ABSTRACT

GROCE, MARY KRISTINA. New Probes for Early Literacy Skills. (Under the direction of Ann Schulte.)

As educators alter their instructional decision making practices to align with a response to intervention (RTI) framework, it becomes crucial that appropriate tools for (a) identifying students at risk of reading failure and (b) monitoring students' responsiveness to intervention are utilized. The assessments currently used for these purposes, such as Curriculum-Based Measurement (CBM; Deno, 1986) and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002) can be time-consuming to administer to large groups of students. The present studies describe the initial evaluation of innovative, group-based progress monitoring measures. In Study 1, the three experimental measures (Reading Fluency, Maze Sentences, and Dolch Word Recognition) were administered to 73 first grade students four times during the academic year and the measures' reliability, validity, and ability to demonstrate students' growth over time were compared to those of two criterion measures, Word Identification Fluency (WIF; Fuchs, Fuchs, & Compton, 2004) and DIBELS Oral Reading Fluency (ORF). In Study 2, Reading Fluency, Maze Sentences, WIF, and ORF were administered weekly to four first grade students who were at risk for reading failure. In the multiple baseline design, the students received an intensive phonics intervention while their progress was monitored with the experimental and criterion measures. Results provide evidence that the Reading Fluency and Maze Sentences tasks are as reliable and valid as other measures in current use for screening, but suggest that they are not sensitive to

students' growth over time. Of the experimental and criterion measures, WIF was the only measure to demonstrate adequate ability to model students' growth. Thus, results suggest that the Reading Fluency and Maze Sentences tasks are promising benchmark/screening assessments within an RTI framework.

New Probes for Early Literacy Skills

by
Mary Kristina Groce

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the
Degree of
Doctor of Philosophy

Psychology

Raleigh, North Carolina

2009

APPROVED BY:

_____	_____
Dr. William Erchul	Dr. John Begeny
_____	_____
Dr. Susan Osborne	Dr. Ann Schulte Chair of Advisory Committee

DEDICATION

To my family, with greatest appreciation

for their boundless support, encouragement, and love.

BIOGRAPHY

Mary Kristina Groce, daughter of Dale and Terrie Groce, was born on November 2, 1978 in Asheville, NC. She attended elementary and middle school at Asheville Catholic School and completed her secondary education at Asheville High School, graduating with honors in 1997.

Kristina graduated *magna cum laude* with a Bachelor of Science in Psychology from Furman University in 2001. She then entered Wake Forest University and completed a Master of Arts in Psychology in 2003. In 2004, she entered the Graduate Program in Psychology at North Carolina State University, with a concentration in School Psychology. Kristina completed a pre-doctoral internship in Psychology with the Virginia Beach City Public Schools and received her Doctorate in Psychology from NC State University in May of 2009.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Ann Schulte, for her support, guidance, and supervision throughout my doctoral training, as well as Drs. William Erchul, John Begeny, and Susan Osborne for their time and feedback reviewing this document. I would also like to thank Jo Naglich, Megan Bennett, Rachel Nice, Anna Anthony, Danielle DeFeo, and Fleming Harris for their assistance with data collection and tutoring for the present study. Finally, I extend special thanks to my fiancé, Dr. Benjamin Brown, for his support and encouragement during the preparation of this document.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Prompted by statistics suggesting that students in the United States are struggling to learn to read, methods of improving children's literacy skills have become a topic of national focus (e.g., Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Juel, 1994;National Center for Education Statistics, 2007). Empirical evidence suggests that early intervention is effective (e.g., Torgesen, 2002), but without the proper methods of identifying students who are at risk and making decisions about their instruction, interventions may not be applied effectively.

Widely used methods of instructional decision making are based on the diagnostic-prescriptive approach, in which students are placed into instructional programs based on the results of pre-intervention assessments (Ysseldyke & Sabatino, 1973). The diagnostic-prescriptive approach relies on the assumption that aptitude-treatment interactions (ATI) exist; under this assumption, instruction should be adapted to groups of students who have similar abilities (Snow, 1989). One example of the diagnostic-prescriptive model is the notion that students have particular learning styles and that instruction should be tailored to these styles. Research, however, has failed to validate the ATI methodology (Arter & Jenkins, 1977).

An additional example of the diagnostic-prescriptive model in the schools is the application of ATI methodology to the identification of students with learning disabilities.

The current definition of specific learning disability uses a discrepancy formula, in which a student's achievement test scores must be lower than his or her IQ scores (e.g., Lyon et al., 2001). The discrepancy method of identifying students with learning disabilities assumes that a student's IQ score (aptitude) is necessary for placement into appropriate instructional programs. Contrary to the assumption that IQ is necessary for placement, research suggests that there are numerous problems with the discrepancy methodology, including findings that there are no differences between poor readers who do and do not have an IQ-achievement discrepancy, and that IQ scores are not relevant to diagnosing learning disabilities (e.g., Fuchs & Young, 2006; Stanovich & Siegel, 1994).

As a result of these and similar findings, educators and researchers have called for reform in the area of instructional decision making. Their call has been answered in the form of response to intervention (RTI), a problem-solving approach in which schools provide specified levels or tiers of increasingly intensive interventions to students at risk of academic difficulties. Most students will respond to interventions that are more intensive than regular classroom instruction; those who do not (about 5-7%) will be evaluated for more specialized services (Compton, Fuchs, Fuchs, & Bryant, 2006; D. Fuchs, Fuchs, & Compton, 2004). Although several variations on the model have been proposed, most involve several tiers of increasingly intensive intervention; students' progress is monitored regularly and those who achieve at a lower level and/or slower rate than their classmates are moved to the next higher tier (e.g., Fuchs, Fuchs, & Speece, 2002). Therefore, the frequent assessment of student

progress, or formative evaluation, is a key component of RTI and will therefore be explored in the present study.

In formative evaluation, assessment of student progress occurs throughout an instructional period in order to monitor instruction effectiveness. Formative evaluation in general has been shown to have a positive impact on student outcomes as well as instructional variables and, in contrast to diagnostic-prescriptive techniques, has been praised for its flexibility and responsiveness to student needs (Deno, 1990; Fuchs, Deno, & Mirkin, 1984). One type of formative evaluation is ongoing progress monitoring, in which assessment occurs at least three times per year using alternate test forms (Kame'enui, 2002). Similar to formative evaluation, progress monitoring is supported by empirical research (Fuchs, 1989; Hoffman & Rutherford, 1984).

As schools and school systems move towards the RTI approach to instructional decision making, methods of progress monitoring are being implemented in educational programming (e.g., Coyne, Kame'enui, & Simmons, 2004). However, as progress monitoring becomes more popular, it also becomes more important that the appropriate assessments be utilized. To help educators and psychologists select appropriate assessments, researchers have begun to set forth standards for assessments used as progress monitoring tools. For example, progress monitoring tools should demonstrate adequate psychometric properties, and should be available in multiple alternate forms, inexpensive, and easy for instructors to administer (Fuchs & Fuchs, 1999; Kame'enui, 2002; Shinn & Bamonto, 1998).

A number of assessments have been utilized for the purpose of progress monitoring; two that will be discussed in the current paper are curriculum-based measurement (CBM) and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). As will be described, both of these tools have evidenced moderate to strong reliability and validity (e.g., Fuchs & Fuchs, 1999; Hintze, Ryan, & Stoner, 2003). However, CBM reading tasks generally are not appropriate for the assessment of early (K-1) literacy skills, as they require the ability to read at least short paragraphs (Fuchs & Fuchs, 1992). DIBELS measures are tailored towards younger students, yet they require individual administration which can become time consuming when the one-minute assessments are conducted with every child in a classroom. If assessments similar to CBM and DIBELS were computerized, then educators could quickly and easily assess groups of students in the same amount of time it would take to administer the assessment to one child.

Computerized assessments of academic skills are, indeed, becoming more prevalent and some evidence exists supporting the validity of such assessments (e.g., Freeze, 1988; Hargreaves, Shorrocks-Taylor, Swinnerton, Tait, & Threlfall, 2004). One particular program, the Discourse Group-Ware Classroom, is a software program that links students' computers to the teacher's computer; as students complete activities on their own computers, their responses can be monitored on the teacher's computer. Students' responses can also be saved and automatically scored (Shin, 2000). The present study utilized Discourse technology as a means of group-level screening for early reading difficulties.

The present study examined the reliability and validity of three new experimental assessments as progress monitoring measures of first graders' reading skills. The three assessments were: Dolch Word Recognition, Maze Sentences, and Reading Fluency. Each of the assessments runs on the Discourse software and therefore can be administered in a group format. The project was comprised of two studies, with Study 2 being contingent upon the results of Study 1. In Study 1, first grade students were assessed on the experimental measures four times evenly spaced throughout the academic year. As a criterion measure, all participants were also assessed using Word Identification Fluency (WIF; administered at all four assessments) and Oral Reading Fluency (ORF; administered at the latter three assessments) within two weeks of measurement on the experimental measures. Results of Study 1 suggested that two of the experimental measures, Reading Fluency and Maze Sentences, had adequate psychometric properties; therefore the project proceeded to include Study 2, an examination of the extent to which these two measures demonstrated individual students' growth. Study 2 utilized a single-subject multiple baseline design. A subset of four first grade students who had been identified as at-risk for reading failure received 7 to 11 weeks of the Sound Partners reading intervention (Vadasy et al., 2005). These students' reading skills were monitored weekly throughout the intervention with the criterion measures and the two experimental measures that demonstrated adequate reliability and validity at the beginning of Study 1. Results of the two studies, described and interpreted in Chapters 5 and 6, add to the literature describing reliable and valid group-based progress monitoring measures of early literacy skills.

CHAPTER 2

A Review of the Literature

The following chapter reviews literature relevant to the development of a group-based progress monitoring measure of early literacy skills. In light of evidence that children in the United States are struggling to learn to read, it will first be suggested that this reading failure is due, in part, to poor instructional decision making practices. The prevailing method of instructional decision making, the diagnostic-prescriptive approach, will be described, and evidence of the strengths and weaknesses of the diagnostic-prescriptive approach will be presented and contrasted with a new approach, response to intervention (RTI). Next, formative evaluation will be highlighted as a key component of RTI, and one type of formative evaluation, progress monitoring, will be described. Following that discussion, standards for progress monitoring measures will be presented and currently available tools for monitoring first grade literacy skills will be described and compared against those standards. Finally, the validity of computerized assessments of early elementary aged children will be discussed.

Children Are Struggling to Learn to Read

The implementation of interventions for improving early literacy skills has become a national priority, propelled by statistics suggesting that many children in the United States are failing to learn to read. For instance, on the 2007 National Assessment of Educational Progress (NAEP), 67% of fourth graders achieved a score at or above "basic" in reading, leaving 33% achieving at the "below basic" level (National Center for Education Statistics,

2007). Additionally, research suggests that students who struggle to learn to read tend to continue to struggle. Juel (1994) found that children who have poor reading skills at the end of the first grade have an 88% chance of still having poor reading skills in the fourth grade. Furthermore, longitudinal studies suggest that the majority of students who have poor reading skills in the elementary grades continue to be classified as poor readers throughout adolescence and adulthood (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Shaywitz et al., 1999).

The Importance of Early Intervention

The statistics presented above paint a dismal picture. However, research on early literacy suggests that improved early instruction and intervention can improve children's literacy skills. For instance, a large body of evidence suggests that explicit, intensive, supportive instruction is effective in increasing students' reading skills (Torgesen, 2002). King and Torgesen (2001) studied a school that implemented preventive reading instruction for students in the first and second grades who were achieving below grade level. After five years of implementation, the proportion of first graders achieving below the 25th percentile on the Word Attack and Word Identification subtests of the Woodcock Johnson Tests of Achievement, Revised (WJ-R) fell from 31.8% to 3.7%; the proportion of second graders scoring below the 25th percentile fell from 14.5% to 2.4% over four years.

Foorman, Francis, Fletcher, Schatschneider, and Mehta (1998) provided additional evidence suggesting that appropriate classroom instruction can reduce the incidence of reading failure among students in the first and second grades. In their study of three early

intervention programs, 285 students were provided either direct instruction in letter-sound correspondence, less direct instruction in sound-spelling patterns, or implicit instruction in the alphabetic code. Results showed faster gains in word reading and better word recognition skills for the students who received direct instruction in letter-sound correspondence. Furthermore, 16% of the students receiving direct instruction showed no significant growth in word reading, compared to 46% and 44% in the implicit instruction and less-direct instruction groups, respectively. Thus, the authors concluded that early intervention employing appropriate strategies can be effective in reducing reading failure.

When discussing the effectiveness of early intervention, it is important to recognize that, across studies, there generally are a few children who do not respond to high-quality treatments. "Treatment resisters" were reported by both King and Torgesen (2001) and Foorman et al. (1998). To investigate the "treatment resister" phenomenon further, Torgesen (2000) summarized five studies of early interventions for reading and reported that each study consistently found small percentages (between 2% and 6%) of children who continue to struggle to read even after receiving treatment. However, Torgesen pointed out that more than 50% of struggling readers who received early intervention demonstrated improved reading skills; therefore, early intervention is effective for most, but not all, struggling readers. Thus, if we know that early intervention strategies are effective for most children, then why are over one-third of fourth graders struggling? The answer may, at least partially, lie within the limitations of prevailing methods of instructional decision making.

The Prevailing Methods of Instructional Decision Making

Knowing that early intervention can help students who are struggling to learn to read, it becomes essential that we be able to identify students in need of intervention and make appropriate decisions regarding their instruction. One goal of the present study is to identify new, more efficient, methods of instructional decision making. Before that can be discussed, however, it is necessary to review current methodology as well as a newer approach to decision making.

The current methods of identification and instructional decision-making for struggling readers are carried out via the diagnostic-prescriptive approach, which emphasizes assessing students for the purpose of classification and then placing them into instructional programs with other similarly-classified students (Ysseldyke & Sabatino, 1973). With the application of the diagnostic-prescriptive model to education, at least two assumptions are made. The first assumption is that aptitude-treatment interactions exist, and the second is that IQ is an important factor in identifying and instructing children with learning disabilities. Following brief descriptions of these assumptions, evidence of their validity will be reviewed.

Aptitude-Treatment Interactions

A fundamental assumption of the diagnostic-prescriptive approach is that individuals may be prescribed a particular treatment according to their particular classification or diagnosis. At the heart of this model is the assumption that aptitude-treatment interactions (ATI) exist. The ATI methodology became a focus of research for many psychologists in the

1950s and was formally introduced by Cronbach in his 1957 address to the American Psychological Association (Snow, 1989). Cronbach integrated correlational and experimental methodology in order to make generalizations about treatment effectiveness for groups of similar individuals. That is, he purported that individual differences identified via correlational research could be related to variation in treatment effectiveness identified in experimental research (Cronbach, 1957; Deno, 1990). Thus, instruction would be adapted to groups of students with similar abilities (Snow, 1989).

One example of the application of the ATI methodology in schools today is the assumption of learning styles, which is the notion that instruction should be tailored to students' unique cognitive abilities. For instance, it has been presumed that students for whom visual perception is a strength should receive whole-word or sight-word reading instruction (Arter & Jenkins, 1977). Another example is the prevailing way in which decisions about special education placement are made: students are assessed, classified based on the assessment results, and placed into educational programs based on their classification. Students with different intelligence and achievement test scores (and therefore, it is assumed, different aptitudes) are placed into different educational programs that are presumed to be able to meet the needs of groups of students with particular patterns of aptitudes (i.e., general versus special education; Reschly & Ysseldyke, 2002).

An Evaluation of the Evidence for the Diagnostic-Prescriptive Model and ATI

Critics of the diagnostic-prescriptive model cite the lack of empirical support for the model (Reschly & Ysseldyke, 2002; Ysseldyke & Sabatino, 1973). For example, in a classic

review study, Arter and Jenkins (1977) examined 13 studies in which beginning readers were classified by learning style (auditory or visual) and provided reading instruction tailored to their styles. Results showed that students were not differentially helped by instruction tailored to their strengths. Speece (1990) and Deno (1990) posed additional arguments against the model which focused on the model's methodology and assumptions.

Speece (1990) analyzed the components of ATI methodology, noting several problems with aptitudes, treatments, and the interaction of the two. These problems included the heterogeneity of treatment groups (e.g., children identified on the basis of IQ scores are likely to be quite variable in other domains); the inadequacy of assessment tools to measure individual-level, as opposed to group-level, differences; the assumption that aptitudes are not amenable to environmental influences; and the lack of attention paid to individual differences in terms of responsiveness to instruction.

Deno (1990) outlined three primary arguments against the ATI approach. First, he argued that the ATI approach places too much emphasis on testing students and results in a disconnect between assessment and intervention. That is, assessment is only used as a means to classify students in order to place them in particular programs, and the individuals who perform the assessments are generally not the same people who implement the educational programs.

Second, Deno (1990) stated that the ATI approach assumes that programs that are effective for a group of students as a whole will be effective for each individual within that group. However, Deno presented findings from the Minneapolis Public Schools' special

education program that suggest that the treatment with the greatest mean effect was not the most effective treatment for all students. That is, just because a treatment was most effective for the largest group of students did not mean that it was the best treatment for every student within that group. Therefore, Deno argued that evidence of effectiveness cannot always be generalized from the group to each member of the group.

Third, Deno (1990) argued that the ATI approach does not place enough emphasis on evaluating students' progress after they have been prescribed a program of instruction. He stated:

> Making predictions about individual student performance in the classroom is made extremely difficult because of the complex interaction of uncontrolled setting variables, teacher variables, method variables, and student characteristics. Thus, an approach to individualizing student performance that must rely heavily on initial diagnostic predictions is one certain to be inaccurate most of the time. Without a mechanism for adjusting programs in progress, they will most certainly fail (p. 166).

Taken together, critics of the diagnostic-prescriptive model and the ATI approach suggest that the diagnostic-prescriptive model for "individualizing" instruction may not be individualized enough. Speece (1990) argued for more attention to be paid to individual differences in response to treatment and in interactions between aptitudes and factors in the environment, and Deno (1990) argued for more emphasis on monitoring and evaluating students after they have been

placed in an instructional program. Reschly and Ysseldyke (2002) noted that Cronbach himself became frustrated with the ATI approach and suggested that applied psychology instead be based on a system of monitoring progress and adjusting treatment as necessary. These arguments foreshadow a new approach to instructional decision making, which will be a later topic of discussion.

