion measure (i.e. "true negatives"). (p. 583)

The ideal screening tool would distinguish 100% of the time between students who are not going to perform adequately on a reading assessment such as OAKS from those who will; however, that is an unreal expectation for a 1-2 minute screening tool. In practice, developing a good screening tool comes down to weighing the tradeoffs between sensitivity and specificity. Sensitivity increases as one type of prediction error (false negatives) decreases. False negatives refer to students not identified as at risk on a screening measure who later perform poorly on the criterion measure. In an RTI system this omission would result in students not receiving targeted or supplemental instruction that would have likely given them the skills they needed to pass statewide reading assessment.

In contrast, specificity increases as another type of prediction error (false positives) decreases. False positives refer to students identified as at risk who later

perform satisfactorily on the statewide reading assessment. A screening tool that produces too many false positives wastes resources and unnecessarily labels students as at risk. In an RTI system the main goal is to find as many students as possible who are at risk of having academic problems as early as possible so that interventions can start as quickly as possible. Because of this sense of urgency, educators who are implementing an RTI model need a screening tool that is accurate, has a high degree of sensitivity, and is pedagogically aligned with the instruction and materials used in the school or district.

Finally, it is very important to note that these screening tools or probes are not full diagnostic assessments. A 1-minute timed reading is neither sufficient nor able to tell a teacher why a student is at risk of not reaching grade level standards. It is also important to note that it cannot be assumed that because there is a strong correlation between accuracy of reading connected text at an arbitrary rate of speed and reaching benchmark standards, there is also a causal relationship between oral reading fluency and comprehension. Not only is this a dubious assumption, it fails to recognize comprehension as a fundamentally important reading skill and one which the OAKS reading assessment measures.

Even though the 1-minute probe may not be particularly useful as a diagnostic tool, it is useful in identifying students who may be at risk of failure. It can alert teachers to conduct a full interim assessment using all its subtests to determine which specific skills a student is lacking. A full DIBELS continues the skills-first emphasis that is present in the ORF subtest. However, Balanced Literacy educators, who work from a meaning-first perspective, have significant concerns about using interim assessments such as DIBELS.

Here is the problem for the Emerald Leadership Team as well as other Balanced Literacy educators: many schools and school districts currently mandate the ORF, and, although Balanced Literacy educators have reasons to question its usefulness, they comply with the mandate because they recognize its predictive value. Although, the screening tool can indicate that a student is not on track to reach benchmark, it cannot indicate what is most important—why the student not. If teachers use ORF as their screening tool, they will need to reassess all of their at risk students with an additional interim assessment which is compatible with their beliefs about how children learn to read. They will then be in a position to diagnose their students' needs and to adjust their instruction accordingly. This is an unhappy predicament for Balanced Literacy educators. They feel as if they are being forced to narrow their instruction, that they subjecting their students to more testing at the expense of time teaching, and that they have had to compromise what they believe are the best instructional practices. For them, a welcome solution would be to have the confidence that an alternate interim assessment such as the DRA has as much predictive value as DIBELS. They then would be able to truly link their instruction to high stakes accountability.

Conclusion

The fundamental differences between the traditionalist perspective and the progressive perspective plays out in the persistent debate about assessment. It actually may not be necessary or helpful to come to a resolution about whether DIBELS or the DRA is the more effective instrument. What was necessary for the Emerald Leadership Team was first to recognize the fundamental differences in philosophies and pedagogies that underlie interim/benchmark assessments. Second, they needed to recognize how

important curricular congruence is. Finally, they needed to realize how critical the role of a comprehensive assessment system is to improving student achievement, especially for the students who are at the greatest risk.

Traditional educators have a useful tool in the DIBELS. It can assess what they believe to be the essential skills of early literacy development and achieve curricular congruence. The DRA has an equally strong instructional utility as the DIBELS, may have equally strong evaluative and predictive utility, and is also congruent with Balanced Literacy. The question then becomes—is it possible to fulfill the assessment needs of all stakeholders by using one interim assessment that has instructional, evaluative and predictive utility and that is also congruent with a Balanced Literacy framework? The study described in the following chapter examines the relationship and predictive value of two subtests of the DRA to student performance on the OAKS. If a strong, positive relationship exists, progressive educators around the state can have confidence in its usefulness as an interim assessment tool. In an attempt to answer this question this investigation explores the correlation and predictive value of a Balanced Literacy interim assessment such as the DRA on the OAKS in order to establish its utility as a viable alternative to DIBELS.

CHAPTER III

METHODOLOGY

This chapter describes the purpose of the study, the need for and significance of the study, the research questions, the context in which the research will take place, the participants, and data sources. This chapter also describes the study's overall design and proposed analyses in relation to the research questions that the Emerald Leadership Team needed to answer in order to continue their work in developing a comprehensive assessment system. Finally, the limitations of the study are described including any potential researcher bias that might be present.

Purpose, Need, and Significance of Study

The purpose of this study is to examine whether or not two subtests of the DRA interim assessment can be used effectively as a screening tool to predict if students are on track to pass the OAKS reading test. If it proves to be the case that the DRA subtests can be used as an alternative to the DIBELS' ORF subtest, Balanced Literacy school districts, such as Emerald, will be able to fully develop a comprehensive early literacy assessment system within their districts that is congruent with their core and supplementary instructional programs. If a philosophically and pedagogically congruent literacy screening tool can be developed and implemented Balanced Literacy school districts, such as Emerald, will then be able to—

- better integrate assessment within their instructional program,
- use these assessments to inform their instructional decisions,
- increase instructional time during a school year because students will not be subjected to duplicate assessments, and
- link their instruction to high stakes accountability test.

Research Questions

This study addresses three research questions:

1. Is there a relationship between the DRA text level at which a student can successfully decode and monitor his own reading and his RIT (Rausch Unit) score on the OAKS Reading Assessment?
2. Are elementary students in grades 3-6 who reach a benchmark level of accuracy when decoding text and who monitor and correct their own decoding errors in the beginning of a school year likely to meet the standard on the OAKS Reading/ Literature assessment by the end of that same year? And conversely, are students with poorly developed decoding skills who do not monitor and correct their own decoding errors in the beginning of a school year unlikely to meet the standard by the end of that same year?
3. Does the DRA level of text at which a student can successfully decode and monitor his own reading add to the predictability of his performance on the Oregon statewide reading test over and above his previous year's performance on the statewide test?

Context for the Research

This research involves six of the seven elementary schools in Emerald School District. Emerald is a middle-class school district in the east county of Portland, Oregon. The district includes seven K-6 elementary schools, one 7-8 middle school, one 9-12 comprehensive high school, and one 7-12 alternative school.

Emerald's enrollment has grown by 1-2% (an average of 75-100 students) annually for more than a decade, while the English Language Learner (ELL) population has grown from 105 students in 1993 to 1,203 students reaching approximately 18% of total district enrollment at the time the data for use in this study were collected. Individual school percentages ranged from 1% to 36% with 45 different languages spoken, the top five being Spanish, Russian/Ukrainian, Romanian, Hmong and Vietnamese. In 2008-2009 the special education population totaled 13%. Seven percent of the students were identified as Talented and Gifted. Forty-seven percent of students district-wide qualified for free/reduced lunch, and the percentages in various schools ranged from 32 to 72.

As part of its efforts to increase academic achievement in the area of literacy, Emerald decided to implement the RTI framework beginning in the 2008-2009 school year. An interim reading assessment is a necessary component in an RTI system. On the face of it, DIBELS would have been a logical choice based on its popularity among many Oregon educators. Although, teachers, building and district administrators wanted to implement an interim reading assessment, ORF was not an acceptable option because it was incongruent with their philosophy, instructional methodology and materials. For many years Emerald has been using the DRA as an interim reading assessment both as a

tool to inform instruction and to report reading progress to parents on report cards in grades 1-3. A similar type of assessment was in place for grades 4-6. Grade level expectations for the DRA reading level at which a student should be able to read independently has been established for grades 1-6. These expectations were based on the DRA materials recommendation (Pearson Education, 2009).

Although Emerald has been using the full DRA to inform teacher instruction and for reporting to parents, the data were incomplete. Because it took so much time to assess each student individually, not all of the students were being assessed. Since the data were incomplete, the DRA could not be used to screen for students who were potentially at risk of not meeting grade level reading expectations. During the 2008-2009 school year the district decided to implement the DRA assessment consistently and to use two subtests (Accuracy and Monitoring of Reading) as a universal screening tool. These two short subtests analyze a student's ability to accurately decode connected text and to monitor and correct decoding errors. They take an average of approximately two minutes to administer to an individual student. Emerald used their results as indictors of whether a student was on target to reach grade level standards based on OAKS. They were chosen in lieu of ORF for several reasons:

1. They can be administered in a very short period of time (less than two minutes) and are comparable in the length of time to administer to the ORF subtest.
2. Monitoring of Reading examines the degree to which a student is engaged in the reading material (Clay, 2000; Forbes et al., 2004). This subtest aligns philosophically with a progressive literacy philosophy which views reading

development as a process of students' acquiring literacy skills through the engagement of high interest reading material that is at the just right reading level.

3. Monitoring was chosen as the probe because Emerald has a relatively high ELL population (more than 20% at the time of this writing). Educators were concerned that, even though the ELL students could accurately decode, they were not understanding the text. Teachers felt that the Monitoring tool would be a better screening tool than ORF for determining which students were at risk readers. Because it analyzes engagement, it was more likely to catch students who could decode but were not understanding what they were reading.
4. In an inter-item correlational analysis of the DRA subtest, both Monitoring and Accuracy are moderately to highly correlated to the comprehension subtests. Both Monitoring and Accuracy have a stronger inter-item correlation to the Comprehension subtests than does the Rate subtest (Pearson Education, 2009).

Participants

The assessment results for 958 students from 2008-2009 academic year are used in this study: 275 third graders, 247 fourth graders, 268 fifth graders, and 168 sixth graders. This study is limited to two groups of these students: those classified as at risk and those classified as low risk for future reading failure. These groups were chosen because the purpose of screening in a RTI framework is to identify which students need more intensive instruction and which students do not.