ATI Methodology Applied to Instructional Decision Making

Despite these criticisms of the diagnostic-prescriptive model, ATI methodology has been the primary means of identifying students with learning disabilities from the beginning of federal special education legislation to the present day. Students who present with a discrepancy between IQ and achievement test scores are classified as having a specific learning disability in the area in which achievement scores fall below what would be expected given the student's IQ (e.g., Lyon et al., 2001). As a consequence, such students are placed into special education programs with other students who show a similar discrepancy; however, low-achieving students who do not meet the discrepancy criteria are excluded from special education services. Thus, the assumption that IQ (aptitude) is important in determining an appropriate treatment links the ATI methodology to our current methods of identifying students with learning disabilities.

The validity of the assumption that IQ is important in determining treatment, however, is a point of debate, particularly in terms of identifying students with learning disabilities. On one hand, researchers such as Kavale (2001) have argued that the IQ-achievement discrepancy should be part of a comprehensive evaluation for learning

disabilities; on the other hand, Stanovich (1999) argued that "a decade's worth of research has undermined the rationale for defining reading disability by reference to aptitude-achievement discrepancies" (p. 352).

Proponents of the discrepancy definition argue that there are significant differences between students who meet the discrepancy criterion and those who are low-achieving but do not meet the criterion (Kavale, Fuchs, & Scruggs, 1994). Kavale et al. analyzed data collected and previously analyzed by Ysseldyke, Algozzine, Shinn, and McGue (1982). Ysseldyke et al. had determined that there were not statistical differences between students classified as having learning disabilities (LD) and those who were low-achieving (LA), based on analyzing the two groups' score distributions and range; thus originally providing evidence *against* the IQ/achievement discrepancy definition. However, Kavale et al. analyzed Ysseldyke et al.'s inferential statistics (t-tests) and determined that students classified as LD had significantly lower achievement scores than students in the LA group. Additionally, via effect size analyses, Kavale et al. determined that, on average, 68% of the LD group could be distinguished from the LA group based on the students' achievement test scores suggesting that the IQ/achievement discrepancy model identified a more impaired group of children. Thus, the Kavale et al. concluded that the IQ-achievement discrepancy *is* a valid method of identifying students with LD and should, therefore, be included as part of a comprehensive LD evaluation.

Despite Kavale, Fuchs, and Scrugg's (1994) analyses, the discrepancy definition has continued to face quite a bit of criticism. Fuchs, Mock, Morgan, and Young (2003) described

the “fall” of the intelligence-achievement discrepancy formula as beginning with arguments that there are inconsistencies in the operationalization of the formula. Established in 1975, PL 94-142 suggested that such a formula be used to define LD, but the law did not specify how to calculate it nor did it specify a requisite magnitude. Therefore, different states developed their own definitions of learning disability, resulting in inconsistent prevalence rates between states. That is, one state may require a difference of one standard deviation between standard scores (i.e., 15 standard score points), whereas another state may require two standard deviations (i.e., 30 standard score points), and a third state may use a regression equation to predict achievement based on intelligence scores. Fuchs and colleagues assert that these inconsistencies contributed to discontent with the definition, and growing research suggesting that the definition is not valid has added additional fuel to the argument.

A second argument against the discrepancy formula is based on additional research suggesting that there are no relevant differences between poor readers with low scores on measures of intelligence (i.e., poor readers without the discrepancy) and those with higher scores (i.e., poor readers with the discrepancy). Despite Kavale, Fuchs, and Scrugg’s (1994) analyses of Ysseldyke, Algozzine, Shinn, and McGue’s (1982) data, more research suggests that these two groups of students do not differ in terms of their performance on reading tasks. For instance, Stanovich and Siegel (1994) compared children with and without a discrepancy on tasks of phonological, orthographic, and language processing skills as well as memory, and their regression analyses showed no differences between the two groups of children. Furthermore, phonological awareness, knowledge of letter names and letter sounds, and

naming speed at kindergarten have been shown to be the best predictors of reading achievement in grades 1 and 2 (Schatschneider, Fletcher, Francis, Carlson, & Foorman, 2004).

Research also suggests that deficits in phonological processing skills predict future reading difficulties (Torgesen, 2002; Torgesen, Wagner, & Rashotte, 1994). Discrepant and nondiscrepant poor readers demonstrate such deficits (e.g., Fuchs, Mock, Morgan, & Young, 2003; Stanovich, 1999); however, these phonological deficits do not differ as a function of level of intelligence. For example, Steubing et al. (2002) conducted a meta-analysis of 46 studies comparing poor readers who did and did not display an intelligence-achievement discrepancy. Their analyses revealed no significant differences between the two groups in terms of several foundational reading skills, including phonological awareness, rapid automatic naming, verbal short-term memory, and vocabulary/lexical skills. Taken together, these studies suggest that the distinction between discrepant and nondiscrepant poor readers that results from use of the discrepancy formula is not meaningful.

A third argument that has been posed against the discrepancy formula stems from the formula's emphasis on scores on measures of intelligence. Users of the discrepancy formula rely on the assumption that intelligence scores predict children's potential or ability to learn. However, research suggests that intelligence scores provide an estimate of current cognitive functioning, not learning potential, and therefore are not relevant to making a diagnosis of a learning disability (Fuchs & Young, 2006; Lyon & Fletcher, 2001). Furthermore, Stage, Abbott, Jenkins, and Berninger (2003) demonstrated that Verbal IQ was not as good of a

predictor of responsiveness to reading instruction as were phonological skills, orthographic skills, and rapid naming. Additionally, they found that a discrepancy between Verbal IQ and scores on a word reading task did not predict responsiveness to instruction.

A fourth argument against the current model is that, in practice, it becomes a "wait to fail" model, as children often do not evidence a discrepancy between IQ and achievement until later in elementary school. Steubing et al. (2002) illustrated this dilemma by describing a child who began struggling to read in the first grade. Her teacher provided individualized help for her and she progressed, but at a rate slower than that of her classmates. The following year, her second grade teacher noticed she was behind her classmates in reading and she was referred for a special education evaluation. However, the child's intelligence and achievement scores were not discrepant enough to qualify the child for special education services. By the time the child had reached the fourth grade, her reading achievement had not improved substantially but her norm-referenced achievement score had decreased because the achievement of her peers had grown at a faster rate. Therefore, her intelligence and achievement scores finally demonstrated the necessary discrepancy and she became eligible for special education services. Thus, the child's teachers had to wait for her achievement to drop far enough below that of her peers before she could be helped in special education.

Summary

Empirical evidence suggests that early intervention is crucial to helping children learn to read, yet using the current methods of instructional decision making, children like the one described above are excluded from special education until they have struggled academically

for several years. There is little empirical evidence to support the diagnostic-prescriptive model and the ATI methodology, and evidence validating the use of IQ scores in identifying and determining treatments for children with learning disabilities is controversial. Furthermore, research suggests that word-level reading skills and phonological processing skills are important to the early identification of reading problems (Schatschneider et al., 2004; Torgesen, 2002). Therefore, educators have searched for alternative approaches to instructional decision making that would allow for the early identification of children with reading skills deficits, regardless of IQ level. One such alternative is the response to intervention (RTI) model, which, as will be discussed, utilizes formative evaluation techniques similar to those employed in the present study.

Responsiveness to Intervention

The reauthorization of the Individuals with Disabilities Education Improvement Act (IDEIA; 2004) provided states with the option of using an alternative approach to identifying students with learning disabilities. Instead of relying on the IQ-achievement discrepancy, states may now use a response to intervention (RTI) approach. RTI is an approach based on problem-solving in which schools provide validated interventions to students who are at risk of academic difficulties. Those children who do not respond to the interventions are presumed to have deficits or disabilities that require more specialized services (Compton, Fuchs, Fuchs, & Bryant, 2006; D. Fuchs, Fuchs, & Compton, 2004). Thus, the RTI model requires that educators emphasize early intervention, monitor students' progress, and alter interventions if students are not progressing (Vaughn, Linan-Thompson, & Hickman, 2003).

In this respect, it is considered to be a "treatment-valid" approach, such that it can "inform, foster, and document the necessity for and effectiveness of special treatment" (Fuchs, Fuchs, & Speece, 2002, p. 34).

A number of models of the RTI approach have been proposed. Models of RTI typically involve several tiers or phases, with increasingly intensive interventions delivered at each tier. For instance, one model is comprised of three tiers. Under Tier 1, all students' progress is monitored within the general education classroom. Students whose achievement level or rate of growth is "dramatically lower" than their peers are identified as being at-risk and in need of more intensive intervention. Thus, they move to Tier 2, where they receive small group instruction and continued progress monitoring. Most children should benefit from the more intensive instruction; the ones who do respond return to their regular classroom instruction, and the ones who do not move on to Tier 3. Tier 3 is what is currently considered as special education; students in Tier 3 are likely to be identified as having learning disabilities and require more specialized and intensive intervention than they received at Tier 2 (Compton et al., 2006). Vaughn et al. (2003) suggested that these students are likely to be the same 5-7% of students who, as prior research indicates, continue to struggle despite explicit and intensive intervention. After reaching Tier 3, these students may be classified as having a learning disability (Fuchs et al., 2003).

Another version of RTI has been proposed by Fuchs et al. (2002) and is a four-phase treatment validity approach to assessment. In Phase I, all students are assessed to determine the overall rate of learning within a classroom. If the average growth in the classroom is low

(compared to the school, district, or nation), then classroom-wide interventions may be in order. In Phase II children who display dual discrepancies (i.e., performing at a lower level and improving at a slower rate then their classmates) are identified. During Phase III, students who demonstrate a dual discrepancy receive some type of change in instruction or intervention, and their responsiveness to that intervention is monitored. Children who do not respond in the general education classroom move to Phase IV, during which special education services are implemented on a trial basis and the students' performance is monitored. Under the authors' recommendation, Phase IV assessments would be used to inform the proper placement for those children who do not respond to these special services.

Support for RTI

Many educators, researchers, and policy-makers have embraced RTI for the improvements it makes over the traditional diagnostic-prescriptive model. Indeed, the Office of Special Education Programs (OSEP) has recently held a series of workgroups, symposia, and LD Summits that concluded in the call for the replacement of traditional methods of identifying LD in favor of RTI. Proponents of RTI suggest that the new model will aid in the early identification of learning disabilities, thereby helping educators to avoid the "wait to fail" model. They also argue that RTI will facilitate the instructional decision-making process and provide an emphasis on effective instruction, both of which are lacking in the traditional model (Compton et al., 2006; Fletcher, Coulter, Reschly, & Vaughn, 2004).

The arguments in favor of RTI are based in a growing body of research evidence. Although a single model of RTI has not yet been agreed-upon, research supports several

models and components of RTI. One model was implemented by Vaughn et al. (2003), who provided up to 30 weeks of small group tutoring to 45 second grade students who were at risk for reading problems. All of the students had been nominated by their teachers, failed a screening measure from the Texas Primary Reading Inventory (TPRI), and were not already receiving supplemental reading instruction. Baseline data regarding students' reading skills were collected using the Word Attack and Passage Comprehension subtests of the Woodcock Reading Mastery Test Revised (WRMT-R), the Phonological Awareness and Rapid Naming composites of the Comprehensive Test of Phonological Processing (CTOPP), and the Test of Oral Reading Fluency (TORF). After baseline data collection, students received 10 weeks of the small group tutoring intervention and then were assessed again using the TORF. Students meeting a pre-established criterion on the TORF were exited from the intervention and labeled "early exit" (n = 10), but they continued to be assessed for the remainder of the study. Students who did not meet the TORF criterion were reassigned to groups based on skill level and received 10 more weeks of the intervention. Students who met exit criteria at the end of the second 10 weeks were labeled "mid exit" (n = 14), and those who met exit criteria at the end of a third 10-week period were labeled "late exit" (n = 10). Those who did not meet exit criteria at the end of the total 30 weeks were identified as "no exit" students (n = 11) and were considered to be in need of more specialized services.

Vaughn et al. (2003) drew two primary conclusions from the results of the study. First, results suggested that it is possible to identify a subgroup of students in need of additional educational support such as special education when criteria for success have been

pre-established and a fixed amount of intervention is provided. Second, their method was a valid indicator of students who had inadequate reading skills. The results showed significant differences between students who did and did not exit the program on pretest measures of rapid naming, fluency, and passage comprehension.

Speece and Case (2001) investigated the validity of a dual-discrepancy definition for identifying treatment nonresponders within an RTI approach. The dual-discrepancy definition specifies that students must be achieving below their peers both in level and rate of growth in order to be identified as needing further evaluation for special education services. The dual-discrepancy procedure requires the regular monitoring of all students' skills in order to gauge achievement levels and growth over time. Speece and Case identified 47 first and second grade students as dually-discrepant in reading based on a minimum of 10 oral reading fluency assessments that occurred throughout the school-year. Students whose level of achievement at the end of the year (based on the mean of the last two data points) and rate of growth across the year were one standard deviation below their class's mean level and growth were classified as dually discrepant. These students were compared to 17 children who displayed an IQ-reading achievement discrepancy (a difference of 1.5 or more standard errors of prediction) and to 28 children who were classified as low achievers, having received a score below 90 on the Woodcock Johnson-Revised Basic Reading Skills cluster. Comparisons of the three groups revealed that the dually discrepant children were younger, had poorer phonological awareness skills, and lower teacher ratings of academic competence than children in the other two groups. The authors also noted that the dually discrepant

group more closely approximated the racial distribution of the school than did the group of students with the IQ-achievement discrepancy. That is, the dual discrepancy formula resulted in the identification of fewer minority students. These results suggest that the dual discrepancy definition, involving the regular monitoring of students' progress and comparison of both progress (growth over time) and level of achievement to that of classmates, is a valid indicator of reading difficulties.

Summary

RTI is an approach to instructional decision making in which children who are at risk of academic failure are provided increasingly intensive levels of validated interventions. It is presumed that the majority of students will respond to intervention, and the 5-7% who do not are those in need of more specialized services and evaluation. Several variations on the RTI model have been proposed and received much praise and empirical support (e.g., Compton et al., 2006; Fletcher et al., 2004; Vaughn et al., 2003). Formative evaluation is one key element of RTI that has been the focus of much research; the present paper will now turn to a discussion of formative evaluation and then to one type of formative evaluation, ongoing progress monitoring. Progress monitoring will be discussed as it relates to the assessment of early literacy skills and, therefore, to the goals of the proposed study.

Formative Evaluation

A key component of RTI is formative evaluation, which is the assessment of student progress that occurs throughout the instructional period in order to monitor the effectiveness of instruction (Deno, 2002; Scriven, 1967). Because measuring responsiveness to

intervention is the defining characteristic of RTI, formative evaluation is at the heart of the approach. As previously described, RTI requires regular evaluation of student progress; progress is then compared to previously-stipulated criterion and a decision is made regarding whether the child is responding appropriately to instruction or whether intervention needs to be changed or intensified (D. Fuchs et al., 2004).

Deno (1986) described formative evaluation as the "quality control mechanism" that provides feedback to educators about the effectiveness of instructional programs or interventions (p. 372). In the formative evaluation approach, interventions are viewed as applied research; educators and school psychologists engage in hypothesis-testing as they modify a student's IEP. Thus, the long term goal of formative evaluation is for the data collection and hypothesis testing process to result in an improvement in the student's rate of growth.

Researchers have called for formative evaluation as part of an alternative to the diagnostic-prescriptive approach. More than two decades ago, Deno (1986) maintained that data gathered within a formative evaluation approach are necessary for documenting the effectiveness of instructional programs. Additionally, Black and Wiliam (1998) also called for formative evaluation to become a regular component in the classroom. They asserted that it is necessary for teaching and learning to be interactive, such that teachers can identify students' needs and adapt instruction to meet those needs. Recent models of RTI, such as that investigated by Speece and Case (2001), have incorporated formative evaluation as a key element. Clearly, formative evaluation is essential to RTI; a model that calls for the regular

monitoring of student progress by definition calls for formative evaluation. The information regarding a student's level of achievement and rate of growth over time that is learned via formative evaluation would then be used to inform intervention. That is, if the student is making good progress compared to his or her classmates, current instructional methods may be continued. However, if the student is showing signs of struggling or is not growing at a rate comparable to his or her classmates, changes may be made to instruction.

Researchers investigating formative evaluation have suggested several advantages it offers over the traditional diagnostic-prescriptive approach. For example, Deno (1990) suggested that formative evaluation allows educators to be more flexible and responsive to students than does the diagnostic-prescriptive approach, emphasizes regular feedback regarding the effectiveness of instructional programs, takes the focus off of the predictive accuracy of pre-intervention testing, and recognizes the uniqueness of individuals. Empirical studies of formative evaluation provide additional support for its utility. For instance, Fuchs, Fuchs, Hamlett, and Ferguson (1992) randomly assigned 33 special education teachers and 63 students with mild to moderate disabilities to three groups: two groups participated in progress monitoring using curriculum based measurement (CBM; one group with expert system consultation regarding instructional decision-making and one without), and the third group did not utilize CBM. Teachers in both CBM groups assessed their students using the CBM-Maze task twice weekly for 17 weeks in order to make instructional decisions; teachers in the control group utilized criterion-referenced tests, daily work grades, unsystematic observation of performance, and teacher-made tests in order to make decisions about

instruction. Results indicated that students in both CBM groups demonstrated higher achievement on reading fluency and comprehension than did students in the control condition.

In a similar study, Fuchs, Deno, and Mirkin (1984) contrasted formative evaluation and typical special education over a period of 18 weeks. Eighteen special education teachers randomly assigned to the formative evaluation group each selected three to four students and measured those students' oral reading fluency twice weekly. The teachers also graphed the data and were instructed to alter instruction when a student's improvement over 7 to 10 data points was inadequate. Teachers randomly assigned to the special education group (n = 21) relied upon their typical methods of establishing IEP goals and monitoring progress, using teacher-made tests, informal observation, and workbook exercises. The results of the study indicated that, at post-treatment, students in the formative evaluation group achieved at a higher level on measures of reading fluency, decoding, and comprehension. Additionally, teachers using formative evaluation increased the amount of structure in instruction, whereas teachers in the special education group decreased their structure. Teachers in the formative evaluation group were also more realistic about and responsive to student progress than were those in the special education group. Finally, students whose teachers used formative evaluation were more knowledgeable about their own learning and goals than those who were in the special education group.

Fuchs and Fuchs (1986) conducted a meta-analysis of investigations evaluating the outcomes of formative evaluation. They excluded studies lacking a control group and those

that evaluated nonacademic behaviors, focused on behavior modification, provided feedback only to students, and involved college-aged students as subjects. The authors operationalized formative evaluation as data collection that occurred at least twice weekly with instructional decisions made on an individual, rather than a group, basis. Their final sample was comprised of 21 studies yielding 96 effect sizes. Their primary finding was a large effect (r = .70) of formative evaluation on increasing mildly handicapped students' achievement. The authors also found that effect sizes were larger when teachers used data-utilization rules to analyze student performance and make instructional changes as warranted. Also, effect sizes were larger when student data were graphed as opposed to recorded, when students were measured twice per week as opposed to three times per week or daily, and when treatment duration was more than 10 weeks.

These early studies utilizing formative evaluation for the purposes of screening and intervention have been the basis for current special education policy. As previously mentioned, the 2004 reauthorization of IDEIA allows states to use alternative approaches to identifying students with learning disabilities; the alternative approaches include methods of using formative evaluation to monitor students' progress during interventions.

Ongoing Progress Monitoring

One type of formative evaluation is ongoing progress monitoring, defined by Kame'enui (2002) as:

> Assessment conducted a minimum of three times a year or on a routine basis (i.e., weekly, monthly, or quarterly) using comparable and multiple test forms

> to (a) estimate rates of reading improvement, (b) identify children who are not demonstrating adequate progress and therefore require additional or different forms of instruction, and/or (c) compare the efficacy of different forms of instruction for struggling readers and thereby design more effective, individualized instructional programs for at-risk learners (p.25).

Progress monitoring can and has been implemented as frequently as twice weekly (e.g., Fuchs et al., 1984); however, as the definition above indicates, high-frequency measurement is not a requirement. Frequent progress monitoring is appropriate when educators are interested in evaluating the effectiveness of interventions so that alterations may be made as soon as a child shows signs of inadequate progress. Within a RTI approach, progress monitoring may occur beginning at a second tier/phase of the model, when students are receiving more intensive instruction or interventions. However, progress monitoring can also be used as a universal screening. As such, assessments of all students would occur three to four times per year in order to identify those children who may be at risk of academic difficulties or failure. Within RTI, universal screening occurs within the first tier/phase in the general education classroom. In the proposed study, progress monitoring will be utilized as universal screening to identify first graders who need additional reading instruction.

Fuchs (1989) stated that "Ongoing, systematic pupil progress monitoring, in general, is associated strongly with effective general and special education practice" (p. 156). A

growing body of empirical evidence exists which supports Fuchs' statement. For instance, Hoffman and Rutherford (1984) reviewed eight studies of schools that had been found to produce high student achievement and identified progress monitoring of student performance as a key ingredient in effective schooling. According to Fuchs (1989), these findings provide the basis of the empirical rationale for educational decision makers to employ formative evaluation as a method of evaluating intervention effectiveness.

As a result of this empirical support, progress monitoring has been incorporated into innovative school programs. For instance, Coyne, Kame'enui, and Simmons (2004) proposed a model of school-wide assessment and intervention in which progress monitoring is a key component. They noted several advantages of progress monitoring, including the ability to inform educators about students' progress toward IEP goals and the effectiveness of interventions, as well as the ability to compare a student's progress to that of his or her classmates and to benchmark goals. These comparisons are important, the authors explained, because they provide information on the amount of absolute progress students are making toward goals, as well as whether or not students are falling behind their classmates.

Summary

Formative evaluation, the regular assessment of student progress throughout an instructional period that is used to monitor the effectiveness of instruction, is key to RTI. Among the advantages of formative evaluation are improved student achievement and an increase in teacher structure and responsiveness (Fuchs et al., 1984; Fuchs & Fuchs, 1986). Ongoing progress monitoring is one type of formative evaluation that has also received much

support from researchers as it relates to both general and special education practices (Fuchs, 1989). Progress monitoring has become a component of innovative educational programs, such as that described by Coyne et al. (2004), and will be implemented in the proposed study as a universal screening to identify students struggling with early literacy skills.

A number of assessments are available for use as progress monitoring tools; however, not all of these assessments are equally appropriate for the task. In order to select the right tools for the job, it is important the school psychologists and educators be familiar with several characteristics that progress monitoring tools must possess. Therefore, researchers have begun to establish sets of standards by which assessments used for progress monitoring can be judged. A discussion of these standards appears below and is followed by a comparison of several tools currently used for progress monitoring to a core set of standards. Later, the tools utilized in the proposed study will be judged against this same set of standards with the aim of establishing them as valid progress monitoring tools.

Standards for Assessments Used in Progress Monitoring

With the call to implement progress monitoring in schools, many researchers have begun to generate standards for the assessment tools to be used in the RTI approach to student evaluation. No single set of criteria has, as yet, been agreed upon; sets of standards have been put forth by The National Center on Student Progress Monitoring (www.studentprogress.org), Kame'enui (2002), Shinn and Bamonto (1998), Fuchs and Fuchs (1999), and Fuchs et al., (2002).

The National Center on Student Progress Monitoring (NCSPM; www.studentprogress.org) has published a set of standards against which progress monitoring tools should be measured. Their seven standards are: Reliability, Validity, Alternate Forms, Sensitivity to Student Improvement, Adequate Yearly Progress Benchmarks, Improving Student Learning or Teacher Planning, and Rates of Improvement Specified. Beyond listing these standards, however, the authors do not specify additional details. For instance, they state that reliability and validity evidence must be reported (Cronbach's alpha, test-retest, and/or interrater reliability; and content, concurrent, predictive, and/or construct validity) and be "relevant and specific to the tool," but they provide no direction for researchers regarding what types or degree of reliability or validity would be adequate for a progress monitoring tool. Alternate forms of the tool must be of "equal and controlled difficulty" and enough forms must be available for monthly monitoring. The tool must also yield "data that are sensitive to children's development of academic competence and/or to the effects of intervention." In terms of establishing AYP benchmarks, the benchmarks for adequate end-of-year performance and the basis for identifying them must be specified and "evidence is based on specifically using this tool or is based on test construction principles from tools used in refereed studies." Additionally, empirical evidence of improved teacher planning and/or student outcomes must exist. Finally, a goal for adequate growth and the basis for measuring growth must be specified, and, as with AYP benchmarks, evidence of growth must be based on using the tool or on previously identified test construction principles.

Others have created their own sets of standards. For instance, in 2002, an eight member Assessment Committee headed by Edward J. Kame'enui, Ph.D. defined a set of criteria for selecting reading assessment tools and then used those criteria to review 29 reading assessments. The Committee based their criteria on the Texas Education Agency (TEA) Criteria for the Evaluation of English Early Reading Instruments. Some of their criteria for progress monitoring tools overlap with those of the NCSPM; however, they also include the following:

1. The assessment must be intended for use in grade(s) K, 1, 2, and/or 3.
2. The length of time needed to administer the assessment must be reasonable and depend upon the purpose of the measure.
3. The assessment instrument must assess early reading skills; instruments that assess book and print awareness alone are not sufficient.
4. Assessments should provide a global index of reading competence; it is not necessary that they provide separate scores for each skill assessed.
5. The assessment must be individually administered.
6. Classroom teachers should be allowed to administer the instruments.
7. Technical data for the instrument must be no more than 15 years old (Kame'enui, 2002, p. 26-29).

Shinn and Bamonto (1998) agreed with the NCSPM that a tool used as a progress monitoring measure should be sensitive to improvement. They also argued that such a measure should be standardized and logistically feasible. In terms of standardization, the

authors stated that the measure should be administered, scored, and interpreted consistently across examiners, so that changes in test scores may be attributed to changes in student performance. It must be feasible and efficient so that assessment does not override instructional time, and it also must be efficient in terms of decision making. That is, in order to achieve the most effect of formative evaluation, educators must be able to make quick decisions regarding instructional changes.

Fuchs and Fuchs (1999) also described several important considerations for selecting assessments for formative evaluation. Similar to other authors, Fuchs and Fuchs suggested that the assessment must have adequate psychometric properties, be sensitive to intraindividual change, be feasible, and have alternate forms. Additionally, they suggest that the tool must be able to model growth trajectories over time. In order to accomplish this, the tool must be based upon an interval scale and be free of floor and ceiling effects in the population for which it is used, have a low standard error of estimate so that growth can be more accurately estimated, and it must make use of time-series data in order to show change. Fuchs and Fuchs also suggest that the assessment tool must not be directly related to the intervention tool, which allows the one assessment tool to be useful in evaluating and comparing many interventions. However, another consideration is that the tool must be able to inform teaching. Thus, the assessment should not be coupled with any specific intervention, but results of the assessment should be able to guide future instruction.

An additional set of standards come from Fuchs et al. (2002), who considered progress monitoring within their treatment validity approach. They suggested that an

assessment tool must meet four primary technical requirements: (a) possesses the ability to model growth, (b) distinguishes between inadequate learning and an inadequate instructional environment (by comparing the rates of growth among children within a classroom), (c) informs the planning of instructional programs, and (d) has utility in evaluating intervention effectiveness.

Summary

Although details of these standards differ by author, the overarching themes that are apparent across standards are the emphases on (a) data than can be used to inform intervention; (b) efficiency in terms of time and cost; (c) sensitivity to individual change over time; and (d) reliability and validity. Given this understanding of the key characteristics of progress monitoring tools, the paper will now turn to a discussion of two primary tools currently being used to monitor the progress of children's academic skills development. These tools are curriculum-based measurement (CBM) and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). CBM and DIBELS are highlighted in the present paper due to the growing research base suggesting that they are valid progress monitoring measures. However, several other measures have been used for ongoing progress monitoring (e.g., Fuchs & Fuchs, 2004; Fuchs, Fuchs, & Compton 2004); the reader is directed to Appendix A, where these measures are compared against several of the above-stated standards for progress monitoring tools.

Curriculum-Based Measurement

Curriculum-based measurement (CBM) is a method of progress monitoring that was designed to assess skills in reading, spelling, writing, and mathematics. It consists of brief (less than 5-minute) probes with multiple alternate forms that test the same skills at the same level of difficulty. CBM allows educators to estimate a student's level of performance as well as rate of growth over time. These procedures were initially developed by Deno and colleagues to be reliable, valid, and practical measures of basic skills that can be used with any curriculum. According to Deno (1986), these measures become "curriculum-based when they are applied to a specific curriculum," but CBM probes can just as easily and effectively be based on "curriculum-free" materials (e.g., newspapers; p. 369). CBM procedures were originally designed to evaluate treatment effectiveness within the special education classroom, and have more recently been applied to measuring student growth in the general education classroom (Marston, Tindal, & Deno, 1984; Shinn, Shinn, Hamilton, & Clarke, 2002).

CBM differs from typical classroom assessments in many ways. First, in typical classroom instruction, teachers focus on single-skill mastery. Many CBM tasks, however, sample broadly from the annual curriculum, requiring students to integrate many skills in each assessment (Fuchs & Fuchs, 2002). Second, CBM is psychometrically-based, with a growing literature base documenting its reliability and validity. Sound psychometric properties help to ensure that data collected using CBM are accurate and meaningful. In contrast, informal observations, worksheets, and teacher-made tests typically have no

documented reliability and validity. A final difference between CBM and typical classroom-based assessment is that the focus of CBM is long-term, such that methods and content remain the same across repeated assessments, so progress can be monitored over time. As noted above, typical classroom assessments tend to focus on specific skills; once a skill is mastered, assessments change focus to the next skill. Therefore, measures over time are not equivalent, which limits the ability to measure individual students' growth (Fuchs et al., 2004).

Figure 1 provides an illustration of the utility of CBM using a case study of a student assessed with oral reading fluency (ORF) probes over a period of 12 weeks. The graphed dots represent scores on individual CBM probes. The trend lines show the child's rate of change; during the first 6 weeks of instruction, his performance was relatively stable over time. The vertical line represents a change the teacher made in instruction and the dotted line represents the teacher's goal for the child. After the first 6 weeks, the teacher determined that the child's trend-line was flatter than his goal-line, so she made an instructional change to the child's reading program. The child's performance improved after this change, evidenced by the increase in his number of words correct per minute. Therefore, by graphing the child's progress, the teacher judged his responsiveness to the change of instruction.

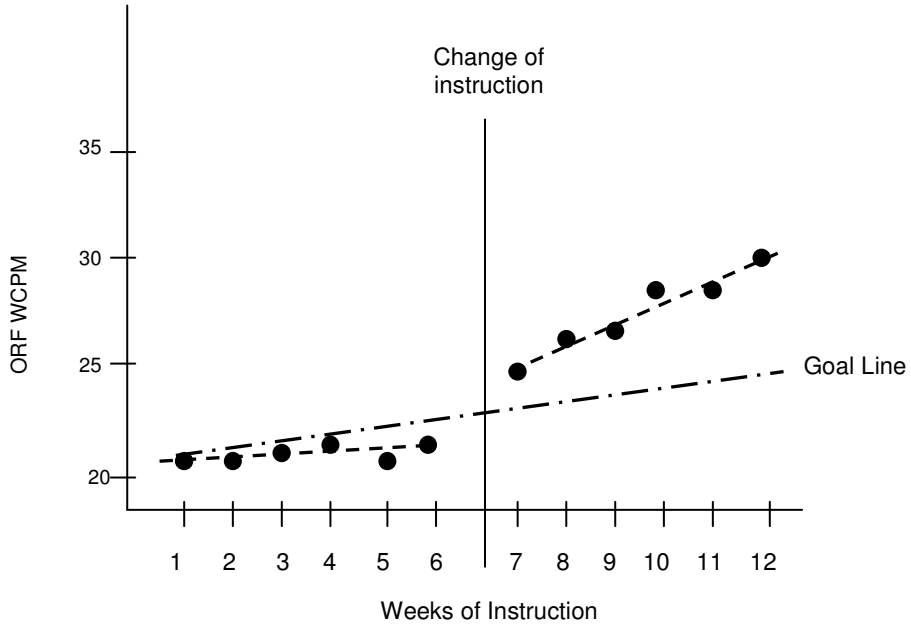


Figure 1. Illustration of graphed CBM data.

Fuchs and Fuchs (1999) critiqued CBM according to their criteria and concluded that CBM "provides a prototype" for a system of progress monitoring. In terms of the strengths of CBM, the authors cite evidence of strong test-retest and interscorer reliability and construct, discriminative, and criterion-related validity combined with the ability for repeated performance sampling which produces time-series data. Furthermore, CBM assessments are based on an interval scale and items presented throughout an instructional period (i.e., the school year) are of the same difficulty level. Because CBM is a progress monitoring tool, alternate forms of CBM probes are available. Also, the authors cite evidence that CBM assessments generally have an acceptably low standard error of estimate relative to slope, and that CBM slopes can differentiate between students. Finally, Fuchs, and Fuchs note that CBM is feasible and efficient and is able to guide teaching and intervention strategies.

CBM Reading Tasks

A number of CBM tasks have been used to measure reading skills, including recall procedures, cloze techniques, maze procedures, and oral reading fluency (ORF; Fuchs & Fuchs, 1992). ORF has been shown to be a reliable and valid assessment of children's reading skills, with test-retest reliability ranging from .93 to .96, and correlations ranging from .54 to .92 with the Woodcock Reading Mastery Tests, Word Identification, and Passage Comprehension Tests (Fuchs et al., 1984). Additionally, ORF has demonstrated the ability to model students' growth; for example, Fuchs, Fuchs, Hamlett, Walz, and Germann (1993) suggested that students in the second grade gain, on average, 1.5 words correct per week. Furthermore, Marston, Fuchs, and Deno (1986) suggested that students' growth on CBM passages was significantly positively correlated with teachers' judgments of student growth. The Maze task has also demonstrated sound psychometric qualities, with test-retest reliability coefficients ranging from .61 to .91 (Shin, Deno, & Espin, 2000) and concurrent validity with the Gates-McGinitie Reading Test (r = .65 to .76) and the Metropolitan Achievement Test of Reading (r = .66 to .76; Jenkins & Jewell, 1993). The Maze task has also demonstrated sensitivity to students' growth (Fuchs et al., 1992). However, these tasks require students to be able to read at least short paragraphs, and therefore are not appropriate for assessing early literacy skills. ORF, for example, "becomes appropriate for most students sometime during the second semester of first grade" (Fuchs et al., 2004, p. 7). The ORF task tends to produce a floor effect when used to measure the skills of very young students; thus, Fuchs et al. stated

that students' scores on the ORF task may indicate that these students have not made progress when, in reality, they have.

As previously noted, it is imperative that educators reliably assess kindergarten and first grade early literacy skills. Research suggests that students who struggle in these early years tend to continue to struggle, but that early intervention can be quite effective. Research also tells us that pre-reading skills, such as phonological awareness and alphabetic knowledge, are predictors of future reading success (Coyne & Harn, 2006). These early literacy skills provide the foundation upon which subsequent reading skills are built; thus, it is essential that we be able to measure their development with tools that will guide and inform intervention (Fuchs et al., 2004).

DIBELS

The DIBELS are a set of standardized, individually administered measures of early literacy skills. They are designed for regular monitoring of the development of pre-reading and early reading skills. There are four kindergarten measures and four first grade measures. The kindergarten measures are: Initial Sound Fluency (ISF), Letter Naming Fluency (LNF), Phoneme Segmentation Fluency (PSF), and Nonsense Word Fluency (NWF). LNF, PSF, and NWF are also first grade measures, in addition to Oral Reading Fluency (ORF). The administration time for each measure is one minute, with the exception of ISF, which requires three minutes (Good & Kaminski, 2002).

A number of studies have provided evidence for the validity of DIBELS kindergarten and first grade measures. For instance, Hintze, Ryan, and Stoner (2003) found moderate

correlations between the measures ISF, PSF, and LNF and the Comprehensive Test of Phonological Processing (CTOPP). Eighty-six kindergarten students were administered the three DIBELS tasks and the CTOPP within a 3-day time-span in early March. Scores generally showed the expected pattern of correlations; LNF was most strongly correlated with the CTOPP Rapid Naming Composite (r = .58), and ISF and PSF were most strongly correlated with the Phonological Awareness Composite (r = .60 and .53, respectively).

In another study, Elliott, Lee, and Tollefson (2001) created and validated a modified battery of DIBELS kindergarten measures (DIBELS-M). These measures were LNF, Sound Naming Fluency (SNF), Initial Phoneme Ability (IPA), Phoneme Segmentation Ability (PSA). IPA and PSA were drawn directly from the original DIBELS measures of ISF and PSF, respectively; IPA and PSA employed simpler stimulus words and allowed children additional time to respond. Thus, the authors perceived these measures to emphasize ability over fluency. Students were administered alternate forms of the DIBELS-M on two occasions, two- to four-weeks apart, at the end of the school year. Criterion measures were the Test of Phonological Awareness (TOPA), the Woodcock Johnson-Revised Broad Reading Skills Cluster (WJ-R), the Kauffman Brief Intelligence Test (K-BIT), the Developing Skills Checklist Pre-Reading Total Score (DSC), and the Teacher Rating Questionnaire (TRQ), on which teachers rated students' levels of pre-reading skills. Results yielded concurrent validity coefficients ranging from .60 to .70 and reliability coefficients ranging from .80 to the mid-.90s. Thus, Elliott et al. concluded that DIBELS are appropriate

tools for monitoring students' progress in reading and for identifying students who are at-risk.