The students were enrolled in five Schoolwide Title I elementary schools and one Title I targeted assisted elementary school. All 2,016 available students, grades 3-5, were tested in these schools. Because of the Emerald's mobility rate and because this was the

first year of implementation of the district-wide interim assessment, some students either did not take the OAKS assessment in the school district or did not have valid Fall and Spring DRA screening assessments. Of the 2,016 students assessed, the 958 participants represent the number of students who took the OAKS assessment in the Emerald school district and had valid Fall and Spring DRA screening assessments. Table 3 displays the demographic make up of the participants.

Table 3

Demographic Profile of Participants

Grade		*% & #*	*ELL*	*SpEd*	*NA*[a]	*Asian*	*Black*	*Hisp*[b]	*White*
3rd	**%**	29%	29%	12%	1%	10%	8%	26%	55%
	#	275	80	35	2	28	21	71	153
4th	**%**	26%	23%	13%	3%	9%	8%	21%	60%
	#	247	57	33	7	23	19	51	147
5th	**%**	28%	19%	12%	1%	10%	7%	25%	57%
	#	268	52	31	2	26	20	66	154
6th	**%**	18%	20%	15%	4%	14%	4%	18%	60%
	#	168	34	26	7	24	7	30	100
Total	**%**	100%	23%	13%	2%	11%	7%	23%	58%
	#	958	223	125	18	101	67	218	554

[a] Native American
[b] Hispanic

All of these students spent most of their educational day in their regular classrooms. Licensed classroom teachers administered the DRA screening test to students individually during their regular literacy block time. The classroom teachers also administered the OAKS either in their classrooms or in the school's computer lab.

Data Sources and Collection

Measures

Two measures were used: The first was the OAKS. The second was the Monitoring and Accuracy subtests from the DRA.

OAKS. The OAKS reading test is administered annually to Oregon's third, fourth, fifth and sixth graders. Students took the OAKS assessment two-to-three times between January and May (ODE, 2008), and the highest of these scores was awarded to each student. This test is designed to measure reading comprehension and to assess whether students are reaching state standards at each grade level. The content standards include the following dimensions of literacy: reading—

- for literary experience (fiction, drama, poetry),
- to gain information (articles, biographies, autobiographies)
- to perform a task (instructions, reference, materials).

These are the reading skills that are assessed:

- Vocabulary
- Read to Perform a Task
- Demonstrate General Understanding
- Develop an Interpretation
- Examine Content and Structure: Informative Text

Although the content standards remain consistent over time, the test questions are adaptive and change from year-to-year. The test items are presented in multiple-choice question form.

Student performance is categorized according to four levels of reading proficiency: exceed, met, close to met, and not met (ODE, 2008). Performance in these categories is reported in scale scores that represent a continuous scale from one grade level to the next. The OAKS scale is designed to measure continuous growth in reading so that performance can be compared from year-to-year. The following questions show the growing sophistication of the test items (OAKS 2008/2009 version) for grades 3, 4, 5, and 6:

> **Third graders** read a passage about *Apollo's trip to the moon* and answered the following question: *Based on what you have read in this story, what do you think is true about Apollo's trip to the moon?* (ODE, 2008, p. 8)
>
> **Fourth graders** read about *beetles* and then answered the following question: *What is the main use for a beetle's front wings?* (ODE, 2008, p. 3)
>
> **Fifth graders** read a passage about *Yosemite National Park* and answered the following question: *Why does the author include the example of the bear in Yosemite National Park?* (ODE, 2008, p. 5)
>
> **Sixth graders** read a passage about *Clouds* and answered the following question: *When clouds form, which of these happens before the other three?* (ODE, 2008, p. 5)

DRA. The DRA interim reading assessment uses two subtests that are designed to measure a student's reading engagement, oral fluency and comprehension. The DRA itself is based on 12 premises about what makes a good reader and that align with a Balanced Literacy approach: good readers—

1. choose reading materials to fulfill different purposes and to reflect their interests. Good readers read well-targeted text (text that is accessible at their level) with a high level of success and accuracy....

2. read for extended periods of time that are consistent with the purpose for reading....

3. preview a book before reading in order to predict events, identify topics or theme, or make real-world connections by relating the content to their own experiences....

4. read aloud with fluency for longer period of time....

5. are aware of and use a variety of strategies to decode words and comprehend reading materials....

6. read for meaning and understanding and are able to summarize text in their own words....

7. read and communicate with others using both oral and written discourse....

8. monitor and develop their reading skills....

9. read, comprehend, and interpret text on a literal level....

10. read, interpret text by making use of inferences, and make connections to personal experiences and existing knowledge....

11. validate their inferences, generalizations, connections, and judgments with information from the text, information from other sources, or personal experiences. (Pearson Education, 2009, pp. 7-11)

Beyond these premises the DRA also aligns with Balanced Literacy in the way in which it is administered. Unlike the DIBELS, the DRA is designed to assess students at their independent or instructional reading level rather than at a predetermined reading level based on their grade and the time of year. Teacher judgment plays a critical role in determining the level at which a student will be assessed; on the other hand, teacher judgment is minimized or even eliminated altogether in the administration of the DIBELS.

The DRA is designed to assess three areas of reading competency: Reading Engagement, Oral Reading Fluency, and Comprehension. Two subtests within each of these areas evaluate a student's knowledge and ability:

Reading Engagement describes the student's level of engagement with reading.

Subtest

1. Wide Reading
2. Self-Assessment/Goal Setting

ORF describes the student's oral reading behaviors.

Subtest

1. Expression
2. Phrasing
3. Rate
4. Accuracy
5. Monitoring
6. Problem-Solving

Comprehension describes the student's ability to retell and understand the text.

Subtest

1. Previewing
2. Retelling
3. Making Connections
4. Using Text Features
5. Interpretation
6. Reflections
7. Prediction/Questioning
8. Summary

9. Scaffolding

10. Literal Comprehension

11. Metacognitive Awareness. (Pearson Education, 2009, pp. 23-24)

Student performance on the DRA is scored according to six levels of reading proficiency: Emerging or Developing in levels A-12 and Intervention, Instructional, Independent and Advanced in levels 14-40. Since the focus of this study is on intermediate level students in grades 3-6, definitions of performance expectations for those levels are described here:

Independent:	The total score for Oral Reading Fluency AND Comprehension must be in the independent range.
Instructional:	The total score for Oral Reading Fluency AND Comprehension must be in the instructional range.
Advanced:	The total score for Oral Reading Fluency AND Comprehension must be in the advanced range. (Pearson Education, 2009, p. 25)

Analysis conducted by the Pearson Education (2009) Corporation concluded that the DRA is a reliable and valid measure. Specifically, the reliability analysis demonstrates that the DRA has relatively little measurement error associated with content, time, and rater. Also, validity analyses indicate that the DRA demonstrates content-related, criterion-related and construct validity (Pearson Education, 2009).

Data Collection

For the purpose of the Emerald school district's universal literacy screening, the Accuracy and Monitoring Subtests of the DRA were administered to all third through sixth grade students three times during the 2008-2009 school year—in the Fall (October/Early November), the Winter (Late February/Early March) and the Spring (Late

May). These assessments served two purposes: they identified students who were not reaching benchmark, and they served as a diagnostic assessment to assist teachers in focusing their instruction for struggling readers. Performance on these subtests were categorized as either—

Instructional: The total score for Accuracy AND Monitoring must be in the instructional range.

or

Independent: The total score for Accuracy AND Monitoring must be in the independent range.

Accuracy Subtest

In order to determine a proficiency level for the Accuracy Subtest, a student classified as independent must accurately decode connected text from leveled passage at or above the Independent Rate based on the passage level. For a student to be classified as instructional the student must accurately decode connected text from a leveled passage at or above the Instructional Rate based on the passage level but below the independent level. DRA administration guidelines determined the independent and instructional cut points (Pearson Education, 2009).

Monitoring Subtest

To determine a proficiency level for the Monitoring Subtest, a student classified as independent must be able to self-correct decoding errors at least 50% of the time while reading a DRA Connected Text Passage. For a student to be classified as instructional the student must self-correct decoding errors at least 33% of the time but less the 50% when reading a Connected Text Passage. These cut scores have been widely used by teachers

conducting Balanced Literacy reading assessments such as Running Records (Clay, 2000).

To determine an overall Reading Proficiency Level the passage level, monitoring level, and accuracy levels were combined to arrive at a single score for each student. Table 4 describes how the overall risk classification was determined for grades 3-6. The district's literacy coaches established these classification using as their guide the recommendations of the DRA technical manual as well as classification standards established by states, such as Utah, that use the DRA within an RTI framework.

Table 4

Risk Classifications for DRA

Grade	Third Grade					
Window	Fall (1st Trimester)		Winter (2nd Trimester)		Spring (3rd Trimester)	
Measures	Scores	Status	Scores	Status	Scores	Status
DRA: Accuracy Refer to DRA for Independent level %	24 or Below 28 30 and up	At Risk Some Risk Low Risk	28 or Below 30 34 and up	At Risk Some Risk Low Risk	30 or Below 34 38 and up	At Risk Some Risk Low Risk
DRA: Monitoring for Meaning (self-corrections)	1sc:4errors 1sc:3errors 1sc:2errors	At Risk Some Risk Low Risk	1sc:4errors 1sc:3errors 1sc:2errors	At Risk Some Risk Low Risk	1sc:4errors 1sc:3errors 1sc:2errors	At Risk Some Risk Low Risk
OAKS					204	passing
Grade	Fourth Grade					
Window	Fall (1st Trimester)		Winter (2nd Trimester)		Spring (3rd Trimester)	
Measures	Scores	Status	Scores	Status	Scores	Status
DRA/QRI Accuracy Refer to DRA/QRI for Independent level %	30 or Below 34 38 and up	At Risk Some Risk Low Risk	34 or Below 38 40 and up	At Risk Some Risk Low Risk	34 or Below 38 40 and up	At Risk Some Risk Low Risk

Table 4 (continued)