Word Identification Fluency (WIF)

WIF is a progress-monitoring measure in which students are shown 50 randomly-selected high-frequency words from Dolch preprimer, primer, and first-grade level lists, and given one minute to read the words aloud. The examiner records the number of words read correctly as the child's score; if a child hesitates on a word for four seconds, he or she is prompted to go on to the next word. Fuchs et al. (2004) found correlations ranging from .52 to .93 between WIF and Woodcock Word Identification, Word Attack, Comprehensive Reading Assessment Battery (CRAB) Fluency, and CRAB comprehension. Fuchs et al. also demonstrated WIF's alternate forms reliability to be $r = .88$.

Despite the evidence of moderate to strong reliability and validity of CBM, DIBELS, and WIF, they are not ideal for progress monitoring for at least one reason: each of these measures is administered individually. It is notable that individual administration is listed among Kame'enui's (2002) criteria for appropriate progress monitoring tools, but individual administration conflicts with the necessary requirement that a tool be feasible for teachers to implement. For instance, a probe that takes one minute to administer to one student could take, at a minimum, 30 minutes to administer to a class. However, if a measure was to be computerized, a group of students in a computer lab could be assessed in a fraction of the time it would take to complete individual assessments. In addition, computerized administration of group-based measures could allow for computerized scoring, further

decreasing the load of tasks demanding teachers' time and effort. In light of these advantages, the goal of the present project is to develop group-based assessments for monitoring first grade students' early literacy skills. Thus, the current paper will now turn to a brief discussion of computerized assessments.

Computerized Assessments

The use of computers within schools brings new possibilities to instruction and assessment. Computers have been used for both standardized and informal assessments, and expert programs have been developed to interpret results (Freeze, 1988; Rickelman, Henk, & McKenna, 1991). Research findings generally suggest that computerized assessments are more time-efficient and more preferred by students than pencil-and-paper tests. However, there is less agreement regarding the validity of computerized assessments; the present studies will help to quell such disagreement. In one study, Hargreaves, Shorrocks-Taylor, Swinnerton, Tait, and Threlfall (2004) found that 10-year-old students given parallel forms of a mathematics assessment on the computer and on paper performed comparably on both assessments, with a trend towards performing better on the computerized assessment. However, when Varnhagen and Gerber (1984) administered written and computerized versions of the Test of Written Spelling to third graders, the students performed better on the written version. One possible explanation for these contrasting results is the extent to which adequate typing skills were required for success on the assessments. In the present project, typing skills were not necessary; students were required only to use the computer mouse to

select their answers from a list of choices and to advance from one item to the next. Practice items were administered until all students were comfortable with this skill.

The present project employed the Discourse Groupware Classroom software system in order to assess students' early literacy skills. With Discourse, all students' computers are networked to the teacher's computer; students respond to test items on their own computers and the teacher can monitor all students' responses from the teacher computer. Responses can be saved, scored automatically, and displayed on a video monitor. The Discourse software was previously utilized by Shin (2000) in a study to determine the validity of maze-based reading probes administered via Discourse in predicting students' end-of-year performance on the California Achievement Test (CAT).

Shin (2000) included 48 second-grade students in his study. Discourse had been used daily within the classroom during the school year; therefore, the students and their teachers were familiar with the Discourse system at the time of the study. Shin's variable of interest, active responding, was operationalized as the number of frames completed per minute over 12 reading and 16 mathematics assessments. Results suggested that active responding was highly correlated with end-of-year CAT scores and was a significant predictor of final performance when initial performance was controlled for.

In another study, Shin et al. (2000) explored the validity of the CBM-Maze reading task administered via Discourse. Results suggested that the computer-administered Maze task was reliable, with alternate-form reliability coefficients ranging from .69 to .91. The validity of the task was demonstrated by using hierarchical linear modeling (HLM) to examine the

relationship between maze scores and California Achievement Test (CAT) reading scores. Results suggested that students with higher CAT scores showed greater improvement on the maze task than students with CAT scores at or below the mean. Therefore, across these two studies, Shin and colleagues provided preliminary evidence for the validity of Discourse as an academic assessment tool; the proposed study will add to this knowledge.

Summary

In light of evidence that early intervention can be successful in helping young children learn to read (e.g., Torgesen, 2002), valid methods of identifying children at risk for reading failure and monitoring the development of their early reading skills are critical. CBM and DIBELS measures are tools commonly used for such identification and monitoring, but each can be improved upon. CBM probes, such as oral reading fluency and maze, are inappropriate for children whose reading skills are not developed to at least the mid-first-grade level. DIBELS measures are appropriate for children in kindergarten and the first grade, but they must be individually administered. Clearly, a group-based system for monitoring children's early literacy skills is warranted. Assessments such as those evaluated in the present project would allow educators to assess a classroom full of children at one time and would provide reliable and valid indicators of each child's progress.

CHAPTER 3

Research Aims

Statement of the Problem

As discussed in the previous chapter, current methods of instructional decision making (i.e., based on the diagnostic-prescriptive approach and aptitude-treatment interaction methodology) are insufficient for the early identification of children who are at-risk for reading failure (Arter & Jenkins, 1997; Fuchs & Young, 2006; Stanovich & Siegel, 1994). As a key ingredient in response-to-intervention (RTI), ongoing progress monitoring offers promise as a new method for not only identifying at-risk students, but also for monitoring their responsiveness to evidence-based interventions (Fuchs, 1989; Hoffman & Rutherford, 1984).

Curriculum-based measurement (CBM) and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) are two assessment tools that have established utility as progress monitoring tools (e.g., Fuchs & Fuchs, 1999; Hintze et al., 2003). However, CBM Oral Reading Fluency (ORF) and Maze tasks are best suited for children who have completed the first half of the first grade, as both tasks require children to read passages. The DIBELS measures are appropriate for children prior to the middle of first grade, but they (as well as ORF) must be individually administered. Although each probe is brief, individually assessing a classroom of children on a quarterly or more frequent basis is likely to become a strain on teachers' time and resources. Therefore, group-administered progress monitoring assessments that measure early literacy skills (i.e., skills learned prior to mid-way through the

first grade) are necessary in order to improve the identification of young children who are at risk for reading failure.

The present project was comprised of two studies. The primary purpose of Study 1 was to demonstrate adequate psychometric properties of three new, group-based progress monitoring measures of early literacy skills: Reading Fluency, Maze Sentences, and Dolch Word Recognition. These tasks were presented to students on computers via the Discourse Groupware Classroom, a software system that links the teacher's computer with all students' computers, such that students may respond to teacher-posed questions on their computers simultaneously and the teacher can monitor their responses. It was predicted that the three new measures would demonstrate adequate test-retest and alternate-forms reliability; concurrent validity with two established measures of early reading skills, Word Identification Fluency (WIF) and ORF; and adequate ability to demonstrate students' growth over time. These predictions were tested utilizing a correlational design that was based on promising results from a recent pilot project (See Appendix B).

It had been proposed that if the results of the first assessment of Study 1 indicated that any of the three experimental measures demonstrated adequate alternate forms reliability and concurrent validity with WIF, then Study 2 would be conducted in order to further investigate those particular measures. Study 2 was indeed conducted using two of the experimental measures, Reading Fluency and Maze Sentences; it employed a single subject design in order to determine whether the new progress monitoring tools demonstrated sensitivity to intraindividual change. A subset of students from Study 1 participated in a

reading tutoring program as their progress was monitored weekly using the experimental (Reading Fluency and Maze Sentences) and criterion (WIF and ORF) measures. The results of both studies add to the literature on early literacy skills, and it is hoped that these studies will be the first step in developing innovative tools for monitoring first grade literacy skills.

Research Questions and Hypotheses

Study 1

Question 1. Is the Reading Fluency task appropriate for the progress monitoring of first grade early literacy skills?

Hypothesis 1. The Reading Fluency task will be a reliable measure of first grade early literacy skills, as indicated by:

A. adequate alternate forms reliability, evidenced by the majority of the correlations between alternate forms of the Reading Fluency task (administered during the same assessment session) at or above $r = .80$; and

B. adequate test-retest reliability, evidenced by the majority of the correlations between administrations of the Reading Fluency task at or above $r = .80$.

Hypothesis 2. The Reading Fluency task will be a valid measure of first grade early literacy skills, in terms of its relation to other variables.

A. The Reading Fluency task will demonstrate adequate concurrent validity, indicated by:

(i) the majority of the correlations with ORF at or above $r = .80$, when Reading Fluency and ORF are administered within two weeks of each other, and

(ii) the majority of the correlations with WIF at or above $r = .80$, when Reading Fluency and WIF are administered within two weeks of each other.

B. The Reading Fluency task will demonstrate adequate predictive validity, indicated by:

(i) the majority of correlations between scores on the first six Reading Fluency probes and the fourth ORF assessment at or above $r = .80$, and

(ii) the majority of correlations between scores on the first six Reading Fluency probes and the fourth WIF assessment at or above $r = .80$.

Hypothesis 3. The Reading Fluency task will model students' growth over time as evidenced by a positive slope across the four data points that is significantly different from zero.

Question 2. Is the Maze Sentences task appropriate for the progress monitoring of first grade early literacy skills?

Hypothesis 4. The Maze Sentences task will be a reliable measure of first grade early literacy skills, as indicated by:

A. adequate alternate forms reliability, evidenced by the majority of the correlations between alternate forms of the Maze Sentences task (administered during the same assessment session) at or above $r = .80$; and

B. adequate test-retest reliability, evidenced by the majority of the correlations between administrations of the Maze Sentences task at or above $r = .80$.

Hypothesis 5. The Maze Sentences task will be a valid measure of first grade early literacy skills, in terms of its relation to other variables.

A. The Maze Sentences task will demonstrate adequate concurrent validity, indicated by:

(i) the majority of the correlations with ORF at or above $r = .80$, when Maze Sentences and ORF are administered within two weeks of each other, and

(ii) the majority of the correlations with WIF at or above $r = .80$, when Maze Sentences and WIF are administered within two weeks of each other.

B. The Maze Sentences task will demonstrate adequate predictive validity, indicated by:

(i) the majority of correlations between scores on the first six Maze Sentences probes and the fourth DIBELS ORF assessment at or above $r = .80$, and

(ii) the majority of correlations between scores on the first six Maze Sentences probes and the fourth WIF assessment at or above $r = .80$.

Hypothesis 6. The Maze Sentences task will model students' growth over time as evidenced by a positive slope across the four data points that is significantly different from zero.

Question 3. Is the Dolch Word Recognition task appropriate for the progress monitoring of first grade early literacy skills?

Hypothesis 7. The Dolch Word Recognition task will be a reliable measure of first grade early literacy skills, as indicated by:

A. adequate alternate forms reliability, evidenced by the majority of the correlations between alternate forms of the Dolch Word Recognition task (administered during the same assessment session) at or above $r = .80$; and

B. adequate test-retest reliability, evidenced by the majority of the correlations between administrations of the Dolch Word Recognition task at or above $r = .80$.

Hypothesis 8. The Dolch Word Recognition task will be a valid measure of first grade early literacy skills, in terms of its relation to other variables.

A. The Dolch Word Recognition task will demonstrate adequate concurrent validity, indicated by:

(i) the majority of correlations with ORF at or above $r = .80$, when Dolch Word Recognition and ORF are administered within two weeks of each other, and

(ii) the majority of correlations with WIF at or above $r = .80$, when Dolch Word Recognition and WIF are administered within two weeks of each other.

B. The Dolch Word Recognition task will demonstrate adequate predictive validity, indicated by

(i) the majority of correlations between scores on the first six Dolch Word Recognition probes and the fourth ORF assessment at or above $r = .80$, and

(ii) the majority of correlations between scores on the first six Dolch Word Recognition probes and the fourth WIF assessment at or above $r = .80$.

Hypothesis 9. The Dolch Word Recognition task will model students' growth over time as evidenced by a positive slope across the four data points that is statistically significant from zero.

The Reading Fluency, Maze Sentences, and Dolch Word Recognition tasks were expected to be appropriate progress monitoring tools for first graders' literacy skills based on their similarities to other progress monitoring tools. For instance, the items that comprise each of the three tasks are based on pre-primer, primer, and first-grade-level high frequency words, as is the WIF task. WIF has demonstrated alternate-forms reliability ranging from .88 to .97 and has been shown to be predictive of students' performance on decoding and reading comprehension tasks (Fuchs et al., 2004). The Dolch Word Recognition task in particular is similar to WIF in that it requires students to read high frequency words one word at a time

(i.e., not in the context of a sentence); however, it is not a measure of students' oral reading skills, but rather assesses a student's ability to recognize high frequency words (as indicated by the ability to match a word to the picture it describes).

The Maze Sentences task is modeled after the CBM-Maze task, which requires students to supply missing words within a passage. The Maze task has demonstrated test-retest reliability coefficients ranging from .61 to .91 (Shin et al., 2000). It has also been shown to have adequate concurrent validity with other reading measures, based on correlations with the Gates-McGinitie Reading Test ranging from .65 to .76 and with the Metropolitan Achievement Test of Reading ranging from .66 to .76 (Jenkins & Jewell, 1993). Maze Sentences differs from the Maze task, however, in its brevity and thus appropriateness for the measurement of early reading skills. That is, rather than being required to read a 150-word (or longer) passage, students were required only to read a series of brief (3- to 6-word) sentences.

Much like ORF, the Reading Fluency task has been designed to be a measure of students' abilities to read connected text. ORF has been shown to be reliable, with test-retest reliability ranging from .93 to .96, as well as valid, with correlations ranging from .54 to .92 with the Woodcock Reading Mastery Tests, Word Identification, and Passage Comprehension Tests (Fuchs, Deno, & Mirkin, 1984). Unlike ORF, the Reading Fluency task does not measure students' oral reading skills; instead, it requires students to read brief sentences to themselves and determine whether the answer is true or false.

Furthermore, it was expected that the three experimental measures would model students' growth over time because each of the measures, similar to existing progress monitoring measures, assess skills that students would be expected to improve upon throughout the course of the academic year. Existing progress monitoring measures, such as ORF, WIF, and Maze, measure reading fluency and comprehension and have been shown to be sensitive to students' growth (e.g., Fuchs & Fuchs, 2004; Fuchs et al., 2004; Marston et al., 1986). The experimental measures are similar to these existing measures in terms of item format and skills assessed, with the exception that they are group-based and target the early literacy skills of students in the first grade.

Although the experimental tasks are similar to existing progress monitoring measures, they are unique in that they are computerized and therefore ideal for group-based assessment. Additionally, the measures utilize pre-primer, primer, and first grade high frequency words, making them specifically targeted for students in the first grade. Therefore, it was a unique aim of the proposed study to demonstrate the psychometric properties of these three group-based assessments of early literacy skills.

Study 2

Question 1. Do the computerized early literacy assessments demonstrate sensitivity to intraindividual change in response to intervention?

Hypothesis 1A. During intervention, students' scores on the Reading Fluency task will show positive linear growth from baseline levels.

Hypothesis 1B. During intervention, students' scores on the Dolch Word Recognition task will show positive linear growth from baseline levels.

Hypothesis 1C. During intervention, students' scores on the Maze Sentences task will show positive linear growth from baseline levels.

There is growing evidence that curriculum-based measures index students' growth. For instance, Marston et al. (1986) demonstrated that oral word and passage reading were sensitive to students' growth over 10 and 16 weeks, respectively. Growth on third through sixth graders' high-frequency word reading fluency was indicated by a significant difference in mean number of words read correctly at week 1 (M = 46.85) and week 10 (M = 60.71). Similarly, third graders' growth on oral passage reading fluency was indicated by a significant difference in mean number of words read correctly during the first three weeks (M = 72.5) and last three weeks (M = 89.7) of assessment.

A number of single-case and small-n studies have investigated the ability of measures similar to the proposed experimental measures to model intraindividual growth over time for the purposes of evaluating interventions. For example, Stoner, Carey, Ikeda, and Shinn (1994) used oral reading fluency to show two students' responsiveness to varying dosages of methylphenidate. In both studies, the students' performance on daily oral reading fluency probes was utilized to inform decisions regarding optimal dosage. In other studies, Fiala and Sheridan (2003) and Stoner, Scarpati, Phaneuf, and Hintze (2002) demonstrated that ORF was useful in monitoring students' improvement throughout reading interventions. Although the proposed measures are in many ways different from ORF, it is expected that they will be

as useful as ORF in demonstrating students' improvements over time. Like ORF, the proposed measures will provide researchers and educators with brief checks on key early reading skills. If the proposed measures are sensitive to students' growth over time in response to intervention, then additional preliminary evidence for the validity of the measures will be demonstrated.

CHAPTER 4

Method

The present chapter describes the research methods used in Studies 1 and 2. For each study, the participants, measures, and procedures are described. Permission to conduct both studies was obtained from the Institutional Review Boards at North Carolina State University and the Wake County Public School System. Because the experimental measures were a part of the school curriculum, passive consent was used for collection of the experimental probes and active consent was used for the criterion probes and Study 2.

Study 1

Participants

Experimental measures. Participants in Study 1 were drawn from the three first grade classrooms at a small, Southeastern elementary school. The group of 73 first grade students was comprised of 40 males and 33 females, with a mean age of 6 years, 8 months (see Table 1). Although all of the 73 students participated in class-wide computerized progress monitoring using the experimental Discourse measures, a number of factors restricted the sample size at each assessment. First, students' absences during each class's assigned computer lab time resulted in smaller sample sizes. Second, some students who were present occasionally unintentionally logged themselves out of the Discourse system before completing assessment probes. Finally, technical difficulties made several probes unscoreable, resulting in reduced sample sizes (by approximately one-third) for the first two assessment periods on Reading Fluency and the first assessment period of Maze Sentences.

Thus, final Reading Fluency samples ranged from 35 to 64 students and Maze Sentences samples ranged from 42 to 63 students. For analyses of growth over time, slope was not calculated for any participant who had fewer than three data points. Therefore, growth analyses for Reading Fluency were based on a sample of 59 students (81% of participants) and growth analyses for Maze Sentences were based on a sample of 61 students (84% of participants). Reading Fluency data from all four assessment points were available for 43 students (59%), and Maze Sentences data from all four assessment points were available for 42 students (58%).

Criterion measures. Parental consent was obtained for 49 students (67%) from three classrooms; one female student was omitted from data analyses due to cerebral palsy, which limited her ability to participate without additional assistance that would have invalidated results. Of the 48 remaining participants, 28 were male and 21 were female and their ages ranged between 6 years, 2 months and 7 years, 11 months, with a mean age of 6 years, 8 months (see Table 1).