Risk Classifications for DRA

Grade	Fourth Grade					
Window	Fall (1st Trimester)		Winter (2nd Trimester)		Spring (3rd Trimester)	
Measures	Scores	Status	Scores	Status	Scores	Status
Monitoring for Meaning (self-corrections	1sc:4errors 1sc:3errors 1sc:2errors	At Risk Some Risk Low Risk	1sc:4errors 1sc:3errors 1sc:2errors	At Risk Some Risk Low Risk	1sc:4errors 1sc:3errors 1sc:2errors	At Risk Some Risk Low Risk
OAKS					211	passing
Fifth Grade						
Window	Fall (1st Trimester)		Winter (2nd Trimester)		Spring (3rd Trimester)	
Measures	Scores	Status	Scores	Status	Scores	Status
DRA/QRI Accuracy Refer to DRA/QRI for Independent level %	34 or Below 38 40 and up	At Risk Some Risk Low Risk	38 or Below 40 50 and up	At Risk Some Risk Low Risk	38 or Below 40 50 and up	At Risk Some Risk Low Risk
Monitoring for Meaning (self-corrections	1sc:4errors 1sc:3errors 1sc:2errors	At Risk Some Risk Low Risk	1sc:4errors 1sc:3errors 1sc:2errors	At Risk Some Risk Low Risk	1sc:4errors 1sc:3errors 1sc:2errors	At Risk Some Risk Low Risk
OAKS					218	passing
Fifth Grade						
Fall (1st Trimester)		Winter (2nd Trimester)		Spring (3rd Trimester)		
Measures	Scores	Status	Scores	Status	Scores	Status
DRA/QRI Accuracy Refer to DRA/QRI for Independent level %	38 or Below 40 50 and up	At Risk Some Risk Low Risk	40 or Below 50 60 and up	At Risk Some Risk Low Risk	40 or Below 50 60 and up	At Risk Some Risk Low Risk
Monitoring for Meaning (self-corrections	1sc:4errors 1sc:3errors 1sc:2errors	At Risk Some Risk Low Risk	1sc:4errors 1sc:3errors 1sc:2errors	At Risk Some Risk Low Risk	1sc:4errors 1sc:3errors 1sc:2errors	At Risk Some Risk Low Risk
OAKS					222	passing

If a student who is at that grade level and in that window can independently read a DRA passage at or above the upper reading level passage (i.e., third grade Fall should be at or above a level 30), the student is classified as low risk of not passing the OAKS assessment. If a student who is at that grade level and in that window can independently read a DRA passage at or below the lower reading level passage (i.e., third grade fall is at or a level 24), then the student is classified as at risk of not passing the OAKS assessment. If a student is in between those two levels he/she is classified at some risk.

All students took the DRA screening test between the months of October and November of 2008. Classroom teachers had been trained in its administration and administered the test. Students' passage level was determined by whether they could decode connected text of the leveled passage with at least a 96% accuracy rate and self-correct reading errors at least 50% of the time. Their performance on these passages was compared with their OAKS scores.

All students took the OAKS assessment between January and May of 2009. All teachers were formally trained in OAKS administration and administered the OAKS to their own class according to OAKS administration guidelines (ODE, 2008). In order to ensure reliability, the school assessment coordinator of the OAKS observed classroom teachers at some point during the OAKS administration. Students had the opportunity to take the OAKS assessment up to three times during the year, and their highest score was compared with their performance on the DRA.

Research Design

This study examines whether the two subtests—accuracy and monitoring—of the DRA are comparable to two subtests—accuracy and rate—of a DIBELS ORF in

predicting passage on the OAKS reading test. The research design and analysis structure is similar to what has been used when other researchers have analyzed DIBELS ORF as a screening tool. In order to conduct the research and analyze the DRA subtests, I used SPSS 18 statistical software and conducted all the analysis myself. Within the SPSS software, I used the analysis tools of simple linear regression and multiple regression modeling. Also, I used the descriptive statics analysis tools of frequency analysis and also the crosstabs utility to produce my kappa and gamma statistics.

Research Question 1

Does a relationship exist between the DRA level of text that a student can successfully decode and monitor his own reading and their RIT (Rausch Unit) score on the OAKS Reading Assessment and is the strength of the relationship comparable to DIBELS ORF?

Research design for question 1. In order to measure this relationship the Pearson correlation coefficient analysis tool is used to examine the relationship between the Fall 2008 Monitoring Level score and the best OAKS scores in each grade level group. Also, the Pearson correlation coefficient analysis tool is used to examine the relationship between the Spring 2009 Monitoring Level score and the best OAKS scores for each grade level group.

Analysis for question 1. Since the main purpose of this study is to examine whether Monitoring could be used as an alternate to ORF as an early literacy universal screening tool, the researcher compares the strength of the relationship of Monitoring to OAKS as compared to the strength of the relationship of ORF to statewide assessments.

The correlation (Pearson's *r*) between ORF and statewide accountability assessments in grades three through five has ranged from .61 to .80 ($p < .001$).

Research Question 2

Are elementary students in each grade level group who reach a benchmark level of accuracy when decoding text and who monitor and correct their own decoding errors in the beginning of a school year likely to meet the standard on the OAKS Reading/ Literature assessment by the end of that same year? And conversely, are students with poorly developed decoding skills who do not monitor and correct their own decoding errors in the beginning of a school year unlikely to meet the standard by the end of that same year?

Research design for question 2. The research design for answering this question uses diagnostic efficiency statistics that provide information about the accuracy of predicting a passing or failing score on the OAKS based on whether a student scores above or below a cutoff point on the monitoring screening. First, a frequency analysis is conducted to examine the accuracy of the Monitoring tool in predicting students in each grade level group who were classified as at risk who did not reach standard as determined by their OAKS best score. The percentage of students in each grade level group who were classified as at risk in the fall of 2008 of not reaching benchmark based on their Fall Monitoring screening level is compared with those who did not reach grade level standard in 2009 as determined by their best OAKS score. Conversely a frequency analysis is conducted to examine the accuracy of the monitoring tool in predicting students in each grade level group who were classified as low risk who did reach standard as determined by their OAKS best score. The percentage of students who were classified

as low risk in the fall of 2008 of not reaching benchmark will be compared with the percent of those who actually did meet the state grade level standard in 2009.

Secondly, this study further examines the accuracy of the monitoring screening tool by examining its positive and negative predictive value. The positive predictive value is an indicator of the proportion of students who were correctly identified as at risk out of all students who were identified as at risk on the screening instrument. Conversely, the negative predictive value is an indicator of the proportion of students who were correctly identified as not at risk out of all students identified as not at risk on the screening instrument (Glover & Albers, 2007; Gredler, 1997, 2000; Jenkins et al., 2007; Severson & Walker, 2002; Stage & Jacobsen, 2001; Wood, 2006). This research calculates the percentage of those found to be at risk in this population using sensitivity and specificity values along with approximate base rates. The positive predictive value is estimated as follows:

- Positive Predictive Value is estimated as:

(base rate x sensitivity)/(base rate x sensitivity) + [(1−base rate) x (1−specificity)]

- The negative predictive value is estimated as:

(1- base rate x specificity)/[(1-base rate) x specificity]+ [(base rate) x (1−sensitivity)]

Third, the overall accuracy of the diagnostic cut score is measured with Gamma and Kappa diagnostic efficiency statistics. Gamma measures agreement versus disagreement by calculating the concordance minus the discordance rates divided by concordance plus discordance (Goodman & Kruskal, 1954; Wood, 2006). Kappa evaluates the overall accuracy of classification after correction for chance (Cohen, 1960; Wood, 2006).

Finally, this study evaluates the degree of sensitivity and specificity of the Monitoring tool to identify students as at risk who actually do not reach grade level standards on OAKS and to identify students as low risk who later reach grade level standards on OAKS. The number of students in each grade level group who were classified as at risk in the fall of 2008 of not reaching benchmark based on their Fall Monitoring screening level is compared with the number of students who should have been classified as at risk based on their performance on their best OAKS reading performance in 2009. The number of students in each grade level group who were classified as low risk in the fall of 2008 of not reaching benchmark based on their Fall Monitoring screening level is compared with the number of students who should have been classified as low risk based on their performance on their best OAKS reading performance in 2009.

Analysis for question 2. Again, since the main purpose of this study is to examine whether Monitoring can be used as an alternate to ORF as an early literacy universal screening tool, the accuracy, positive predictive power, negative predictive power, sensitivity and specificity of the monitoring screening tool benchmarks are compared to the accuracy, positive predictive power, negative predictive power, sensitivity and specificity of the ORF benchmarks on high-stakes statewide assessments. The ORF ranges are listed below:

- ORF at risk accuracy ranges from 72% to 82%
- ORF low risk accuracy ranges from 91% to 96%
- ORF accurately classified 86% and 88% of students being assessed
- ORF Positive Predictive Power ranges from 36% to 72%

- ORF Negative Predictive Power ranges from 85% to 97% (Buck & Torgesen, 2002; Good et al., 2001; Hammer, 2003; Hosp & Fuchs, 2005; Jenkins et al., 2007; Shaw & Shaw, 2002; Silberglitt et al., 2006; Wood, 2006).

Research Question 3

Does the Monitoring screening level add to the predictability of performance on the Oregon statewide reading test over and above the previous year's performance on the statewide test?

Research design for question 3. Not only is it important to be able to determine whether Monitoring screening level benchmarks have predictive utility, it is also important to examine whether the DRA reading levels add to the predictability of performance on the Oregon statewide reading test over and above the previous year's performance on the statewide test. A multiple regression analysis is used to examine the correlation between Monitoring screening level and OAKS for each grade level group and the proportion of variance accounted for by each variable—Fall Monitoring screening level and previous year OAKS score. The first statewide proficiency test is given in third grade. Thus, prior year performance was not available for third graders. However, prior year performance on the statewide test for fourth, fifth and sixth grade was available for use in a multiple regression equation to determine whether Monitoring screening level adds to the predictability of current year performance on the statewide reading test.

Analysis for question 3. First the previous years OAKS and current year fall Monitoring screening level is entered simultaneously into a multiple regression analysis to examine whether both independent variables were significant and independent

predictors of current year best OAKS score. Next the current year fall Monitoring screening level were entered into the multiple regression equation after previous year OAKS score to examine if the fall Monitoring screening level accounted for an additional variance in the current year OAKS score. Finally, how much the previous year's OAKS score and the current fall Monitoring screening level account for the variation in current year in OAKS scores is examined.

Limitations and Potential Researcher Bias

Limitations of Study

Some limitations of this study should be noted. First, even though reading-endorsed literacy coaches trained all teaching staff who administered the DRA, it was only the first year of implementation. Furthermore, because the training was not standardized, the level of expertise among the teachers who were administrating the DRA might very well have been inconsistent. This issue is important since teacher judgment plays an important role in determining at which reading passage level a student should be assessed.

Second, it would be important to know how ELL students performed on this screening in comparison to non ELL students because of the high ELL population within the district and because Monitoring was chosen as a subtest over Rate because of this significant population. There was not a large enough sample of ELL students within each grade level to do this analysis. Identifying a sufficiently large sample and conducting similar analysis for this population will be an important next step for the district in its efforts to refine instruction for ELL learners.

Third, the DRA assessment breaks into two strands of reading skills: ORF and Comprehension. Even though both subtests demonstrate moderate inter-item reliability to comprehension, neither the Accuracy subtest nor Monitoring subtest are subtests of the Comprehension strand. Time was the main factor in not choosing a subtest of the Comprehension strand. The district decided that it would take much longer to assess all students using a Comprehension subtest and would have taken too much time away from instruction.

The final limitation of note is the length of the OAKS assessment window. Students can take the OAKS assessment up to three times between the months of January and May; consequently the time frame between the Fall Monitoring screening in October and the best OAKS score for the year varies from student to student. Furthermore, establishing concurrent validity is an important part of the process in establishing predictive validity. It is important to note that the OAKS best score could have been achieved as early as January whereas the spring Monitoring score would not have been recorded until May.

What this Study Does Not Propose to Do

At their best interim assessments have two functions: they can predict future performance on summative assessments, and they can inform educators and evaluate instruction. It is important to understand just how they fulfill these two roles and to distinguish between them. Typically a teacher, a school, or a district chooses an interim assessment based on the belief that the subtests within the assessment will provide valid and reliable indicators of reading competence. The results of these subtests can be used to tell teachers which skills students have in their command and which they still need to

master in order to become successful readers. Administrators and teachers together can also draw on these results to determine the effectiveness of a school-wide or district-wide literacy program.

Although both of these applications are very important, this study is not designed to examine the construct validity of the relationship between the interim assessment and an instructional program. Rather, this study examines the ways in which the DRA can help predict present and future performance on summative assessments such as OAKS.

Potential Researcher Bias

The purpose of this dissertation is to examine the Emerald School District's attempt to develop and implement an interim literacy assessment that is aligned with its progressive pedagogical philosophy. As Emerald's Director of Assessment, I facilitated the development and implementation of this assessment because I was eager to address the claim that there is no empirical research that has analyzed and evaluated the predictive validity of the DRA on passing an end of year high stakes accountability test. It is my belief that, without findings that demonstrate its efficacy, the Emerald School District will not be able to rely on the information that it provides.

Because I serve both as the Director of Assessment for the Emerald School District and as the researcher for this study and because my own pedagogical beliefs in the area of early literacy instruction are that of a progressive educator, obvious concerns about researcher bias must be addressed. Although I was the Director of Assessment while data were being collected and continued in that position as the data was being analyzed, I have not nor will I gain any financial benefit from the success of the

implementation of the universal screening tool. Furthermore, the results of this study will not affect my employment or my performance evaluation.

The intent of this study is to inform policy makers, first and foremost the leadership of the Emerald School District, as to whether this DRA can be used to evaluate student learning. If the DRA cannot serve this purpose, only time has been lost in the endeavor. The Leadership Team will continue to examine other options for its interim assessment in order to reach its goal of achieving congruency between all the components of an instructional program and a comprehensive assessment system.

Conclusion

This research design and analysis is intended to examine Emerald's effort to establish a universal reading screening tool that is part of an interim assessment. If this screening tool demonstrates both concurrent and predictive validity to the OAKS assessment, then the DRA assessment has the potential to meet the needs a district has for an assessment of this type. If so, the DRA will be both predictive of future performance on summative assessments and help educators inform and evaluate instruction. Furthermore, the DRA interim assessment can then be integrated into Balanced Literacy instructional programs, increase instructional minutes by not having to duplicate assessments and finally, and perhaps most importantly, progressive educators in Oregon will have access to an assessment that is congruent with their philosophic and pedagogical beliefs and which also gives them the ability to link their instruction to high stakes accountability results without being forced to narrow their instruction and to compromise what they believe are the best instructional practices.

CHAPTER IV

RESULTS

The purpose of this dissertation is to examine one school district's attempt to develop and implement an interim literacy assessment that is aligned with its progressive pedagogical philosophy which, in this case, was the DRA interim literacy assessment. The leadership of this school district directed me, in my capacity as the Director of Assessment, to facilitate the development of this tool. Over a span of 2 years, I worked with the literacy coaches, reading specialist and the curriculum director to facilitate its implementation. However, I was concerned about the lack of empirical evidence that the DRA was as strong an interim assessment as DIBELS. I felt that it was important to study the comparative effectiveness of these two interim assessments in order to reassure myself as well the Leadership Team and teachers of Emerald that the DRA was an effective interim assessment tool for progressive schools in Oregon. The results of this study are described in this chapter, and their purpose is to examine whether the two subtests—Accuracy and Monitoring—of the DRA can be used as a screening tool to predict passage on the OAKS reading test. The results of these findings, which will be discussed in chapter 5, were used to evaluate whether the DRA is a viable alternative to the DIBELS ORF interim literacy assessment.

This chapter restates each research question and reports the results of the statistical analyses performed to address the questions. Finally, I summarize the results as a whole.

Research Question 1

Is there a relationship between the DRA text level at which a student can successfully decode and monitor his own reading and his RIT (Rausch Unit) score on the OAKS Reading Assessment?

Analysis: Correlation Between OAKS and DRA Screening Tool Text Level

The Pearson correlation coefficient analysis tool was used to measure this relationship. Table 5 shows the Pearson correlation coefficient among the fall and spring DRA screening tool text level and OAKS scores.

Table 5

Correlation Coefficients

	Standardized Coefficients		Standard			
Grade/Window	Beta		Error		Significance	
	Fall	Spring	Fall	Spring	Fall	Spring
Third Grade	.74	.77	.039	.029	.000	.000
Fourth Grade	.67	.75	.033	.028	.000	.000
Fifth Grade	.65	.69	.032	.029	.000	.000
Sixth Grade	.62	.73	.045	.039	.000	.000
Overall	.75	.79	.016	.014	.000	.000

Note: Correlation is significant at the 0.01 level (two-tailed).

Evaluation: Correlation Between OAKS and DRA Screening Tool

The correlation analysis above examined the relationship between DRA screening tool for Fall 2008 and the OAKS best scores for 2009 as well as the relationship between the DRA screening tool for Spring 2009 and the OAKS best scores for 2009. The

Correlation coefficients demonstrated a significant relationship between the DRA screening tool for both Fall 2008 score and spring 2009 score and OAKS Reading RIT in grades 3-6.

Since the main purpose of this study is to examine whether Monitoring can be used as an alternate to the ORF screening tool as an early literacy universal screening tool, it is necessary to compare whether the strength of the relationship between the DRA screening tool and OAKS is comparable to the relationship between the ORF screening tool and statewide assessments. The correlation (Pearson's *r*) between ORF and statewide accountability assessments in grades three through five has ranged from .61 to .80 ($p < .001$), while the correlation between the monitoring and OAKS in grades three through six ranged from .62 to .79 ($p <. 001$). Based on this analysis, it is clear that the DRA screening tool is comparable to the ORF screening tool in its relationship to statewide assessments.

Research Question 2

Are elementary students in grades 3-6 who reach a benchmark level of accuracy when decoding text and who monitor and correct their own decoding errors in the beginning of a school year likely to meet the standard on the OAKS Reading/ Literature assessment by the end of that same year? And conversely, are students with poorly developed decoding skills who do not monitor and correct their own decoding errors in the beginning of a school year unlikely to meet the standard by the end of that same year?

The second research question posed was whether or not the DRA screening tool was accurate in predicting a passing or failing score on the OAKS accountability test.

Specifically, are elementary students in grades 3-6 who reach a benchmark level of accuracy when decoding text and who monitor and correct their own decoding errors in the beginning of a school year likely to meet the standard on the OAKS Reading/ Literature assessment by the end of that same year? And conversely, are students with poorly developed decoding skills who do not monitor and correct their own decoding errors in the beginning of a school year unlikely to meet the standard by the end of that same year?

Analysis: Accuracy and Predictive Validity

Diagnostic efficiency statistics were used on the monitoring tool to evaluate its accuracy and predictive validity. Accuracy was evaluated by first conducting frequency analysis to examine the accuracy of the Monitoring tool in predicting which students in each grade level group were at risk and which were low risk of not meeting grade level standards as determined by OAKS. Table 6 displays the results of this analysis:

Table 6

Accuracy of DRA Screening Tool

Grade		At Risk & Did Not Meet Benchmark	Low Risk & Did Meet Benchmark	Accurate Prediction
Third	%	75%	96%	92%
	#	40	196	136
Fourth	%	82%	91%	90%
	#	31	178	209
Fifth	%	93%	86%	86%
	#	26	185	211
Sixth	%	86%	96%	93%
	#	31	108	139

Secondly, Kappa and Gamma diagnostic efficiency statistics were applied to determine the level of agreement between the DRA screening tool and OAKS on the reading achievement of student's determined to be at risk or low risk of meeting grade level reading standards. Table 7 shows the results of this analysis.