Table 1

Number, Gender, and Age of Participants by Classroom and Measure

		Classroom 1	Classroom 2	Classroom 3	Total
Experimental Measures	N	23	24	26	73
	Male	13	12	15	40
	Female	10	12	11	33
	Mean Age	6 yrs, 7 mos	6 yrs, 8 mos	6 yrs, 9 mos	6 yrs, 8 mos
Criterion Measures	N	12	19	17	48
	Male	8	9	10	27
	Female	4	10	7	21
	Mean Age	6 yrs, 7 mos	6 yrs, 8 mos	6 yrs, 8 mos	6 yrs, 8 mos

Experimental Measures

The experimental measures are administered via the Discourse Groupware Classroom (Shin, 2000). Discourse is a software system that allows for the authoring of several types of classroom tasks, such as multiple choice, fill-in-the-blank, and true/false questions. With Discourse, student computers are networked with the teacher's computer, allowing the teacher to monitor all students' responses at once. Under "social mode," the teacher controls when and which items in a lesson are presented to students; under "self-paced mode," each student can advance his or her screen to the next item in a lesson at his or her own pace. The three experimental tasks are administered under self-paced mode. Typing skills are not necessary for responding to tasks; students use the computer mouse to select their answers to all items. All experimental measures are scored by subtracting the total number of incorrect items from the total number of correct items.

Reading Fluency. The Reading Fluency task requires students to read brief (3-5 word) sentences and use the computer mouse to select the word "yes" if the statement is true or "no" if the statement is false. For each probe, students are allowed one minute to respond to as many items as they can.

Items for the Reading Fluency task were created using preprimer, primer, and first-grade level Dolch words, as well as Dolch nouns. A test bank of 100 items was developed and probes were created by randomly sampling 25 items, with replacement, from the test bank. The first five items of each probe of the Reading Fluency task were comprised of 3-word sentences to help assure a sensitive floor to the measure for beginning readers.

Maze Sentences. Students are required to read a series of brief (3-5 word) sentences that each have a blank space where the final word should be. Students use the computer mouse to select one of four Dolch words to correctly fill in the blank. For each probe, students are allowed one minute to respond to as many items as they can.

Items for the Maze Sentences task were created using preprimer, primer, and first-grade level Dolch words, as well as Dolch nouns. A test bank of 100 items was developed and probes were created by randomly sampling 25 items, with replacement, from the test bank. The first five items of each probe of the Maze Sentences task were comprised of 3-word sentences, and each of the four answer choices began with a different letter.

Dolch Word Recognition. Students' computer screens picture an object and four Dolch words. Students use the computer mouse to select the word that best describes the picture. For each probe, students are allowed one minute to respond to as many items as they

can. Items for the Dolch Word Recognition task were created using preprimer, primer, and first-grade level Dolch words, as well as Dolch nouns and clip-art picturing Dolch nouns. A test bank of 100 items was developed and probes were created by randomly sampling 25 items, with replacement, from the test bank.

It had been proposed that two Dolch Word Recognition probes would be administered on each of the four assessments throughout the year; however, technical difficulties that could not be remedied emerged during the first administration of the probes. The image files used in each item were different sizes, therefore items loaded onto the computer screen in varying amounts of time. Also, for an unknown reason, the images that appeared on the students' screens did not match with the answer choices provided on the same screen. Therefore, the probes that were administered during the first assessment period were unscoreable and Dolch Word Recognition was eliminated from the study.

Criterion Measures

DIBELS Oral Reading Fluency (ORF). The ORF task is comprised of grade-level passages of equivalent difficulty that the student should be able to read fluently by the end of the academic year. Students read aloud from a grade-level passage for 1 minute to a trained examiner, who recorded the number of words read correctly as the score. Words omitted, substituted, and hesitations of more than 3 seconds are counted as errors. Research supports the reliability and validity of ORF as a measure of overall reading performance as well as its utility for instructional decision making. Fuchs, Deno, and Mirkin (1984) stated that test-retest reliability of ORF ranged from .93 to .96 and that the concurrent validity of ORF with

the Woodcock Reading Mastery Tests, Word Identification and Passage Comprehension Tests ranged from .54 to .92. Additionally, the internal consistency reliability ranged from .66 to .79. DIBELS ORF was administered to students during the second, third, and fourth assessments because the measure is not intended for use prior to mid-way through the first grade.

Word Identification Fluency (WIF). With WIF, students are presented with a list of 50 high-frequency words and given one minute to read aloud from the list. If a child hesitates on an item for 4 seconds, he or she is prompted to try the next word. A child's total score is the number of words read correctly. Alternate forms of the measure are random samplings (with replacement) from a list of 100 preprimer, primer, and first-grade level high frequency words. Past research by Deno, Mirkin, and Chiang (1982) and Fuchs et al. (2004) has shown that the alternate-forms reliability for Word Identification Fluency ranges between .88 and .97; concurrent validity with the reading comprehension subtest of the Peabody Individual Achievement Test and the phonetic analysis and the inferential and literal reading comprehension subtests of the Stanford Diagnostic Reading Test ranges between .68 and .71.

Procedure

Consent forms for individual assessments (WIF and ORF, the criterion measures) were distributed to all parents of first graders at the beginning of the 2007-2008 academic year by placing the study description in the students' weekly parent folders. All first grade students with permission were administered two probes of the Reading Fluency and Maze Sentences tasks in their computer classes four times throughout the academic year as part of

the school's reading program. Assessments were administered in early December, early February, late March, and mid-May. When each class entered the computer classroom for Discourse assessments, the computer resource teacher introduced students to the task and identified the university assistants who would be helping with the assessments, and explained that the University assistants would control what the students would be seeing on each of their computer screens. She then led the students through the steps of logging into the Discourse program and modeled how to select answers to items of each measure. The teacher was not asked to read aloud a set of instructions because pilot testing showed that standardized instructions caused problems when the teacher could not deviate from instructions in order to enhance students' understanding. Instead, the teacher was instructed to incorporate the following points into her instructions to students, and instructions were audiotaped and verified that all components were present:

1. Go to the Reading/Writing folder (a location on the computer with which students were already familiar) and click on the blue "Discourse Assessment" icon.
2. Make the window bigger, because when it first opens it is small; the square in the top right corner will make that window bigger.
3. Click on "File," then "Connect." A box that says, "Select your teacher" and has the word "Administrator" in it pops up; click on the word "Administrator" and then click on "OK."
4. Next, a box with a list of all of the students' names pops up; each child should click on his or her own name, and then click "OK."

5. The screen will then say, “Please wait for your teacher to continue.” Everyone should have their screens looking like this, and then the University assistant will begin the assessment.

The teacher also led students through three sample items for each task. For the Maze Sentences task, the teacher demonstrated for students that they should read the sentence, select the best word to complete the sentence, click on the word so that the circle (radio button) next to the word becomes filled in with a dot, and then click the arrow that says “Next” on the toolbar at the top of the screen. Instructions were the same for the Reading Fluency task, except that students were instructed to select “Yes” if the statement was true and “No” if the statement was false. During the first assessment period, in which the Dolch Word Reading task was administered, students were instructed to select the word that best matches the picture. For all probes, students were told that after a minute of working on each task, the university assistants would pause their computers. Students were instructed to wait quietly while the next task was prepared. After the tasks were completed, students would return to their usual class work.

Once students’ understanding had been ensured, students logged into the Discourse system with the help of the computer resource teacher, the principal investigator, and university student research assistants. Once all students were ready to begin, the principal investigator and university assistants began the first task; the order of the experimental measures was randomized at each assessment. When students had had one minute to complete the first probe, their computers were paused while the second probe was opened.

Students then had one minute to complete the second probe, before moving on to the two probes of the second and third tasks. Upon completion of the probes, students then resumed their typical computer classroom activities and data were saved to the computer.

Those students who had written parental permission to participate in the individual assessments were also assessed on the two criterion measures within two weeks of the experimental assessments (with the exception that ORF was not included in the first assessment). Students were pulled out of their classrooms (at times acceptable to their teachers) by trained university students to complete the ORF and WIF measures. The order in which these measures were presented was counterbalanced within and across students. When administering ORF, examiners provided the following directions to examinees:

> Please read this (point to passage) out loud. If you get stuck, I will tell you the word so you can keep reading. When I say "Stop," I may ask you to tell me about what you read, so do your best reading. Start here (point to the first word of the passage). Begin.

Verbal instructions provided to examinees prior to the WIF assessment were:

> When I say "Go," I want you to read these words as quickly and correctly as you can. Start here (point to the first word) and go down the page (run your finger down the first column). If you don't know a word, skip it and try the next word. Keep reading until I say stop. Do you have any questions? Ready? Go.

Examiners provided a sample probe for the purposes of illustrating the task while instructing participants on the task. Just before triggering the stopwatch, the examiners flipped a page to show the testing probe. Students completed one 1-minute probe from each of the two measures, and students' responses to both were audiotaped. Upon completion of the two probes, students were returned to their classrooms.

Study 2

Study 2 was contingent upon the results of Study 1; the decision about whether or not to proceed with Study 2 was made after the first round of data collection for Study 1. It had been proposed that Study 2 would be conducted using any tasks that demonstrated adequate alternate forms reliability and adequate concurrent validity with WIF after the first data collection period. The Reading Fluency and Maze Sentences tasks met criteria, thus, Study 2 proceeded with these two tasks.

Participants

Students. Four first grade students participated in Study 2. These students were identified by (a) performance in the bottom 20% of students on WIF and/or the experimental measures, and (b) verification from their teachers that they were at risk of reading failure and could benefit from additional reading assistance using the Sound Partners program. Once the four students had been identified, written parental consent was obtained. Of the four participants, three were male and one was female; they ranged in age from 6 years, 3 months to 6 years, 5 months at the beginning of the study. The students came from two different classrooms; two from one classroom and two from another. Parental consent for Students 3

and 4 was received later in the study, such that they were added to baseline data collection two to three weeks after Students 1 and 2 had begun the study.

Tutors. A total of six volunteers participated in providing Sound Partners tutoring: the principal investigator and her faculty supervisor, one graduate student in psychology, one post-baccalaureate psychology student, and two undergraduate psychology majors. The university student volunteers were trained on the Sound Partners program by the principal investigator (see Tutor Training and Treatment Fidelity, below).

Intervention

Sound Partners. Sound Partners is a phonics-based tutoring program designed for students in the first grade and above who are having difficulty learning decoding skills. The program consists of 100 scripted lessons that are each made up of six to nine activities that teach letter sounds, segmenting, decoding, spelling, sight words, and fluency. Lessons 1-30 focus on phonemic awareness and include instruction in common sounds, letter pairs, and sight words. These lessons also provide students with practice spelling words, reading word lists, and reading storybooks. Phonics instruction is an area of focus in lessons 31-60; children receive more reading and spelling practice as they learn words with initial/final blends, silent e's, and word endings such as –s, -ed, and –ing. Lessons 61-100 teach students strategies for reading longer words; story reading becomes more of a focus in these lessons. Mastery tests are conducted after every 10th lesson; the authors suggest that students performing lower than 90% accuracy may need to work at a slower pace and spend more

time reviewing before moving on to new skills. Mastery tests may also be utilized for the placement of students who do not need to begin at lesson 1 (Vadasy et al., 2005).

Sound Partners underwent a 5-year period of development and refinement and has demonstrated the ability to improve students' reading achievement (Jenkins, Vadasy, Firebaugh, & Profilet, 2000). For example, Jenkins et al. (2000) reported a range in effect sizes from .51 on reading fluency to 1.16 on decoding skills. Similarly, Vadasy, Jenkins, Antil, Wayne, and O'Connor (1997) demonstrated that, when tutors implemented Sound Partners with high treatment integrity, effect sizes ranged from 0.70 to 1.40 on measures of word reading (WRAT-R Reading, Dolch, and the Analytical Reading Inventory) and nonword reading (Woodcock Johnson Word Attack, Bryant, and the Pseudoword List).

Measures

The Reading Fluency and Maze Sentences tasks, administered via the Discourse program, were the experimental measures utilized in Study 2; WIF and ORF were the criterion measures. A tutor observation checklist was utilized to evaluate the fidelity of implementation of the Sound Partners program. As used by Vadasy, Sanders, Peyton, and Jenkins (2002), the checklist includes core components of the Sound Partners program, such as: letter-sound instruction, segmenting, phoneme blending, spelling, sight-word instruction, phonics instruction in silent-e words and word endings, and scaffolding during storybook reading.

Procedure

Study 2 was conducted during the students' spring semester of first grade. In this multiple-baseline study, students participated in progress monitoring assessments (using the criterion and experimental measures) weekly and received tutoring using the Sound Partners tutoring program for 7 to 11 weeks. At the beginning of the study, three probes of each criterion measure were administered weekly; however, not enough ORF passages had been published to ensure that different passages were used every week. Therefore, beginning with the third week of baseline data collection, only two probes of each measure were administered weekly. As only 20 DIBELS ORF passages have been published at the first grade level, passages were recycled such that the portion of each passage that students had previously read (approximately the first 40 words, rounding to the nearest whole paragraph) was cut from each passage; cut passages were then administered to participants at least three weeks after the students had initially read the original passage. Thus, participants did not re-read the identical sentences they had previously read and there was a period of at least three weeks' time that elapsed between readings from the same passage.

Participants began Sound Partners tutoring after at least three weeks of baseline data collection and after at least one of the progress monitoring measures showed relative stability. The exceptions to this protocol were Students 3 and 4, whose baseline and intervention schedules were affected by the school's spring break and the researcher's ethical responsibility to provide intervention within a timely manner. These two students' baseline

data were not as stable as one would hope, but delaying intervention would not have been appropriate given the upcoming school holidays.

During baseline data collection, each student was pulled out of his or her class once weekly by the primary investigator or a trained university student. Students were taken to a quiet location and were administered the WIF and ORF assessments; one morning each week, all four of the participants were taken to the computer classroom to engage in Discourse assessments. During each student's intervention phase, he or she was pulled out of his or her classroom three times each week by either the primary investigator or university student trained on the Sound Partners program. A period of 25 minutes was allotted to Sound Partners tutoring sessions; during the first tutoring session of each week, students were assessed on the criterion measures (ORF and WIF) prior to beginning tutoring. Discourse assessments took place weekly at a time agreed upon by the child's classroom teacher and the computer technology teacher.

Tutor Training and Treatment Fidelity

Tutors received one hour of training from the principal investigator prior to the beginning of the study. The principal investigator described the rationale for the project, modeled tutoring activities, and observed and provided feedback as tutors practiced. Tutors audiotaped each of their tutoring sessions and a random sample of one-third of the recordings (drawing equally from each participant's tutoring sessions) was evaluated by the principal investigator using the tutor observation form.

The principal investigator and tutors stayed in regular contact via email throughout the duration of the study so that tutors could have questions answered and share pertinent information regarding tutoring. After approximately one month of tutoring, the principal investigator met with the tutors to answer additional questions that had arisen and to correct minor errors that became evident during initial review of audio recordings.

CHAPTER 5

Results

The current chapter presents the results for the research hypotheses for Studies 1 and 2 presented in Chapter 3. Preliminary analyses, including descriptive data, are presented first, followed by results of analyses pertaining to each hypothesis.

Study 1

Preliminary Analyses

Criterion measures. Descriptive data for WIF and ORF scores are presented in Table 2. WIF mean scores show an overall increase in words correct over time, from approximately 39 words at the first assessment to nearly 66 words at the fourth assessment, whereas ORF mean scores increased from 81 to 89 words.

Table 2

Mean Score, Standard Deviation, Range, and Sample Size for WIF and ORF at Assessment Periods One Through Four

Measure		1	2	3	4
WIF					
	M	39.09	45.84	54.98	65.99
	SD	28.97	31.02	35.70	40.58
	Range	1 – 107	6 – 109	7 – 109	8 – 120
	N	46	44	46	42
ORF					
	M	N/A	81.07	82.12	89.42
	SD	N/A	50.19	50.88	53.53
	Range	N/A	15 – 159	8 – 168	15 – 171
	N	N/A	44	46	42

Table 3 presents correlation coefficients representing the stability of ORF and WIF scores between assessment periods. Correlations between administrations of ORF ranged from r = .92 - .95, indicating strong reliability. Findings for WIF were similar, with correlations ranging from r = .87 - .94.

Table 3

Stability of ORF and WIF Between Assessments

Assessment Period	2	3	4
		ORF Probes	
2	———	.95	.92
3	———	———	.95
		WIF Probes	
1	.94	.88	.93
2	———	.87	.93
3	———	———	.93

Table 4 presents correlation coefficients representing the concurrent and predictive validity of ORF and WIF scores relative to each other. All correlations were above r = .85, suggesting strong validity of the criterion measures.

Table 4

Validity of Criterion Measures Relative to Each Other

Probe	WIF 1	WIF 2	WIF 3	WIF 4
ORF 2	.90	.92	.91	.90
ORF 3	.86	.89	.88	.89
ORF 4	.89	.86	.85	.90

Experimental measures. Table 5 shows the mean score (in words correct per minute), standard deviation, and range for Reading Fluency and Maze Sentences at each of the four assessment periods. Reading Fluency mean scores did improve across the assessment periods, from between 5 and 6 words correct at the first assessment to 8 to 9 words correct at the final assessment. Maze Sentences mean scores also improved across the assessment periods, from approximately 4 words correct at the first assessment to approximately 7 words correct at the final assessment. Across all four assessment periods, students answered significantly more Reading Fluency items ($M = 9.60$) than Maze Sentences items ($M = 7.07$), $t(72) = 11.45$, $p < .01$.

Table 5

Mean Score, Standard Deviation, Range, and Sample Size for the Reading Fluency and Maze Sentences Tasks at Assessment Periods One Through Four

	1		2		3		4	
	A	B	A	B	A	B	A	B
Reading Fluency								
M	5.47	6.37	5.95	6.77	8.67	8.75	9.05	8.35
SD	3.81	4.57	3.97	4.41	5.12	5.35	5.41	5.47
Range	0 –13	0 –14	0 –13	0 –15	0 - 16	0 - 16	0 - 17	0 - 17
N	43	43	38	35	64	64	63	62
Maze Sentences								
M	4.79	4.65	5.73	6.27	6.19	6.50	7.46	7.61
SD	3.59	3.77	4.07	4.16	4.13	4.36	4.25	4.75
Range	0 –11	0 –13	0 –12	0 –12	0 –13	0 –14	0 –14	0 –14
N	42	63	56	56	62	62	63	62

Inter-scorer reliability. Inter-scorer reliability coefficients for WIF and ORF were calculated for a random sample of 30% of protocols from each of the four assessment periods by dividing the total number of scoring agreements between two independent scorers by the total number of items scores. The analysis yielded 99% reliability between 2 scorers on both WIF and ORF.