Table 7

Efficiency of DRA Screening Tool

Diagnostic Efficiency Grade/Window	KAPPA				GAMMA			
	Fall	Sig	Spring	Sig	Fall	Sig	Spring	Sig
Third Grade	.716	.000	.737	.000	.978	.000	.973	.000
Fourth Grade	.623	.000	.659	.000	.935	.000	.958	.000
Fifth Grade	.544	.000	.541	.000	.923	.000	.975	.000
Sixth Grade	.581	.000	.817	.000	.982	.000	.985	.000

Predictive validity was evaluated by first, calculating the proportion of students who were correctly identified as at risk and low risk in the fall using the DRA screening tool based on the positive and negative predictive values. The sensitivity and specificity of the monitoring tool were calculated in order to identify students as at risk who do not reach grade level standards on OAKS and to identify students as low risk who later reach grade level standards on OAKS. The sensitivity and specificity results were used as part of the positive and negative predictive value calculations which determined the over all predictive validity of the DRA screening tool. Table 8 displays the results of this analysis.

Table 8

Positive and Negative Predictive Power of the DRA Screening Tool

Analysis	Third	Fourth	Fifth	Sixth
Positive Predictive Power	75%	59%	73%	96%
Negative Predictive Power	97%	95%	91%	80%
Sensitivity	91%	80%	52%	48%
Specificity	90%	88%	96%	99%
Passing Base Rate	75%	82%	83%	68%
Failing Base Rate	25%	18%	17%	32%

Evaluation: Accuracy and Predictive Validity

Grade 3: Accuracy. Of the 204 students who scored at the benchmark goal of 38 or above on the Monitoring screening in the spring of 2009, 196 (96%) scored meeting benchmark or exceeding benchmark on the OAKS assessment. However, only 13 out of the 53 (25%) students who scored less than 30 on the Spring Monitoring screening met benchmark. 257 out of 275 (93%) of the third grade students assessed fell within the classifications of at risk or low risk.

Furthermore, the diagnostic efficiency statistics of Gamma indicated a strong agreement between the Monitoring screening and the OAKS best score in 2009 in identifying which students were at risk and which students were low risk of not reaching grade level standards. Finally, Kappa indicated that the level of diagnostic efficiency above chance, in grade three, was 72% in the fall screening and 74% in the spring screening.

Grade 3: Predictive validity. Sensitivity and specificity calculations were used to determine the predictive power of the monitoring tool. The positive predictive power

(i.e. the probability of correctly predicting who would not reach grade level expectations based on OAKS) of 75% was above the base rate in grade three of 25%, while the negative predictive power (i.e. the probability of correctly predicting who would reach grade level expectations on OAKS) of 97% was also above the rate in grade three of 75%.

Grade 4: Accuracy. Of the students who scored at the benchmark goal of 40 or above on the Monitoring screening in the spring of 2009, 178 of the 195 students (91%) scored meeting benchmark or exceeding benchmark on the OAKS assessment. However, only 7 out of the 31 (18%) of the students who scored less than 34 on the Spring Monitoring screening met benchmark on the OAKS assessment. Furthermore 233 out of 247 (94%) of the grade four students assessed fell within the classifications of at risk or low risk. It would appear that the grade four benchmarks of 34 for at risk and 40 for low risk in the spring are sufficient for establishing a reasonable probability of meeting or exceeding benchmark on the OAKS assessment or not meeting or exceeding OAKS assessment.

Gamma, for grade 4 also indicated a strong agreement between the DRA screening tool and the OAKS best score in 2009 in identifying which students were at risk and which students were low risk of not reaching grade level standards. Finally, Kappa indicated that the level of diagnostic efficiency above chance was 62% in the fall screening and 66% in the spring screening for grade four.

Grade 4: Predictive validity. The positive predictive power of 59% was above the base rate in grade four of 18%, while the negative predictive power of 95% was also above the rate in grade four of 82%.

Grade 5: Accuracy. Of the students who scored at the benchmark goal of 50 or above on the DRA in the spring of 2009, 185 of 216 students (86%) scored meeting benchmark or exceeding benchmark on the OAKS assessment. However, only 2 out of the 28 (7%) of the students who scored less than 38 on the Spring DRA met benchmark on the OAKS assessment. Furthermore 244 out of 268 (91%) fell within the classifications of at risk or low risk. It would appear that the third grade benchmark of 38 for at risk and 50 for low risk in the spring are sufficient benchmarks for establishing a reasonable probability of meeting or exceeding benchmark on the OAKS assessment or not meeting or exceeding OAKS assessment.

Gamma, for grade 5, also indicated a strong agreement between the DRA screening tool and the OAKS best score in 2009 in identifying which students were at risk and which students were at low risk of not reaching grade level standards. Finally, Kappa indicated that the level of diagnostic efficiency above chance was 54% for both the fall and spring screening for grade five.

Grade 5: Predictive validity. The positive predictive power of 73% was above the base rate in grade five of 17%, while the negative predictive power of 91% was also above the rate in grade five of 83%.

Grade 6: Accuracy. Of the students who scored at the benchmark goal of 60 or above on the DRA in the spring of 2009, 108 of 113 (96%) scored meeting benchmark or exceeding benchmark on the OAKS assessment. However, only 5 out of the 36 (14%) of the students who scored less than 40 on the Spring DRA met benchmark on the OAKS assessment. Furthermore 149 out of 168 (89%) of the sixth grade students assessed fell within the classifications of at risk or low risk. It would appear that the third grade

benchmarks of 40 for at risk and 60 for low risk in the spring are sufficient for establishing a reasonable probability of meeting or exceeding benchmark on the OAKS assessment or not meeting or exceeding OAKS assessment.

Gamma, for grade 6, also indicated a strong agreement between the DRA screening tool and the OAKS best score in 2009 regarding which students were at risk and which students were at low risk of not reaching grade level standards. Finally, Kappa indicated that the level of diagnostic efficiency above chance was 58% for the fall screening and 82% for the spring screening for grade six.

Grade 6: Predictive validity. The positive predictive power of 96% was above the base rate in grade six of 32%, while the negative predictive power of 80% was also above the rate in grade six of 68%.

Grade level summary. The above analysis appears to demonstrate that the linkage between Emerald school district's DRA screening tool benchmarks and performance on Oregon's high stakes accountability assessment is sufficient to inform practitioners about the progress students are making toward meeting or exceeding the State's reading standards. However, since the main purpose of this study is to examine whether Monitoring can be used as an alternate to the DIBELS ORF screening tool as an early literacy universal screening tool, it is also important to compare the accuracy and predictive validity of the tool to DIBELS ORF screening tool ranges.

In this comparison the DRA screening tool compares favorably to DIBELS ORF screening tool in accurately classifying students at being at risk or low risk. The DRA screening tool accurately classified at risk student correctly 75% to 93% of the time compared to DIBELS ORF screening tool which accurately classified at risk students

correctly 72% to 82% of the time. The DRA screening tool accurately classified low risk students correctly 86% to 96% of the time, while the DIBELS ORF screening tool accurately classified low risk students 91% to 96%. The only grade level that fell below the DIBELS ORF screening tool range was for the classification of low risk students in grade 5.

With regard to predictive validity, the DRA screening tool again compared favorably to ORF in its predictive validity. The positive predictive power of the DRA screening tool ranged from 59% to 96% compared to DIBELS ORF screening tool's positive predictive power which ranged from 36% to 72%. The negative predictive power of the DRA screening tool ranged from 80% to 97% compared to the DIBELS ORF screening tool's negative predictive power which ranged from 86% to 97%.

In summary, it appears that the accuracy and predictive validity of the Emerald School District's DRA screening tool is comparable to DIBELS ORF screening tool.

Research Question 3

Does the Monitoring screening level add to the predictability of performance on the Oregon statewide reading test over and above the previous year's performance on the statewide test?

Analysis: Value Added Predictability of the DRA Screening Tool

Not only is it important to be able to determine whether the DRA screening tool has predictive utility; it is also important to examine whether the tool adds to the predictability of performance on the Oregon statewide reading test over and above the previous year's performance on the statewide test. A multiple regression analysis was used to examine the correlation between the DRA screening tool and OAKS and the

proportion of variance accounted for by each variable—fall Monitoring screening level of performance and the previous year's OAKS score.

Because, in Oregon, the first statewide proficiency test is given in grade 3, prior year performance was not available for grade 3 students. However, prior year performance on the statewide test for grades 4-6 was available to use in a multiple regression equation to determine whether the Monitoring screening level adds to the predictability of current year performance on the statewide reading test. The results are presented in Table 9, which provides the standardized regression coefficients (β) and Table 10 which provides the proportion of variance accounted for by each variable.

Table 9

Standardized Multiple Regression Coefficients

		2007-08 OAKS Reading Score				DRA Fall 2008 Screening Level			
		Standardized Coefficients				Standardized Coefficients			
Grade	R^2	β	B	*t*	Sig	B	B	*t*	Sig
4th	.62	.54	.43	9.107	.000	.32	.23	5.352	.000
5th	.64	.52	.39	9.351	.000	.35	.21	6.401	.000
6th	.80	.83	.81	16.409	.000	.10	.07	1.896	.060

Table 10

Multiple Regression Equation

		Proportion of Variance Accounted For When Entered into the Equation			
		Fall DRA		Previous Year OAKS Score	
Grade	Total Variance	First	Second	First	Second
4th	62%	46%	6%	56%	16%
5th	64%	49%	7%	57%	15%
6th	80%	41%	< 1%	80%	< 1%

Fourth grade. A multiple regression was performed using the grade 4 OAKS data in this study as the dependent variable. There were two predictors of grade 4 OAKS performance: third grade best OAKS scores and grade 4 fall Monitoring screening level. Of the grade four students, 34 were new to the district at the beginning of fourth grade and so did not have grade three OAKS scores. Thus, the sample size for this analysis was 213 instead of the original 247 fourth graders. The correlation between the grade three OAKS and grade four OAKS was .75. The correlation between grade four fall Monitoring screening level and grade four OAKS was .68. When grade three OAKS and grade four fall Monitoring screening level were entered simultaneously, regression analysis indicated that both were significant predictors of grade four OAKS (grade three OAKS, $t = 9.107$, $p < .001$; grade four fall Monitoring screening level, $t = 5.352$, $p < .001$). When fall Monitoring screening level was entered into the equation after grade four OAKS, it accounted for an additional 6% of the variance in grade four OAKS. Together, grade four OAKS and grade four fall Monitoring screening level accounted for 62% of the variation in fourth grade OAKS scores.

Fifth grade. A multiple regression was performed using the grade 5 OAKS data in this study as the dependent variable. There were two predictors of grade 5 OAKS performance: grade 4 best OAKS scores and grade 5 fall Monitoring screening level. Of the grade five students, 26 were new to the district at the beginning of grade five and so did not have grade four OAKS scores. Thus, the sample size for this analysis was 242 instead of the original 268 grade four students. The correlation between grade four OAKS and grade five OAKS was .76. The correlation between grade five fall Monitoring screening level and grade five OAKS was .71. When grade four OAKS and grade five

fall Monitoring screening level were entered simultaneously, regression analysis indicated that both were significant predictors of grade five OAKS (grade four OAKS, $t = 9.351$, $p < .001$; grade five fall Monitoring screening level, $t = 6.401$, $p < .001$). When fall Monitoring screening level was entered into the equation after grade four OAKS, it accounted for an additional 7% of the variance in grade five OAKS. Together, grade four OAKS and grade five fall Monitoring screening level accounted for 64% of the variation in grade five OAKS scores.

Sixth grade. Finally a multiple regression was performed using the grade 6 OAKS data in this study as the dependent variable. There were two predictors of grade 5 OAKS performance: grade 5 best OAKS scores and grade 6 fall Monitoring screening level. Of the grade six students, 14 were new to the district at the beginning of sixth grade and so did not have grade five OAKS scores. Thus, the sample size for this analysis was 154 instead of the original 168 grade six students. The correlation between grade five OAKS and grade six OAKS was .89. The correlation between grade six fall Monitoring screening level and grade six OAKS was .64. When grade five OAKS and grade six fall Monitoring screening level were entered simultaneously, regression analysis indicated that the previous year's test score was a significant independent predictor of grade six OAKS, but the fall Monitoring screening level was not (grade five OAKS, $t = 16.409$, $p < .001$; grade six fall Monitoring screening level, $t = 1.896$, $p < .057$). When fall Monitoring screening level was entered into the equation after grade five OAKS, it accounted for less than 1% of the variance in sixth grade OAKS. Together, grade five OAKS and grade six fall Monitoring screening level accounted for 80% of the variation in grade six OAKS scores.

Evaluation: Value Added Predictability of the DRA Screening Tool

In evaluating the analysis of whether the DRA screening tool adds to the predictability of performance on the Oregon statewide reading test over and above the previous year's performance on the statewide test, it appears that overall it does for grades 4 and 5 but not for grade 6. In grades 4 and 6 both the DRA screening tool and the previous years OAKS performance were significant predictors of current year OAKS performance, while in grade 6 only the previous year OAKS score was a significant predictor of current year OAKS performance independently of the other independent variable used in this study.

Conclusion

This study examined whether the two subtests—Accuracy and Monitoring—of the DRA can be used as a screening tool to predict passage on the OAKS reading test, and, if so, whether or not it is comparable to ORF screening tool as an early literacy universal screening.

The preceding discussion first looked at whether a significant relationship existed between DRA screening tool and OAKS and whether that relationship was comparable to the DIBELS ORF screening tool. Secondly it evaluated the accuracy and predictive validity of the DRA screening tool in classifying at risk and low risk students and whether its accuracy and predictive utility were comparable to the DIBELS ORF screening tool. Finally it analyzed whether the Emerald screening tool adds to the predictability of performance on the Oregon statewide reading test over and above the previous year's performance on the statewide test.

The results of this study appear to demonstrate that (a) the accuracy, predictive validity and the correlation to the OAKS high stakes accountability of the DRA screening tool is comparable to DIBELS ORF screening tool, (b) the DRA screening tool adds to the predictability of performance on the Oregon statewide reading test over and above the previous year's performance on the statewide test for grades 4 and 5 but not for grade 6, and (c) the DRA screening tool is a significant independent predictor of OAKS performance in grades 4 and 5.

The following discussion in chapter 5 builds on these findings to suggest further lines of research and make recommendations to the policy makers of Emerald School District as well as to other districts wishing to adopt the DRA as an interim literacy assessment.

CHAPTER V

DISCUSSION AND CONCLUSIONS

The Significance of the Findings

This study investigated the correlation and predictive value of the DRA on a statewide reading proficiency test such as OAKS in order to establish its utility as a viable alternative to DIBELS. Specifically, it examined whether or not the Accuracy and Monitoring subtests of the DRA could be an effective screening tool to predict if students were on track to pass the OAKS reading test. The study demonstrated a strong relationship between this screening tool as used by the Emerald School District and performance on OAKS. The Correlation coefficients demonstrated a significant relationship between the screening tool for both Fall 2008 and Spring 2009 Monitoring scores and OAKS Reading RIT in grades 3-6. These results are comparable to previous studies that have demonstrated a significant relationship between DIBELS ORF and high stakes reading tests (Buck & Torgesen, 2002; Good et al., 2001; Schilling et al., 2007; Shaw & Shaw, 2002; Silberglitt et al., 2006; Vander Meer et al., 2005). It seems safe to conclude that the DRA screening tool is comparable to the DIBELS ORF screening tool in the strength of its relation to student performance on statewide assessments

Results from the study also support the claim that brief curriculum-based measures, such as DRA screening tool and DIBELS ORF, can provide meaningful indicators of a student's ability to perform well on high stakes accountability tests. The

question of whether the DRA text level predicts performance on reading achievement was answered. It was found that the DRA text level at which students could accurately decode and monitor their reading correlated to their performance on the OAKS assessment. Additionally, the diagnostic efficiency statistics indicated that the Monitoring screening cut scores provided valuable information about whether students would pass or fail the OAKS. These predictors were significantly above chance levels, and the overall accuracy rates were above base rates. These findings suggest that the DRA can be used as a screening tool to detect potential reading difficulties.

The final purpose of this study was to evaluate whether the screening tool predicts performance on the statewide reading test over and above the predictability of prior year performance on the statewide test. Here the results were mixed. Third grade was not evaluated because it is the first year students are assessed and therefore no previous scores were available. In grades 4 and 5 the screening tool was a significant predictor of performance on the OAKS even after including prior year OAKS performance in the multiple regression analysis. Even though the screening tool's contribution to the proportion of variance for grades 4 and 5 is relatively small, it is still significant. These results are comparable to a previous study demonstrating that DIBELS ORF is a significant predictor of performance on high stakes statewide reading tests (9% for grade four and 4% for grade five) over and above prior year performance for grades four and five (Wood, 2006). In grade 6 the results indicated that the screening tool produced no additional variation of performance on current OAKS performance over and above the prior year performance. One might hypothesize that the difference between grades four and five and grade six in relation to additional variation is due to the difference in degree

of variation that is already explained by the previous year's test score in grade six (80%) compared to grades four and five (62% and 64%). This high degree of correlation between fifth grade performance on OAKS and sixth grade performance might make it difficult for any measures to explain any additional variation.

The multiple regression analysis yielded some of the most interesting findings in the study. Although school districts can gain additional information about individual students in grades 4 and 5 from curriculum-based measures, such as the DRA screening tool or the DIBELS ORF, the amount of information is relatively small.

Implications for Policy and Leadership

The research conducted in this study is potentially important. If the DRA screening tool is comparable to DIBELS ORF with regard to correlation and predictive utility to the OAKS assessment, then progressive districts in Oregon who are currently using the full DRA will now be able to use the Accuracy and Monitoring subtests as their screening tool. Just as districts that have adopted a Direct Instruction model rely on DIBELS, districts using a Balanced Literacy framework can rely on the DRA. They will be able to integrate their assessment practices with their instructional programs and also use their assessments to inform their instructional decisions.

Policy makers who are considering using the DRA as an interim assessment should be aware that, even though this study suggests that it will is an effective screening tool, there are also cautions. First of all, unlike DIBELS, the DRA does not have a significant body of research with regard to its use in an RTI model. Districts would be well advised to continue to invest in the research necessary to ensure that the DRA is as viable tool as it appears to be.

As the Emerald school district continues work toward the implementation of the DRA as an interim assessment, I have advised the policy makers, such as the superintendent and curriculum director, to be cautious about touting the success of the DRA until further research is conducted. In the world of educational politics it seems that educational leaders are understandably anxious to promote their school district's accomplishments. Many times this desire causes them to rush these successes to the public before they can actually be verified. Emerald policy makers are no different in this way, so asking them continue to spend money on research on the tool while maintaining reserve about its success is a tough sell.

Second, the findings from this study—and others like it (Wood, 2006) that have evaluated the added value of using a screening tool over and above the previous year's statewide assessment—should evoke serious discussion about its use. A district needs to consider the cost/benefit of gathering this additional information. It should further examine the differences in the variation among grades that these curriculum-based measures are able to explain. There may be multiple explanations. One possibility is that the scores may have differing degrees of variation from one year of OAKS scores to the next. For example, 57% of the variation in fifth grade scores is explained by the previous year's scores, but 80% of the variation in sixth grade scores is explained by the previous year's scores. If the screening tool provides only a small amount of additional variance over and above the previous year's test score, policy makers should consider whether it will suffice just to use the test score as a screen. In such an approach an interim assessment (i.e. the complete DRA or the DIBEL) would be administered only to students who do not have a previous year's score or who have been identified by last

year's score as at risk of not reaching benchmark. I am not advocating for either position. I do suggest, though, that policy makers closely examine this issue and make a decision about which approach best serves their needs and that they clearly communicate rationale for their decision to their staff.

When I began this discussion with policy makers and the instructional leadership of the Emerald School District , the complexity these decisions became apparent. To the instructional leadership abandoning the use of the DRA would be catastrophic in their view for their attempts to improve literacy instruction. Literacy coaches and the elementary principals view the DRA as a way of not only assessing the students' progress but also to evaluating whether teachers actually understand their students' reading abilities. Coaches and principals have begun analyzing the assessments at the building level to determine which teachers need targeted professional development in the area literacy. On the other hand the human resource director and the superintendent are concerned about labor issues related to work load because of what they have seen happening in neighboring districts that have mandated too much testing.