Analyses of Specific Questions

Question 1. Is the Reading Fluency task appropriate for the progress monitoring of first grade early literacy skills?

Hypothesis 1. It was hypothesized that the Reading Fluency task would be a reliable measure of first grade early literacy skills, as indicated by adequate alternate forms reliability (Hypothesis 1A) and adequate test-retest stability (Hypothesis 1B). Hypothesis 1A was supported; as can be seen in Table 6, all of the correlations between alternate forms of the Reading Fluency task were at or above the criterion level of $r = .80$. Correlation coefficients ranged from .80 to .88. Hypothesis 1B was not supported; only eight of the 24 correlations between administrations of the Reading Fluency task were at or above $r = .80$. Correlation coefficients ranged from .60 to .88.

Table 6

Alternate Forms and Test-Retest Reliability Coefficients for the Reading Fluency Task

Probe		1		2		3		4	
		A	B	A	B	A	B	A	B
1	A	—	.82	.73	.69	.74	.76	.74	.76
	B		—	.77	.60	.84	.81	.77	.83
2	A			—	.81	.79	.71	.74	.74
	B				—	.79	.80	.77	.77
3	A					—	.88	.83	.88
	B						—	.85	.87
4	A							—	.80
	B								—

Note. Shaded cells denote alternate forms reliability.

Hypothesis 2. It was hypothesized that the Reading Fluency task would demonstrate adequate concurrent validity (Hypothesis 2A) and predictive validity (Hypothesis 2B) when correlated with WIF and ORF scores, evidenced by the majority of correlation coefficients at or above $r = .80$. As can be seen in Table 7, results support Hypothesis 2A. The majority of correlation coefficients between Reading Fluency and WIF (range = .78 - .89) and between Reading Fluency and ORF (range = .76 - .83) were at or above $r = .80$. However, results only partially support Hypothesis 2B. The majority of correlations between Reading Fluency and WIF, with Reading Fluency scores predicting WIF scores at the fourth assessment, were at or above the criterion level of .80 (range = .78 - .87). Only half of the correlation coefficients describing the strength with which Reading Fluency scores predict ORF scores reached the criterion level (range = .71 to .84).

Table 7

Concurrent and Predictive Validity Coefficients for the Reading Fluency Task Relative to WIF and ORF

Probe	1		2		3		4	
	A	B	A	B	A	B	A	B
WIF1	.89	.86	.81	.80	.81	.80	.78	.83
WIF2	.85	.86	.78	.81	.79	.76	.76	.80
WIF3	.88	.86	.79	.88	.80	.78	.81	.81
WIF4	.87	.87	.78	.85	.82	.81	.81	.84
ORF2	.86	.86	.76	.80	.83	.78	.83	.82
ORF3	.82	.80	.69	.82	.83	.77	.84	.80
ORF4	.82	.79	.71	.84	.82	.77	.80	.80

Hypothesis 3. It was predicted that the Reading Fluency task would model students' growth over time as evidenced by a positive, non-zero slope across the four data points. The slope of Reading Fluency scores was calculated based on the one probe from each

assessment period that had better psychometric properties (1B, 2B, 3A, 4B). Changes in scores over time yielded a mean slope that was positive, but not statistically different from zero ($M = 0.06$, $SD = 0.21$), $t(59) = 0.28$, $p > .05$ (one-tailed; see Table 8). Thus, the hypothesis is not supported.

Table 8

Mean Slopes and Standard Deviations for Experimental and Criterion Measures

Measure	*M*	SD	Range	*t*
Reading Fluency	0.06	0.21	-0.67 – 0.47	0.28
Maze Sentences	0.13	0.16	-0.50 – 0.50	0.81
WIF	1.12*	0.64	-0.13 – 2.70	1.75**
ORF	0.53	1.20	-2.00 – 3.70	0.44

*based on total number of words correct
**significant at $p = .05$

Question 2. Is the Maze Sentences task appropriate for the progress monitoring of first grade early literacy skills?

Hypothesis 4. It was hypothesized that the Maze Sentences task would be a reliable measure of first grade early literacy skills, as indicated by adequate alternate forms reliability (Hypothesis 4A) and test-retest stability (Hypothesis 4B). Hypothesis 4A was supported; the majority of the correlations between alternate forms of the Maze Sentences task were at or above $r = .80$ (range = .76 - .88; see Table 9). Hypothesis 4B, however, was not supported; only five of the 24 correlations between administrations of the Maze Sentences task were at or above $r = .80$ (range = .61 - .86).

Table 9

Alternate Forms and Test-Retest Reliability Coefficients for the Maze Sentences Task

Probe		1		2		3		4	
		A	B	A	B	A	B	A	B
1	A	—	.86	.79	.69	.61	.68	.72	.63
	B		—	.84	.77	.75	.85	.78	.79
2	A			—	.88	.83	.86	.79	.80
	B				—	.72	.75	.71	.71
3	A					—	.83	.72	.65
	B						—	.78	.73
4	A							—	.76
	B								—

Note. Shaded cells denote alternate forms reliability.

Hypothesis 5. It was hypothesized that the Maze Sentences task would demonstrate adequate concurrent validity (Hypothesis 5A) and predictive validity (Hypothesis 5B) when correlated with WIF and ORF scores. Results did not support these hypotheses. As can be seen in Table 10, the majority of correlation coefficients between Maze Sentences and WIF (range = .54 - .86) and between Maze Sentences and ORF (range = .48 - .81), when the measures were presented concurrently, were not at or above $r = .80$. When Maze Sentences scores were used to predict WIF scores, correlation coefficients ranged from .65 - .87; when Maze Sentences scores were used to predict ORF scores, correlation coefficients ranged from .54 - .83. As with Hypothesis 5A, the majority of correlations did not reach the criterion level predicted for support of the hypothesis.

Table 10

Concurrent and Predictive Validity Coefficients for the Maze Sentences Task Relative to WIF and ORF

Probe	1		2		3		4	
	A	B	A	B	A	B	A	B
WIF1	.78	.86	.80	.77	.58	.65	.74	.85
WIF2	.61	.76	.73	.67	.49	.57	.71	.80
WIF3	.68	.82	.84	.80	.54	.65	.73	.82
WIF4	.77	.87	.83	.75	.65	.67	.74	.85
ORF2	.63	.79	.81	.77	.46	.62	.78	.82
ORF3	.64	.79	.79	.76	.48	.63	.75	.76
ORF4	.74	.83	.80	.72	.54	.62	.74	.80

Hypothesis 6. It was predicted that the Maze Sentences task would model students' growth over time as evidenced by a positive slope across the four data points that was significantly different from zero. The slope of Maze Sentences scores was calculated based on the one probe from each assessment period that had better psychometric properties (1B, 2A, 3B, 4B). Changes in scores over time yielded a mean slope that was positive, but not statistically different from zero ($M = 0.13$, $SD = 0.16$), $t(63) = 0.81$, $p > .05$ (one-tailed; see Table 8). Thus, the hypothesis was not supported.

Study 2

Preliminary Analyses

Treatment integrity. Twenty-four tutoring sessions (one-third of each student's sessions) were selected by random draw to be coded for treatment integrity. Across the sample, tutors completed a mean of 96.67% of the treatment components and 96.02% of the

session management strategies included on the Sound Partners Tutor Observation Form (see Table 11).

Table 11

Percentage of Sound Partners Treatment Components and Session Management Strategies Completed by Tutors

Student	Number of Sessions Sampled	Treatment Components	Session Management
1	9	96.61	99.14
2	6	98.83	92.38
3	4	95.50	96.15
4	5	93.36	95.46
Total	24	96.67	96.02

Descriptive statistics. Table 12 displays the number of Sound Partners sessions in which each student participated, the average amount of time spent on Sound Partners tutoring sessions, the range of session length, and the lessons that each student completed. Twenty-five minutes were allotted for each tutoring session; however, class activities, fire drills, and the tendency for teachers to take their classes to lunch early limited the ability of tutors to keep students for the full time period. Amount of time spent on WIF and ORF assessments was subtracted out of total session time, such that the times presented reflect time allocated to Sound Partners only.

Student 1 was the first student to begin intervention, therefore he participated in the greatest number of tutoring sessions. He was learning English as a second language. Student 2 was the second student to begin intervention; he tended to have difficulty attending to the task, even in the one-on-one sessions. He often talked with the RAs about non-related topics

and required frequent redirection to the task. Student 3 also required multiple behavioral redirections during every tutoring session due to inattentiveness; RAs engaged her in "games" to help her sit still and focus on the task at hand. However, progress during sessions tended to be slow; for instance, during one session it took 9 minutes for her to read 12 words (which she was able to sound out correctly) with the RA. Finally, Student 4 and attended well to his tutors and required few redirections.

Table 12

Gender, Age, Number of Intervention Sessions, Time Spent in Intervention, and Sound Partners Lessons Completed

Student	Gender	Age	Number of Sessions	Lessons Completed	Session Length (mins)	
					Mean	Range
1	Male	6 yrs, 5 mos	26	1 - 21	16.40	10 – 22.0
2	Male	6 yrs, 3 mos	19	61 - 71	15.88	8 – 22.5
3	Female	6 yrs, 4 mos	11	41 - 43	17.87	10 – 25.5
4	Male	6 yrs, 4 mos	14	17 - 27	16.80	5 – 24.0

Analyses of Specific Questions

Question 1. Do the computerized early literacy assessments demonstrate sensitivity to intraindividual change in response to intervention?

Hypothesis 1A. It was predicted that, during intervention, students' scores on the Reading Fluency task would show positive linear growth from baseline levels.

As can be seen in Table 13, the average performance of Student 1 did improve from baseline (M = 3.33 wcpm) to intervention (M = 6.43 wcpm). However, his slope of

improvement during baseline and intervention reveals a trend towards decreasing scores (slopes of -0.50 and -0.03, respectively); therefore, it cannot be concluded that Student 1 showed positive linear growth on the Reading Fluency task. Lack of growth is also evident via visual inspection of his progress, available in Figure 2.

Student 2 showed no change in mean performance on RF from baseline (M = 4.00) to intervention (M = 4.00). However, his slope did improve from baseline (-2.40) to intervention (0.65), but it is not clear whether the change in slope from negative to positive represents true improvement in skills or random variability among his scores. Inspection of standard deviations reveals an increase in variability from baseline to intervention.

Review of Student 3's mean level of performance and slopes suggest that she did show improvement from baseline to intervention (M = 4.67 during baseline to M = 9.75 during intervention; slope = 0.00 during baseline to slope = 0.35 during intervention). However, visual inspection of her graph reveals more stable performance, with one lower score during baseline which likely decreased her baseline mean score and slope. Thus, it cannot be concluded that Student 3 showed positive linear growth on the RF task.

Student 4's mean level of performance did improve from baseline (M = 8.00) to intervention (M = 11.33). His baseline slope (0.88) suggests improvement during baseline, but his intervention slope (-1.38) suggests a decrease in performance during the tutoring program. Visual inspection of his graphed data reveals considerable variability in his performance; therefore, data from Student 4 do not support the hypothesis.

Hypothesis 1B. It was predicted that, during intervention, students' scores on the Maze Sentences task would show positive linear growth from baseline levels.

Student 1 did not show improvement on the Maze Sentences task from baseline to intervention. Neither his mean scores at baseline and intervention (M = 1.67 and 1.00, respectively) nor his slopes (-0.50 during baseline and 0.18 during intervention) reflected improvement. Furthermore, visual inspection of his graphed data do not provide evidence that would support the hypothesis that scores would show positive growth from baseline levels. Student 1 was learning English as a second language and demonstrated consistent difficulty with the letters b, d, p, and q throughout Sound Partners lessons; research assistants reported that his difficulty distinguishing between b, d, p, and q began to hamper his progress during the latter half of the intervention phase of the study.

There is some evidence that Student 2 showed mild improvement from baseline to intervention. His mean level of performance improved from 2.50 items correct at baseline to 4.33 items during intervention, and his baseline slope of 0.00 improved to 0.75 during intervention. However, inspection of his graph reveals substantial variability among his scores; his graph shows that two of his scores during baseline were higher than the majority of his scores during intervention. Thus, one cannot conclude that his scores on the Maze Sentences task showed positive linear growth.

Inspection of Student 3's mean level of performance (M = 4.67 during baseline and M = 7.00 during intervention) as well as her graphed data suggest some mild improvement in Maze Sentences performance over time. Furthermore, variability that was present in her

baseline data is not present to the same extent in her intervention scores (*SD* = 2.73 during baseline and *SD* = 1.00 during intervention). However, her change in slope from baseline to intervention (-0.07 to -0.14) does not suggest positive linear growth.

Student 4 did show positive linear growth in Maze Sentences scores from baseline to intervention both in terms of mean level of performance (*M* = 7.50 during baseline and *M* = 10.00 during intervention) and growth over time (slope = 0.57 during baseline and slope = 1.31 during intervention). His growth is also evident via visual inspection of the student's graph. Therefore, data for Student 4 does provide support for Hypothesis 1B.

Supplemental Analysis: To what degree did criterion measures demonstrate positive linear growth over time?

ORF. Mean scores on ORF did improve from baseline to intervention for each of the four participants, as can be seen in Table 13. Furthermore, visual inspection of graphed data points indicates that Students 1, 2, and 3 showed improvements in ORF scores over time. Student 4 also appears to have shown improvement, but growth is less clear. Despite these indications of improvement, analyses of students' slopes failed to provide evidence of growth over time on ORF.

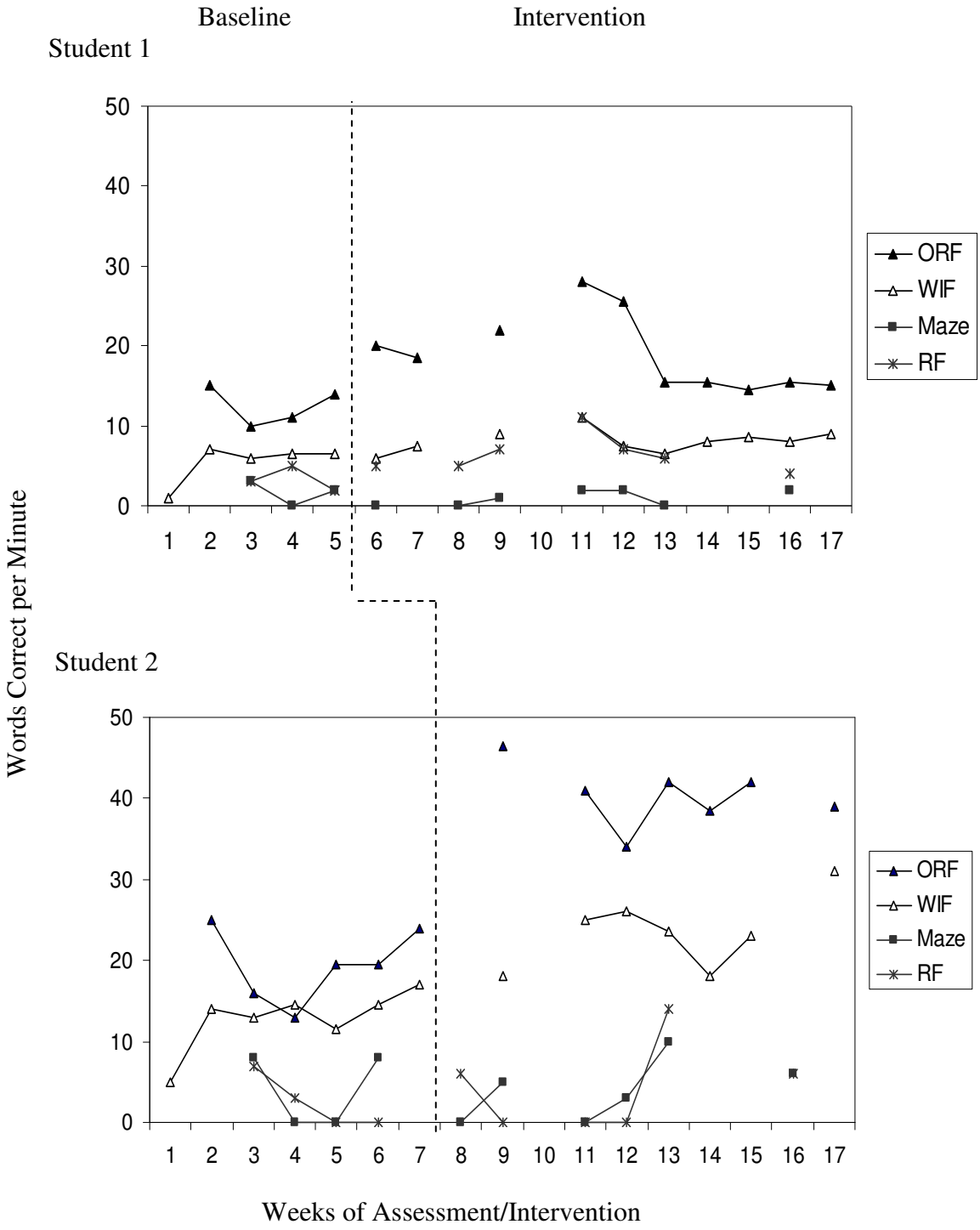
WIF. Similar to analyses on ORF, WIF data provide contradictory evidence regarding students' growth over time. Inspection of mean levels of performance on WIF does suggest that all four students improved over time. However, Student 2 is the only participant whose graph shows clear improvement. Analysis of students' slopes failed to demonstrate improvement in scores over time.