Finally, if the instructional leadership wins this political battle over the use of the DRA, then districts such as Emerald, need to make sure that they invest the infrastructure necessary to implement, develop, and sustain a DRA interim assessment so that teachers and policy makers can have an assessment tool in which they have full confidence. All elementary teachers will need to have access to ongoing professional development so that they are fully capable of assessing students' reading levels and abilities. This policy decisions brings in other political players into the discussion. The finance director, superintendent and the school board are currently looking at a significant shortfall in next

year's budget. Dollars for professional development will pitted against extracurricular activities and class size. Nobody disputes the importance high quality imbedded professional development for teachers, but at what cost? Politics is essentially the distribution of resources. When resources are plenty political tension is low; when resources are scarce political tension runs high. For the policy makers of Emerald the political tension is currently extremely high. The instructional leadership of the district will be watching closely to see what decision is made.

Emerald's Search for Congruency

Emerald School District's journey to develop an interim literacy assessment as one important part of an over arching goal of realizing congruent curricular and instructional programs is woven throughout this dissertation. This effort has received so much attention because achieving congruence is a complex and difficult process that frankly eludes a great many school systems. Congruency is achieved by doing the difficult work of arriving at a common philosophy and then developing goals and instructional strategies that flow from this philosophy (Allington, 2009; Cuban, 2007a, 2007b; Wilson-Bridgman, 2003). Ultimately, it requires the clarity that comes from having arrived at a common language that all the stakeholders understand and use. The Emerald Leadership Team was firmly convinced that an interim reading assessment that was aligned with its long history and culture of constructivism and progressive education was an essential component of its effort to develop this congruency. They believed that it would lead to a shared understanding about students' reading progress across grade levels and programs. Since this study demonstrates that the DRA has the potential to serve both as an interim assessment and also as a high quality screening tool, the district will now be

able to move forward with its reform effort being confident in the DRA's efficacy and utility.

With an interim assessment that is philosophically and instructionally aligned with the Emerald's core reading program, the supplemental programs of Title I and Special Education can be synchronized with the core. The Leadership Team and staff will have a common assessment tool at its disposal that fits with their belief system and that reliably predicts performance on OAKS. It will then be in a position to eliminate cognitive confusion for students and staff in a several ways:

- Supplemental reading programs can be put in place that have philosophies, goals and instructional strategies that are well-matched with the core program.
- Goals both in the Title I program and goals on student's special education IEP's can now be based on the information derived from the DRA assessment.
- Because the instructional goals are the same for both the core reading instruction and the supplemental programs, the instructional strategies will be more readily aligned.
- The reforms can be expected to result in improved reading performance for at risk readers.
- When curricular and instructional programs are not only congruent but reflect evidence based pedagogical procedures, there is even greater potential for eliminating the cognitive confusion struggling readers often experience (Allington, 2009; Wilson-Bridgman, 2003).

A further benefit that an interim assessment that is philosophically and instructionally aligned with the Emerald's core and supplemental reading program has is

the opportunity for district-wide staff development. Developing a high quality expert teaching staff is the key to any serious instructional reform effort. An interim assessment such as the DRA not only assesses the areas of strength and weakness of a student's reading ability, it can also provide a focus of instruction. For teachers to be able to fully utilize the power of a high quality interim reading assessment, they need to have a deep understanding of sound Balanced Literacy instructional practices. The Emerald school district can use the DRA to establish a district-wide professional development plan that helps teachers deepen their understanding of the key areas of reading development and instructional strategies that will increase their students' literacy skills.

Based on the results of this study I recommend that the instructional leadership of the Emerald school district analyze the results of the assessment not just to assess student learning but also teacher understandings in the key areas of reading. To assess well is to understand what one is assessing. I believe Emerald could dramatically improve student learning for at risk students by using the information gleaned from the interim assessment to evaluate which teachers struggle to understand their students reading abilities and then provide targeted professional development to those teachers. Using an interim reading assessment to improve teachers' understanding of the curricular components that underlie effective instruction has the potential to significantly improve instruction for all students and especially those who are at risk of reading failure.

In summary, I strongly recommend that the Emerald School District continue its use of the DRA as part of its effort to arrive at instructional congruency between its core reading program and its supplemental programs. Using the DRA as a basis for developing this congruence can lead both teachers and students to greater understanding about what

makes an effective reader. For students this understanding is related the nature of reading and how they go about making meaning from text. For teachers, this understanding is related to the underlying principals of sound teaching practices grounded in a Balanced Literacy framework. If the programs are congruent and teachers are using a common language that allows them to talk with one another rather than past one another, both students and teachers will understand more and perform better. The Emerald School District's leadership can send a strong message to its staff that achieving congruence is essential to improving achievement for all students. Using a high quality interim assessment well needs to be an essential part of that message.

Future Research

Correlation Between OAKS and DRA Screening Tool Text Level

Some limitations of this study should be noted. These limitations could become the basis for forming significant research questions that would lead to future studies. First, this study is the initial analysis of the DRA text level and its relationship to performance on the OAKS assessment. While the results of this study are promising, further studies should be conducted to verify these findings. Also, this study did not control for subgroups of students such as English Language Learners, Special Education students, and different demographics of students. Research that analyzed the relationship between OAKS and the DRA while controlling for these variables would be very valuable contributions to the body of knowledge regarding the DRA.

The final limitation, in the area of correlational analysis, that is the most exiting area for future study is in the relation between OAKS and the DRA across grade levels and teachers. This study did not control for grade level or different assessors (i.e.

teachers). If it had, then individual students, teachers, and schools could be meaningfully compared, and the effects that are unique to particular grade levels, teachers or schools could be explored. Measures of effects of a grade level, teacher or school could help answer more question both about the DRA as a interim assessment and screening tool, but also an analysis of this type could help school and district policy makers answer some important curricular and instructional questions such as the following:

- How effective is an individual teacher, school or grade at assessing students reading abilities? and
- How effective is an individual teacher, school or grade at using assessment information to develop a focus of instruction based on assessment results?

The above questions require estimates of the variability among teacher, grade or school effects. If the data and statistical models can accurately describe the contributions of teachers to interim assessment results, the models could provide estimates of the variability among teacher effects and determine the proportion of variability in achievement or growth that is attributable to teachers.

Answers to these questions could have significant impact on both policy and practice. Unlike DIBELS ORF, the administration of the DRA requires teacher judgment when conducting the assessment. If, the DRA is an reliable and valid measure of student progress, even when controlling for teacher judgment, then the DRA can be used both as a screening tool for students and for teachers. School or district leaders could evaluate which teachers are using their formative interim assessments effectively to inform their instructional practice and which need further professional development.

Accuracy and Predictive Validity

Limitations of this study also produced new research opportunities in the area of cut scores to determine which students are at risk and low risk for reaching benchmark. The leadership of the Emerald school district based their cut scores on the technical manual for the DRA (Pearson Education, 2009). Even though these cut scores have adequate positive and negative predictive power and were comparable in accurately classifying at risk and low risk students to DIBELS ORF, further analysis could improve their usefulness. Screening cut scores may be adjusted depending on the needs of the school and or district. A balance must be achieved when setting cut scores for prediction of proficiency on the high stakes accountability test. For example, setting a lower cut score increases the probability that a student scoring below the cut will have a failing score on OAKS. However, this will also increase the chance that a student scoring above the cut score will also have a failing score on OAKS. The rule of thumb for cut scores is that setting a higher cut score will increase the probability of predicting success on the reading test but will also decrease the probability of predicting failure. Cut scores should error on the side of failure identification at the expense of success identification (Glover & Albers, 2007; Jenkins et al., 2007; Wood, 2006). A perfect screen would accurately classify every child who did not reach benchmark as at risk and every child who did reach benchmark as low risk. Obviously it is unreasonable to expect a perfect screen to be developed, but Emerald school district's classification cut scores should be studied carefully to be sure that they are the best that they could be.

Conclusion

I conclude this dissertation by reflecting back to its beginning and particularly to my exploration of the broad philosophic debate between traditional and progressive thinkers. We have been debating the nature and purpose of schooling for a very long time, and there is no reason to believe that the conversation will end any time soon. Educational philosophers, policy makers, leaders, and politicians are all too eager to enter the discussion and to press their points of view and agendas. This debate—that all too often becomes strident and polarizing—continues to play out at the national, state and local level. Right now we are watching it in live time as law makers in Washington, DC work on the reauthorization of the Education and Secondary Education Act (ESEA), as the Oregon State Board of Education determines which standards to adopt and what assessment systems best evaluate these standards, and as local districts attempt to implement reform efforts such as RTI.

The implementation of RTI offers an excellent example of how educational philosophy and reform efforts intersect. If both are not clearly defined, teachers and administrators find themselves mired in program and policy confusion. As I reflect on my own response to presentations about RTI, I realize that, like the other progressive educators of the Emerald School District, their resistance has nothing to do with RTI but rather with how Oregon has intertwined philosophy with this reform effort (Howe, 2008). The implementation of RTI on the State level has a very strong behaviorist agenda, and this focus has led progressive districts such as Emerald to resist putting it into place.

On first glance it may seem that educators, such as myself, are resistant to RTI principals, but this is not the case. In fact, RTI is embraced by both progressive and

traditional educators (Allington, 2009; Howe, 2008). Our resistance stems from our concern that in Oregon RTI implementation recommendations are based on a traditional perspective on learning. I, along with other educators such as Allington, believe that we know why this is the case: Oregon's approach to RTI is so strongly rooted in traditional pedagogical and epistemological assumptions because Oregon is the birthplace of DIBELS and Direct Instruction.

What progressive educators, who are struggling with how to implement RTI in a manner that is compatible with their beliefs, need to do is to take a step back. I urge them to stop arguing about which is better—Direct Instruction or Balanced Literacy, DIBELS or the DRA. Instead, I invite them to think carefully and deeply about their philosophy as it relates to their efforts to develop a high quality comprehensive assessment system. Once they are clear about where they are coming from, they will be in a position to scrutinize their assessment tools to ensure that they are both compatible with their philosophy and are instruments that will work well within an RTI system. The research conducted for this dissertation has demonstrated the strength of the DRA as an interim assessment that is compatible with Balanced Literacy and also robust enough to become an essential component of a comprehensive assessment system. Although this study is straightforward and, at bottom, a simple comparative analysis, it is a beginning and lays a foundation on which progressive school leaders can build as they work with their staffs to achieve congruence throughout their instructional programs.

REFERENCES

Allington, R. L. (1990, March). Curriculum and at-risk learners: Coherence or fragmentation? *PRISE Reporter, 21*(3), 1-3.

Allington, R. L. (2009). *What really matters in response to intervention*. Boston, MA: Allyn and Bacon.

Allington, R. L., & Johnston, P. (1986, June 17-18). *The coordination among regular classroom reading programs and targeted support programs*. Paper presented at the Designs for Compensatory Education Conference, Washinton, DC.

Blume, H. (2009, January 28). L.A. teachers' union calls for boycott of testing. *Los Angeles Times*. Retrieved from http://articles.latimes.com/2009/jan/28/local/me-lausd28

Buck, J., & Torgesen, J. K. (2002). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test*. Tallahassee, FL: Florida Center for Reading Research.

Calkins, L. M. (2001). *The art of teaching reading* (1st ed.). New York, NY: Addison-Wesley Educational Publishers Inc.

Christman, J. B., Neild, R. C., Bulkley, K., Blanc, S., Liu, R., Mitchell, C., & Travers, E. (2009). *Making the most of interim assessment data: Lessons from Philadelphia*. Philadelphia, PA: Research for Action.

Clay, M. M. (2000). *Running records for classroom teachers*. Auckland, New Zealand: Heinemann.

Clune, W. H., & White, P. A. (2008). *Policy effectiveness of interim assessments in Providence Public School*. Madison, WI: University of Wisconsin-Madison, School of Education, Wisconsin Center for Education Research.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cuban, L. (2007a). The never-ending quest. *Education Week, 26*(29), 28-29.

Cuban, L. (2007b). No more magical thinking: Leading from top or bottom. *School Administrator, 64*(3), 6.

Donmore, R. (1999). The continuing quest for a knoweldge base, 1976-1998. In J. Murphy & K. S. Louis (Eds.), *Handbook of research on educational administration* (2nd ed.; pp. 25-44). San Francisco, CA: Jossey-Bass.

Engelmann, S. (2008). *Machinations of What Works Clearinghouse*. Retrieved from http://www.zigsite.com/PDFs/MachinationsWWC(V4).pdf

Ernest, P. (1994, March). Varieties of constructivism: Their metaphors, epistemologogies and pedagogical implications. *Hiroshima Journal of Mathematics Education, 2*, 1-14.

Forbes, S., Poparad, M. A., & McBride, M. (2004). To err is human; to self-correct it to learn. *The Reading Teacher, 57*(6), 566-572.

Glover, T. A., & Albers, C. A. (2007, April). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*(2), 117-135.

Good, R. H., III, & Kaminski, R. A. (2002). DIBELS Oral reading fluency passages for first through third grades (Technical Report). Eugene, OR: University of Oregon.

Good, R. H., III, Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5*(3), 257-288.

Goodman, L. A., & Kruskal, W. H. (1954, December). Measures of association for cross classifications. *Journal of American Statistical Association, 49*(268), 732-764.

Gredler, G. R. (1997). Issues in early childhood screening and assessment. *Psychology in the Schools, 34*, 99-106.

Gredler, G. R. (2000, January). Early childhood education–assessment and intervention: What the future holds. *Psychology in the Schools, 37*(1), 73-79.

Hammer, K. B. (2003). *The effect of oral reading fluency on analysis skills*. San Rafael, CA: Dominican University of California, School of Business, Education, and Leadership.

Hirsch, E. D. (2009, March 23). Reading test dummies. *New York Times*, p. A21.

Hosp, M. K., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review, 34*(1), 9-26.

Howe, K. (2008). *Effective behavior and instructional support systems.* Paper presented at the PBIS/EBISS State Conference, Eugene, OR.

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

Kame'enui, E. J., Carnine, D. W., Dixon, R. C., Simmons, D. C., & Coyne, M. D. (2002). *Effective teaching strategies that accommodate diverse learners* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Kame'enui, E. J., & Simmons, D. C. (1990). *Designing instructional strategies: The prevention of academic learning problems*. Columbus, OH: Merrill Publishing Company.

Kosanovich, M., Ladinsky, K., Nelson, L., & Torgesen, J. (2007). *Differentiated reading instruction: Small group alternative lesson structures for all students. Guidance document for Florida "Reading First" schools.* Tallahassee, FL: Florida Center for Reading Research.

Kucer, S. (2008). Speed, accuracy, and comprehension in the reading of elementary students. *Journal of Reading Education*, *34*(1), 33-8. Retrieved 9 May 2010, from Education Full Text database.

Langer, E. J. (1997). *The power of mindful learning*. Reading, MA: Perseus Books.

Lyon, G. R., Fletcher, J. M., Torgesen, J. K., Shaywitz, S. E., & Chhabra, V. (2004). Preventing and remediating reading failure: A response to Allington. *Educational Leadership, 61*(6), 86.

Madelaine, A., & Wheldall, K. (2005). Identifying low-progress readers: Comparing teacher judgment with a curriculum-based measurement procedure. *International Journal of Disability, Development, and Education*, *52*(1), 33-42.

Manzo, K. K. (2005). National clout of DIBELS test draws scrutiny. *Education Week, 25*(5), 1.

Marshall, K. (2008, September). Interim assessments: A user's guide. *Phi Delta Kappan, 90*(1), 64-68.

Mathes, P. G., & Torgesen, J. K. (1998). All children can learn to read: Critical care for the prevention of reading failure. *Peabody Journal of Education, 73*, 317-340.

Meier, D. (2000). *Will standards save public education*. Boston, MA: Beacon Press.

Meier, D. (2010, March 9). Bridging differences: What I did not recant or abandon [Web log post]. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/Bridging-Differences/2010/03/what_i_did_not_recant_or_aband.html

Murphy, J., & Louis, K. S. (Eds.). (1999). *Handbook of research on educational administration* (2nd ed.). San Francisco,CA: Jossey-Bass.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction [on-line]*. Retrieved from http://www.nichd.nih.gov/publications/nrp/smallbook.htm

Newmann, F. M., Smith, B., Allensworth, E., & Bryk, A. S. (2001). Instructional program coherence: What it is and why it should guide school improvement policy. *Educational Evaluation and Policy Analysis, 23*(4), 297-321.

Oregon Department of Education. (2008, October 1). *Oregon assessment of knowledge and skills.* Salem, OR: Author. Retrieved from http://www.ode.state.or.us

Owens, R. G. (2001). *Organizational behavior in education: Instructional leadership and school reform* (7th ed.). Boston, MA: Allyn & Bacon.

Pearson, P. D. (2004). The reading wars. *Educational Policy, 18*(1), 216-252.

Pearson Education. (2009). *DRA2 technical manual*. Retrieved from http://www.pearsonschool.com

Perie, M., Marion, S., & Gong, B. (2007). *A framework for considering interim assessments*. Washington, DC: Center for Assessment.

Phillips, D. C., & Soltis, J. F. (1998). *Perspectives on learning* (3rd ed.). New York, NY: Teachers College Press.

Popham, J. W. (2008). *Transformative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Roberts, G., Torgesen, J. K., Boardman, A., & Scammacca, N. (2008). Evidence-based strategies for feading instruction of older students with learning disabilities. *Learning Disabilities Research and Practice, 23*(2), 63-69.

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology, 46*(3), 343-366.

Schilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *Elementary School Journal, 107*(5), 429-448.

Severson, H. H., & Walker, H. M. (2002). *Pro-active approaches for identifying children at risk for socio-behavioral problems*. Boston, MA: Allyn & Bacon.

Shaw, R., & Shaw, D. (2002). *DIBELS oral reading fluency-based indicators of third grade reading skills for Colorado State Assessment Program (CSAP)* (Technical Report). Eugene, OR: University of Oregon.

Shelton, N. R., Altwerger, B., & Jordan, N. (2009). Does DIBELS put reading first? *Literacy Research and Instruction, 48*(2), 137-148.

Silberglitt, B., Burns, M. K., Madyun, N. H., & Lail, K. E. (2006, May). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in the Schools, 43*(5), 527-535.

Simmons, D. C., Kame'enui, E. J., Good, R. H., III, Harn, B. A., Cole, C., & Braun, D. (2000). Building, implementing, and sustaining a beginning reading model: School by school and lessons learned. *OSSC Bulletin, 43*(3), 6-30.

Sirotnik, K. A., & Kimball, K. (1999, November). Standards for standards-based accountability systems. *Phi Delta Kappan, 81*(3), 209-214.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children: Executive Summary*. Washington, DC: National Academy of Sciences. Retreived from http://www2.ed.gov/inits/americareads/ReadDiff/read-sum.html

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*(3), 407-419.

Stiggins, R. J., Arter, J. A., Chappuis, J., & Chappuis, S. (2006). *Classroom assessment for student learning: Doing it right - using it well*. Portland, OR: Assessment Training Institute Inc.

University of Oregon. (2010). *DIBELS data system*. Eugene, OR: University of Oregon Center on Teaching and Learning. Retrieved from https://dibels.uoregon.edu/

Vander Meer, C. D., Lentz, F. E., & Stollar, S. (2005). *The relationship between oral reading fluency and Ohio proficiency testing in reading*. Eugene, OR: University of Oregon.

Wilson, J. (2005). *The relationship of Dynamic Indicators of Basic Early Literacy Skills (DIBELS) oral reading fluency to performance on Arizona Instrument to Measure Standards (AIMS)* (Research Brief). Tempe, AZ: Tempe School District No. 3.

Wilson-Bridgman, J. (2003). *Curricular congruence at an implementation level: Are two interventions of a district's early literacy project (improved classroom instruction and Reading Recovery) working congruently to address the needs of low-achieving readers?* Niagra University, NY. Retrieved from http://www.eric.ed.gov/PDFS/ED478118.pdf

Wood, D. E. (2006). Modeling the relationship between oral reading fluency and performance on a statewide reading test. *Educational Assessment, 11*(3), 85-104.