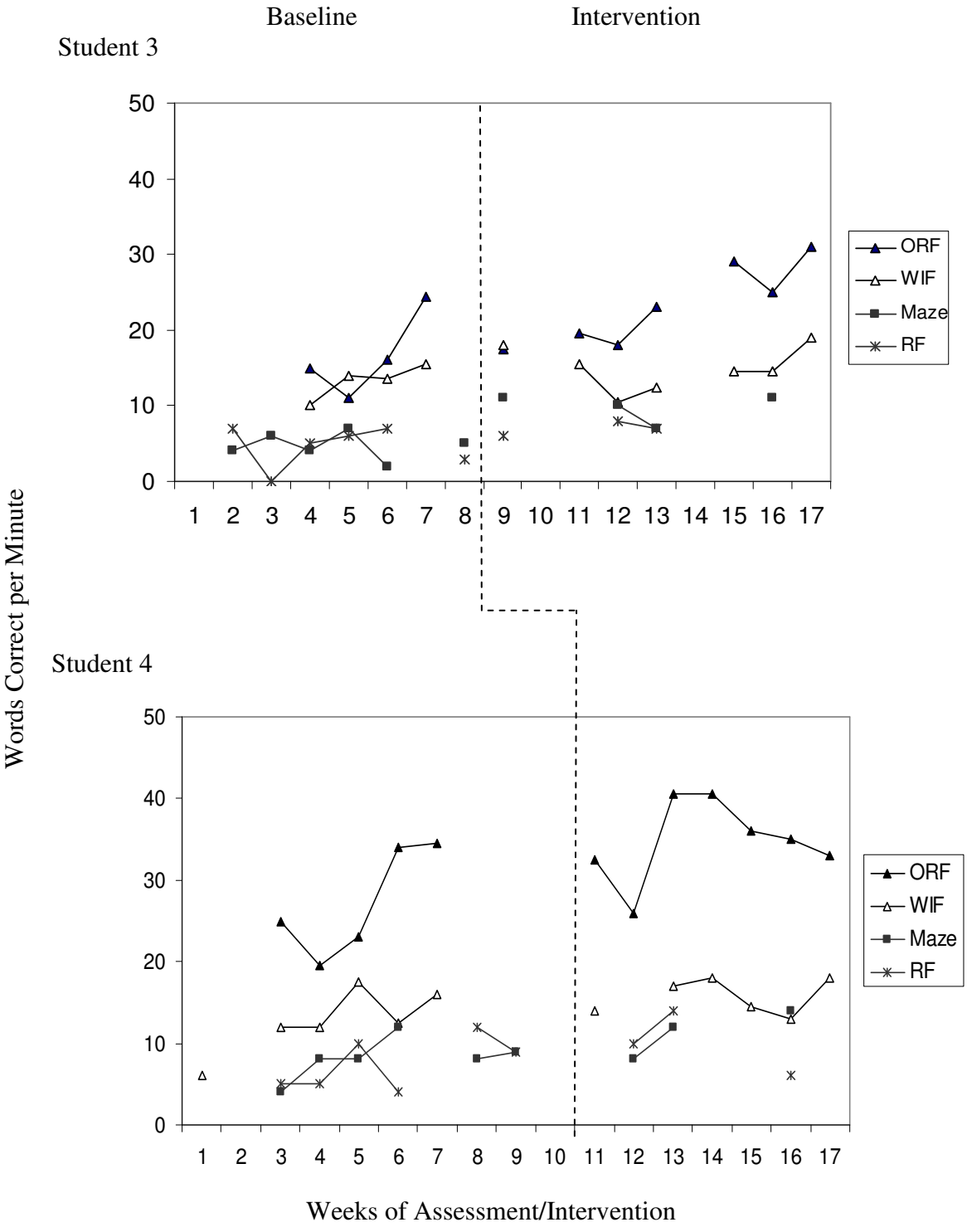
Table 13

Mean Scores, Standard Deviations, and Slopes of Improvement Across Measures Administered during Baseline and Intervention

	ORF		WIF		RF		Maze	
	Base.	Int.	Base.	Int.	Base.	Int.	Base.	Int.
Student 1								
Mean	12.50	19.00	5.40	8.10	3.33	6.43	1.67	1.00
SD	2.38	4.79	2.49	1.41	1.53	2.30	1.53	1.00
Slope	-0.20	-0.63	0.15	0.12	-0.50	-0.03	-0.50	0.18
Student 2								
Mean	19.50	40.43	12.79	23.50	4.00	4.00	2.50	4.33
SD	4.50	3.86	3.65	4.57	3.32	5.57	4.62	3.85
Slope	0.34	-0.56	0.13	-0.55	-2.40	0.65	0.00	0.75
Student 3								
Mean	16.63	23.29	13.25	14.93	4.67	9.75	4.67	7.00
SD	5.68	5.34	2.33	2.95	1.75	1.89	2.73	1.00
Slope	3.35	1.69	1.60	0.10	0.00	0.35	-0.07	-0.14
Student 4								
Mean	27.20	34.79	12.67	15.75	8.17	11.33	7.50	10.00
SD	6.73	5.04	4.00	2.18	2.56	3.06	3.27	4.00
Slope	3.35	0.54	1.51	0.17	0.88	-1.38	0.57	1.31

Figure 2. Students' scores on ORF, WIF, Maze Sentences, and Reading Fluency (RF) during baseline and intervention.


Baseline
Intervention
Student 1
Student 2
Words Correct per Minute
Weeks of Assessment/Intervention
ORF
WIF
Maze
RF
0
10
20
30
40
50
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17


Baseline
Intervention
Student 3
Student 4
Words Correct per Minute
Weeks of Assessment/Intervention
ORF
WIF
Maze
RF
0
10
20
30
40
50
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

CHAPTER 6

Discussion

Educational reform has initiated a shift from the traditional approach to instructional decision-making, the diagnostic-prescriptive model, to an alternative approach, RTI. Within RTI, students who are at risk for academic failure are identified and provided with increasing levels of intensive intervention. Two factors that are integral to the effectiveness of RTI are (a) the proper identification (screening) of at-risk students who are in need of additional or alternative instruction, and (b) accurate monitoring of these students' progress throughout the course of intervention in order to ensure that they are receiving the level and type of intervention that they need.

Formative evaluation is a method of monitoring the effectiveness of instruction and has been shown to positively affect student outcomes (e.g., Deno, 1990; Fuchs, Deno, & Mirkin, 1984). One type of formative evaluation, progress monitoring, involves assessment that occurs at least three times a year using alternate but equivalent test forms (Kame'enui, 2002). A number of tools, such as CBM and DIBELS, are currently utilized for progress monitoring; however, they are not ideally suited for the screening progress monitoring of early academic skills. Nearly all of the measures within CBM and DIBELS require individual assessment, which is time-consuming when large groups of students are to be screened, and measures such as ORF require reading skills beyond a beginning first grade level. Thus, these assessments are not necessarily conducive to *early* screening and progress monitoring.

Therefore, the aim of the present project was to introduce innovative progress monitoring measures that were intended to be more appropriate for the screening and progress monitoring of first grade literacy skills. The experimental measures were computerized in order to facilitate group assessments and targeted toward early literacy skills, such as sight word fluency and reading connected text composed of high-frequency words. The objectives of Study 1 were to establish the reliability and validity of two experimental measures and to determine to what extent the measures were able to model students' growth over time. The objective of Study 2 was to determine how sensitive the measures were to individual changes in growth, in order to evaluate whether the experimental measures are appropriate for ongoing progress monitoring within an RTI framework. In both studies, the experimental measures were compared to two established and routinely-utilized progress monitoring measures: ORF and WIF.

The current chapter presents a discussion of the results of Studies 1 and 2. The discussion of Study 1 begins with a discussion of results as they relate to the six research hypotheses and existing research. The section closes with a general discussion of how the overall findings of Study 1 relate to the current literature base, and limitations of that study. The next section follows a similar organization but focuses on the findings of Study 2. In the third section, Studies 1 and 2 are discussed as a whole, and findings are discussed in terms of contributions made within the broader contexts of research and practice. Finally, future directions for research and implications for practice are presented.

Study 1

Hypothesis 1

Hypothesis 1 predicted that the Reading Fluency task would be a reliable measure, both in terms of alternate forms stability (Hypothesis 1A) and test-retest reliability (Hypothesis 1B). Although the Reading Fluency task has not been previously researched, there was reason to believe that the task would demonstrate adequate reliability based on prior research suggesting that ORF, another progress monitoring measure of the ability to read connected text, is a reliable measure (Fuchs, Deno, & Mirkin, 1984).

It is interesting that results met criteria for support of Hypothesis 1A but not for Hypothesis 1B. However, within the present study, test-retest reliability was calculated using different probes, rather than the same probe, administered on separate occasions. Therefore, our "test-retest" reliability may be more appropriately conceptualized as "delayed alternate forms" reliability. It is not surprising that two different probes would tend to demonstrate higher reliability when administered on the same day than two probes administered several weeks apart.

Additionally, although results failed to meet the criterion necessary to support Hypothesis 1B, review of the literature indicates that the Reading Fluency task may still demonstrate a generally acceptable magnitude of reliability. The criterion magnitude of r = .80 that was established for the present study is within the high range of reliability coefficients that have been reported for progress monitoring measures currently being utilized. For example, reliability coefficients for ORF range from .56 to .96 (Hintze &

Shapiro, 1997; Hintze, Shapiro, & Lutz, 1994). Therefore, although Reading Fluency did not meet the standards for reliability that were established for the present study, the obtained reliability coefficients ranging from .60 to .88 are within a range that, based on current use of progress monitoring tools, appears to be acceptable to educators. However, it should be noted that correlation coefficients as low as .60 are considered to indicate only a *moderate* effect whereas the criterion level set for the present study is associated with a *strong* effect (Cohen, 1988). Thus, although the Reading Fluency task is as reliable as other progress monitoring measures, research should continue to develop the measure to a higher standard of reliability.

Hypothesis 2

Hypothesis 2 predicted that the Reading Fluency task would be a valid measure of first grade early literacy skills, demonstrated by evidence of adequate concurrent (Hypothesis 2A) and predictive (Hypothesis 2B) validity with ORF and WIF. The prediction was based on evidence that ORF had been shown to be a valid progress monitoring tool (Fuchs, Deno, & Mirkin, 1984). Thus, support of Hypothesis 2A suggests that the Reading Fluency task did, indeed, measure the same reading skills that ORF and WIF both measure, namely fluency and accuracy in reading connected text and sight words.

With respect to Hypothesis 2B, predictive validity coefficients did not meet the criterion established for the present study; results of correlations between Reading Fluency and WIF do support the hypothesis, but those between Reading Fluency and ORF do not. However, predictive validity correlations between Reading Fluency and ORF are comparable to the correlations between ORF and other reading measures. For instance, previous research

indicates correlations ranging from .54 to .92 between ORF and the Woodcock Reading Mastery Tests, Word Identification, and Passage Comprehension Tests (Fuchs, Deno, & Mirkin, 1984). Thus, Reading Fluency shows promise as a valid measure of early literacy skills, but the measure should be further developed in order to improve upon its predictive validity.

Hypothesis 3

Hypothesis 3 predicted that the Reading Fluency task would demonstrate growth over time, as evidenced by a positive linear slope of Reading Fluency scores. The prediction was made based on the fact that the Reading Fluency task was designed to measure skills that would be expected to improve over time, as well as on previous finding suggesting that ORF is sensitive to students' growth over time (e.g., Fuchs & Fuchs, 2004). Fuchs, Fuchs, Hamlett, Walz, and Germann (1993) found that first grade students improved by two words correct per week, on average, on ORF; it was hoped that the Reading Fluency task would reflect similar progress. However, Hypothesis 3 was not supported; although mean Reading Fluency scores showed a trend toward increasing over time, improvement was not reflected in a positive, non-zero slope.

The failure to yield a positive slope could be partially explained by the limited number of items included in the Reading Fluency task; students were, on average, responding to fewer than 10 items, whereas they answered two to three times as many items on WIF and five times as many on ORF. It is possible that a longer assessment would have been more sensitive to changes in students' performance over time. Related to this is the fact that on the

Reading Fluency task, students only had an opportunity to earn a point towards their score after every sentence (i.e., after every 3 – 6 words read), compared to ORF and WIF, which credit students for every word read correctly. Thus, students' slope does not represent the number of *individual words* that the student gained over time, but rather growth in the number of *whole items* the student answered correctly. Thus, perhaps comparison of Reading Fluency slope to slopes demonstrated for the CBM-Maze task would be more appropriate; CBM-Maze credits students for replacing a missing word that occurs after every seven words in a passage. Indeed, CBM-Maze slopes have been shown to be lower than ORF slopes; Fuchs et al. (1993) reported an average slope of 0.39 correct replacements per week for first through sixth graders in their study. In Fuchs et al.'s methodology, students are given 2.5 minutes to complete Maze probes, so it is possible that lengthening the Reading Fluency assessments would improve the measure's ability to detect growth.

Hypothesis 4

Hypothesis 4 predicted that the Maze Sentences task would be a reliable measure, both in terms of alternate forms stability (Hypothesis 4A) and test-retest reliability (Hypothesis 4B). The prediction was based on findings suggesting that CBM-Maze, a measure similar to the Maze Sentences task utilized in the present study, demonstrated adequate reliability, with reliability coefficients ranging from .69 to .91 (Shin, Deno, & Espin, 2000).

Results of the present study mirrored results for the Reading Fluency task; in both cases the measures showed adequate alternate forms stability but not evidence of adequate

test-retest reliability was not demonstrated. Although test-retest correlation coefficients did not reach the criterion level for support of the hypothesis, they were comparable to those reported by Shin et al. (2000). Thus, as discussed relative to the Reading Fluency task, results suggest that the Maze Sentences task is as reliable as other progress monitoring measures, but that the task does not demonstrate strong test-retest reliability.

Hypothesis 5

Hypothesis 5 predicted that the Maze Sentences task would be a valid measure of first grade early literacy skills, demonstrated by evidence of adequate concurrent (Hypothesis 5A) and predictive (Hypothesis 5B) validity with ORF and WIF. The prediction was made based on evidence that CBM-Maze has been shown to demonstrate adequate validity when compared to other reading measures. For instance, Jenkins and Jewell (1993) reported correlation coefficients ranging from .65 to .76 between CBM-Maze and the Gates-McGinitie Reading Test and the Metropolitan Achievement Test of Reading.

Results failed to support Hypotheses 5A and 5B; however, analyses yielded correlation coefficients that were generally comparable to those reported by Jenkins and Jewell (1993). Therefore, although it cannot be concluded that the Maze Sentences task is a valid measure based on the standards proposed in the current study, current findings suggest that it is no less valid than similar, established measures. Failure to reach the criterion magnitude established for the present study may be due to a relatively limited range of scores. The largest range of scores, ranging from 0 to 14, occurred during the third and fourth assessments; on the short end, scores ranged from 0 to 11. It is possible that such a limited

range did not allow for enough differentiation between students of varying skill levels compared to scores on the criterion measures (scores had greater than 100-point ranges on both ORF and WIF, thus allowing for much more differentiation between students).

Hypothesis 6

Hypothesis 6 predicted that the Maze Sentences task would demonstrate growth over time, as evidenced by a positive linear slope of Maze Sentences scores. The hypothesis was based on the fact that the task was developed to measure skills that are expected to improve over time and on previous research suggesting that the CBM-Maze task is sensitive to students' growth over time (Fuchs et al., 1993). However, the Maze Sentences task was not found to adequately model students' growth over time. As discussed previously relative to the Reading Fluency task, the lack of sensitivity to growth may be due, in part, to the length of the assessment; not only would including more test items likely improve reliability and validity, but it may also improve the measure's ability to make finer determinations of growth.

Discussion of Study 1

Findings of Study 1 extend the knowledge base on progress monitoring tools in two important ways. First, the present study provides preliminary information regarding the reliability and validity of two innovative progress monitoring measures; and second, the study extends the literature on two established measures, ORF and WIF.

Experimental measures. With respect to the experimental measures, the present study revealed that the Reading Fluency and Maze Sentences tasks are generally as reliable and

valid as other established progress monitoring measures, although correlation coefficients did not reach the criterion level established *a priori* for support of every hypothesis. Thus, it can be concluded that Reading Fluency and Maze Sentences are promising measures of early literacy skills, pending further development. Overall, the Reading Fluency task outperformed the Maze Sentences task; the two measures were comparable in terms of reliability, but Reading Fluency consistently demonstrated larger validity coefficients than did the Maze Sentences task. One would expect the Maze Sentences task to be a more sensitive (and therefore, reliable) task, given that students had a one in four chance of correctly answering each question, compared to a one in two chance on the Reading Fluency task. It is logical that more of the students' correct responses on Reading Fluency were due to chance compared to the Maze Sentences task. However, it is notable that there was a slightly larger range of scores on the Reading Fluency task than on the Maze Sentences task. Perhaps the yes/no format of Reading Fluency allowed students to progress through more items within the allotted amount of time, resulting in a relatively longer test. Indeed, students did answer significantly more items on the Reading Fluency task than on the Maze Sentences task. It is possible that more items allowed for finer distinctions to be made between students, which contributed to higher correlations between Reading Fluency scores and scores on ORF and WIF.

Currently, there is an obvious gap in the literature on group-based progress monitoring tools that are appropriate for students demonstrating early (i.e., first grade) reading skills. As discussed in Chapter 2, the tools currently being utilized either require

individual administration, which can be time consuming for groups of students, or require students to be able to read connected text fluently enough to be tested on passage reading. First grade readers, particularly during the first half of the year, may not be at a level in which they can read passages; therefore such measures often demonstrate floor effects (e.g., Fuchs et al., 1993). The results of Study 1 suggest that the experimental measures, with their focus on sight word reading, are worthy of continued research in order for further development. The correlations between the experimental and criterion measures provide preliminary evidence that students' relative reading achievement can be determined via the Reading Fluency and Maze Sentences tasks. Therefore, if benchmark norms were established, these measures could be utilized as quick screening or benchmark measures in order to identify those students who are at-risk of reading failure and who could, perhaps, benefit from more intensive assessment and intervention. However, preliminary results suggesting that Reading Fluency and Maze Sentences do not demonstrate sensitivity to growth over time indicate that the measures may not be suitable for ongoing progress monitoring in their present form. A more in-depth evaluation of the measures' sensitivity to growth was the focus of Study 2; therefore this topic will be revisited in the discussion of Study 2 and in the General Discussion of the entire project.

Criterion measures. Although the experimental measures were the primary focus of the present study, the results of Study 1 also add to the literature on ORF and WIF. The current literature base suggests that ORF and WIF are reliable and valid measures of students' reading skills (e.g., Fuchs, Deno, & Mirkin, 1984; Hintze & Shapiro, 1997), and

those findings were corroborated by the present study. ORF and WIF are also accepted progress monitoring tools on the basis that they have been shown to be sensitive to students' growth over time, but data from the present study do not entirely corroborate these findings. With respect to ORF, Fuchs et al. (1993) reported that a realistic expectation for first graders is growth by 2 words per week. However, the students in the present study gained, on average, 0.53 words per week. The discrepancy between findings of the present study and those reported in previous research may be due to methodological differences. First, different ORF probes were used in the two studies; Fuchs et al. used probes developed by Deno and colleagues (Deno, Deno, Marston, & Marston, 1987, cited in Fuchs et al.). Second, Fuchs et al. collected weekly ORF data once each week from October through April, whereas data in Study 1 were collected only four times between early December and late May. It is possible that the present study did not employ enough data points to yield a reliable slope that would be comparable to that demonstrated by Fuchs et al. That is, could we conclude, based on four data points, that our estimation of growth would be as reliable as if we had collected 20 or more data points?

The question of how many data points are necessary has arisen in the literature and results indicate that four data points do not result in the most reliable growth estimates. In one study, Fiala and Sheridan (2003) investigated the effect of parent involvement in reading and noted that the child with the longest intervention and therefore the most data points showed the most stable data. Allinder and Fuchs (1994) also questioned reliability of growth estimates in their evaluation of the effect of a 3-week break from school on students'

mathematics performance. The authors based most students' growth estimates on only 5 data points, but about one-third of the sample did have six data points. Analyses on these two groups of students suggested that about 67% of the variability was shared between trend lines based on five scores and those based on six scores. Thus, the authors stated that the slopes of the trend lines were not as reliable as they had anticipated, which limited the generalizability of their results.

Good and Shinn (1990) also investigated the question of how many data points produce a reliable estimate of slope. The results of their study suggested that 10 data points produce an acceptably reliable slope, but that 20 data points are significantly more accurate. Their results are corroborated by Fiala and Sheridan's (2003) study, in which those children with fewer than 10 data points showed less reliable and less interpretable data than the child with 13 data points. These findings suggest that the data collected in Study 1 were not sufficient for the reliable estimation of students' growth over time, and that one could expect the slopes of Students 1 and 2 in Study 2 to be more reliable than those of Students 3 and 4, simply because they have more data points during intervention.

Unlike ORF, WIF data from the present study do corroborate previous research. Fuchs, Fuchs, and Compton (2004) report WIF slopes of 0.90, 1.31, and 1.02 for the fall, spring, and entire year, respectively. Therefore, the slope of 1.12 obtained in the present study is very much in line with Fuchs et al.'s findings. The finding that WIF did model students' growth, however, begs us to revisit the question of why findings for ORF were so discrepant from previous research, whereas those for WIF were not. It is logical that if the

four data points were insufficient to yield a reliable slope, then we would have encountered similar problems with WIF as we did with ORF. One likely explanation is the amount of variability in ORF and WIF slopes. Indeed, the standard deviation of ORF slopes was more than twice the mean, whereas the standard deviation of WIF slopes was approximately one-half of the mean. A large amount of variability in the present data is not surprising in light of recent research revealing variability in ORF readability levels, and, as a result, ORF scores. Francis, Santi, Barr, Fletcher, Varisco, and Foorman (2008) pointed out that ORF passages differ dramatically in terms of reading level depending on which formula is utilized. (Indeed, the passages used in the present study ranged in Spache difficulty from 2.3 to 2.7). As a result, growth patterns look very different depending on the order in which passages are administered. Therefore, increases or decreases in ORF scores could mistakenly be attributed to gains or losses in skill, when in reality these fluctuations could simply be a function of more or less difficult passages. Francis et al. obtained more reliable data when they equated passages by converting raw scores (WCPM) into percentile ranks.

The finding that ORF passages vary in difficulty has significant implications for research and practice with ORF. For instance, the current study relied upon previous findings from studies on ORF to establish hypotheses about the experimental tasks; however, if ORF scores are more variable and less valid than previous research indicates, then our criterion in developing new, improved assessments is lower. Additional variability that is not thoroughly understood and accounted for would also reduce correlations between ORF and other measures, leading to the conclusion that the other measures were lacking in validity.

Furthermore, one of the key characteristics of progress monitoring tools is that individual probes are equivalent in terms of difficulty level. If probes differ, then changes in students' scores may be attributed to changes in the students' skills or responsiveness to interventions, when in reality, they are due to a change in relative difficulty of the task. That is, a child may appear to be making progress on ORF probes, but in reality the child may just have been administered relatively easy probes. Thus, until this issue has been dealt with appropriately (by developing a method to create equivalent probes or equate scores on different probes), ORF slope estimates will continue to include variance due to non-equivalent probes. With that in mind, the amount of variability within slopes in the present study is not surprising.

Limitations of Study 1

A number of limitations potentially reduced the degree to which valid conclusions can be drawn from Study 1. They are each discussed briefly, below, and will be further addressed in the discussion of Directions for Future Research.

Missing data. Missing data became a significant problem in Study 1, due in part to students' absences on data collection days and also to technical difficulties with the Discourse software. Unfortunately, due to the availability of students and the computer classroom, it was not feasible to make up missed assessments. It has been long-established that lower sample sizes reduce statistical power, thereby making it more difficult to detect a relationship between variables if it truly exists; therefore, with a larger sample, the experimental measures might have demonstrated stronger reliability and validity. Missing data also resulted in the inability to calculate a slope from every student's data; students were

excluded from growth analyses if they had more than one missing data point. Twenty-seven percent of Reading Fluency slopes and 31% of Maze Sentences slopes were calculated based on only three data points instead of four, which likely lowered the reliability of growth estimates. Difficulties with Discourse software resulted in the loss of approximately one-third of the data from the first and second assessment periods.

Timeline of the study. It was originally intended for data collection to begin in early October (when the school system would allow us to begin); however, difficulties with the installation of the Discourse software resulted in data collection being pushed back until early December. Therefore, three of the four assessments occurred during the spring semester; originally, all four assessments were to have been evenly distributed across the full academic year. Thus, we were not able to calculate growth estimates across the entire year. Past research has indicated that more reliable estimates of growth have been obtained during the fall semester (Fuchs, Fuchs, & Compton, 2004), but we were not able to obtain data across the fall semester in the current study.

An additional limitation to the timeline of Study 1 is the fact that data were collected only four times throughout the study. Study 1 was developed as a study of benchmark assessments and mimics how benchmark screening is conducted in practice; however, as previously discussed, the methodology resulted in fewer data points than are necessary to achieve reliable estimates of growth. If data had been collected more frequently, perhaps the experimental measures would have been more able to demonstrate sensitivity to students' progress. The limitation was addressed by conducting Study 2, in which the progress of a

small sample of students was monitored over a period of 17 weeks to further evaluate whether the experimental measures could reliably model students' growth.

Definition of test-retest reliability. Test-retest reliability is typically defined as the correlation between scores on one probe administered at two different points in time. However, in the present study, different probes were administered at each of the four assessment periods. Therefore, the test-retest reliability discussed in the present study is perhaps better conceptualized as delayed alternate forms reliability. The different way of describing test-retest reliability makes it difficult to compare the results of the present study with findings of previous research. Furthermore, it is likely that the test-retest reliability coefficients presented in the present study are lower than they would have been had true test-retest reliability analyses been conducted on the same probe at two different times. That is, it is logical to predict that the correlation between scores on the same probe at two different points in time would be larger than that between scores on two different probes at two different points in time. Thus, although the findings of the present study are realistic in that our methodology mimicked the way the measures would be used in real life, they might underestimate the test-retest reliability that would have been found had the traditional definition been used.

Study 2

Hypothesis 1

In Study 2, it was hypothesized that the Reading Fluency (Hypothesis 1A) and Maze Sentences (Hypothesis 1B) tasks would demonstrate sensitivity to intraindividual change in

response to intervention, such that students' scores on each task would show positive linear growth from baseline levels. However, results did not support Hypothesis 1A and revealed contradictory evidence in support of Hypothesis 1B. Only one of the four students in the sample (Student 4) showed improvement on the Maze Sentences task; his mean scores and slopes of improvement both increased from baseline to intervention phases. Although data from Student 4 do provide some evidence that the Maze Sentences task is sensitive to intraindividual change, the fact that data from the other three participants did not support the hypothesis makes it difficult to conclude that the Maze Sentences task is sensitive to intraindividual changes in response to intervention.

There are a number of factors that likely contributed to failure to support the hypotheses. First, as previously discussed with respect to Study 1, the experimental measures were developed to be very brief tasks, in an effort to make them more feasible for teachers to implement on a regular basis. However, the brevity of the measures likely reduced their ability to discriminate between fine changes in growth.

Second, variability among scores made the interpretation of each child's graphed data difficult, and reduced the extent to which slopes could be interpreted. For instance, the standard deviations of Reading Fluency and Maze Sentences scores for Student 2 are, in some cases, greater than his mean scores. This indicates the presence of significant variability, especially when compared to the variability present within criterion measures. One would expect greater variability among scores during the intervention phase, in which scores are predicted to improve, but not in the baseline phase, during which scores should be

stable. However, Maze Sentences score standard deviations tended to be greater during baseline than during intervention. One possible explanation for this is that students may have guessed on more items during the baseline phase, which would likely have increased variance among scores.

Additionally, part of this variance is likely due to the attention and motivation levels of the children. Reports made by the students' teachers and *Sound Partners* tutors as well as observations made by the primary investigator suggest that some of the students' attentional and motivational difficulties negatively affected their performance in the study. For instance, during baseline data collection and the beginning of the intervention phase, Child 2 was observed by research assistants to be purposefully selecting the *in*correct answers to Reading Fluency and Maze Sentences items; this activity seemed to amuse the child. The child did not respond to research assistants' prompts to try to answer each item correctly; however, once research assistants had mentioned his response pattern to his teacher, the child showed an increase in items answered correctly (as is evident on his graph at week 13). Throughout the assessments, the children had to be redirected to maintain attention to the task (instead of playing with zippers, shoelaces, or the keyboard and getting out of their seats). In such a brief, timed task, it is essential for students to sustain their attention to the task, for even a quick redirection could expend valuable seconds that a child could have spent reading and responding to an item. It is plausible that scores could fluctuate from one assessment to the next with a child's ability to attend to the task, thereby muddying interpretations of growth.

A third factor contributing to the experimental measures' demonstration of growth over time is the question of whether or not the students did, in reality, improve their reading skills during the course of the intervention. That is, given the type and amount of intervention provided, could we reasonably expect to see growth? Due in part to the aforementioned attentional difficulties, some of the children progressed through more Sound Partners lessons than did others. Child 3, for instance, demonstrated the most difficulty sitting still and attending to assessment and intervention tasks; this led, in part, to her completion of only three lessons across 11 intervention sessions. Thus, one would not expect her to have grown as much as a result of the intervention than a child who had completed 10 or more lessons.

Another important consideration is not only the amount of progress made in terms of number of lessons completed, but also the content of the lessons completed. For instance, Child 1 completed 21 lessons, but these were the first lessons of the program, which focused on single-letter sounds, consonant-vowel-consonant words, and extremely short, repetitive connected text. Perhaps this child would not be expected to make as much progress on the experimental measures as a child who had focused on later Sound Partners lessons, which place more emphasis on sight words and reading higher-level connected text. Indeed, Student 2 completed the highest-level lessons and his ORF data showed the most impressive and clear-cut improvement over baseline levels of all of the participants.

It is also possible that the intensity of intervention provided in Study 2 was not adequate to result in measurable growth. Previous research on *Sound Partners* has demonstrated that students who received 30 minutes of Sound Partners tutoring four days a

week for 23 weeks made significant gains on multiple reading measures, including the WRAT-R reading and a Dolch word list, when the program was implemented with high integrity (Vadasy, Jenkins, Antil, Wayne, & O'Connor, 1997). The length and intensity of intervention in Vadasy et al.'s study was substantially greater than that of the present study. Therefore, perhaps it was an unrealistic expectation that we would have seen each of the four students in the present study make detectable progress given only 7 to 11 weeks' worth of intervention, particularly with a maximum of 3 25-minute tutoring sessions each week. Furthermore, it is more likely, given the length of Vadasy et al.'s intervention, that students completed most if not all of the *Sound Partners* lessons. Vadasy et al. did not indicate which or how many lessons students in their study completed, but it seems logical that all or close to all 100 lessons would have been reached, given that students were provided with 92 tutoring sessions (however, the authors did not describe the impact of student or tutor absences on the implementation of intervention; 92 sessions represents a maximum).

In order to gain more insight into the question of whether or not the students had improved their reading skills throughout the study, one can compare their changes in performance over time on the experimental measures to that on the criterion measures. Indeed, this was the primary purpose of including the criterion measures in Study 2. However, only Child 2 demonstrated clear improvements in both ORF and WIF; Children 1, 3 and 4 showed improvements in ORF, but not as clearly in WIF. With criterion measures that are difficult to interpret, it is difficult to draw conclusions about what we could reasonably expect to learn from the students' progress on experimental measures. In the case

of Student 2, it seems logical to expect his Reading Fluency and Maze Sentences scores to mirror the pattern of improvement that is evident in ORF and WIF scores; however, data did not follow the expected pattern, possibly due to the child's tendency to select the incorrect responses on the experimental measures.

Discussion of Study 2

Although neither Hypothesis 1A nor 1B was supported, the results of Study 2 contribute to the progress monitoring literature in several ways. First, the fact that results provided limited evidence suggesting that the Reading Fluency and Maze Sentences tasks are sensitive to individual students' growth is important information in light of the increasing emphasis on RTI for educational decision making. The results of Study 2 suggest that more research and development should take place with the experimental measures before they are used for individual students' educational planning. A second, related, issue is that of whether or not growth on a particular measure would be expected given a particular intervention. That is, if the intervention program and assessment measures are different, can generalization from the intervention to the measure be expected? As previously discussed, it might not have been a reasonable expectation for some of the participants of Study 2 to show improvement, given that the lessons they completed focused on lower-level skills than did the assessments. Therefore, the present data point out the acute need for progress monitoring measures of pre-literacy and very early literacy skills. Clearly this is a topic that should be further investigated in order for researchers and educators to successfully implement RTI.

A third observation made in Study 2 that is relevant to the progress monitoring literature centers on the appropriateness of slope as an indicator of students' progress. When interpreting data from the experimental and criterion measures, it becomes evident that slopes of improvement are not necessarily sufficient information for forming accurate judgments of individual students' progress. For instance, Student 2's mean ORF scores improved from baseline to intervention and inspection of his graphed data indicate no overlapping data points, such that all of his intervention-phase ORF scores are higher than his baseline-phase ORF scores. However, his intervention slope indicates a decrease in scores. Thus, an individual looking at only slope of improvement would determine that this child was not making progress during the intervention phase, whereas a more thorough interpretation of his data clearly shows that he did, indeed, improve ORF scores from baseline to intervention. This finding is consistent with recent research on the utility of CBM slopes, which suggests that variability in data points introduces error that clouds the interpretation of slopes. For instance, Hintze and Christ (2004) stated that data collected over shorter periods of time tend to have more measurement error, which could have the result that "practitioners have no way of knowing whether decisions that are made are more a product of the actual construct of interest (i.e., slope) or of measurement error brought about by brief progress monitoring periods or variations in the material used for assessment" (p. 205).

Limitations of Study 2

As with Study 1, a number of limitations are potential threats to the validity of the results of Study 2 and will now be addressed.

Student and tutor absences from intervention and assessment. Perhaps the most significant limitation of Study 2 was that students' absences limited the amount of data that could be collected and reduced the number of *Sound Partners* tutoring sessions. During the 17 weeks of baseline and intervention, the first grade classes went on a number of field trips and had special assemblies during their designated intervention and assessment days and times. Such absences resulted in a number of missing data points, which may have decreased the ability to interpret students' growth over time. Missing data was particularly a problem with Students 3 and 4; these two students had the least amount of time in the intervention phase of the study, therefore fewer weeks' worth of data collection during the intervention phase. Absences also pulled all four of the students out of several tutoring sessions, which reduced the amount of intervention that the students received. Additionally, occasional absences on the part of tutors (due to illness, injury, and end-of-semester scheduling conflicts) further reduced the number of tutoring sessions that took place, such that some of the students only met with a tutor once a week toward the end of the study, rather than the scheduled three times per week. If students were not receiving the intervention, then we would be much less likely to see growth if growth was occurring as a result of intervention.

Characteristics of the sample. A second limitation of Study 2 has to do with characteristics of the participants that made them less likely to (a) benefit from the

intervention, and (b) make measurable progress on assessments. Of the four students who participated, two had significant difficulties with attention and motivation, and a third student, who was learning English as a second language, had reading difficulties that appeared to exceed the scope of the Sound Partners intervention. Thus, three of the four students sampled presented with difficulties above and beyond having reading delays that appeared to impact their acquisition of reading skills. Furthermore, the difficulties that two of the children had focusing their attention made the assessments quite demanding for them and likely reduced the validity of results. That is, incorrect responses may have been just as much a reflection of the inability to attend to the task as they were a reflection of poor reading skills. Therefore, the experimental measures' lack of sensitivity to growth might be due to invalid responses rather than to the psychometric properties of the measures.

Unstable baseline data. A third limitation of Study 2 relates to achieving stable baseline data prior to beginning intervention. Failure to accomplish this was partially due to the variability among measures and partially to methodological limitations. Particularly with Students 3 and 4, the school schedule and upcoming spring break created an urgency to begin intervention before a stable baseline had been established. It seemed unethical to withhold intervention until after the students came back from spring break in mid-April, which would leave them with only six weeks of intervention. Thus, students were provided *Sound Partners* tutoring before stable baselines had been established, which further complicated analyses of whether or not improvements were made during intervention.

General Discussion

The aim of the current project was to establish the reliability and validity of new, computerized progress monitoring measures and to determine whether the measures were sensitive to interindividual and intraindividual growth over time. Taken together, Studies 1 and 2 provide preliminary evidence that the experimental Reading Fluency and Maze Sentences tasks are as reliable and valid as other well-established progress monitoring tools (although reliability and validity coefficients did not reach the criterion level set for the present study for every analysis). Therefore, Reading Fluency and Maze Sentences show promise as measures of early literacy skills. Studies 1 and 2 also indicate that the measures are not as sensitive to change over time as are ORF and WIF. These outcomes will now be considered within the broader contexts of research and practice.

Reading Fluency and Maze Sentences as Screening/Benchmark Assessments

In light of results of Studies 1 and 2, Reading Fluency and Maze Sentences appear to be more appropriate for use as screening and benchmark measures, useful in the quick identification of students who are at-risk for reading failure, rather than for use as progress monitoring tools to be administered frequently to assess changes in students' performance over time. Reading Fluency and Maze Sentences meet some but not all of the core standards used for the evaluation of progress monitoring tools, discussed in Chapter 2 and provided in more detail in Appendix A. Both Reading Fluency and Maze Sentences demonstrated reliability and validity estimates that are comparable to established progress monitoring measures (although additional research should continue to develop these measures to higher

standards of reliability and validity in order to approve the confidence with which conclusions can be drawn from measurement data) and they are available in alternate forms. Additionally, they were developed as group-based measures, which enhances their feasibility, although a delivery system that is more user-friendly for this purpose than the Discourse Groupware Classroom should be developed (see below for further discussion of Discourse). However, Reading Fluency and Maze Sentences did not demonstrate sensitivity to growth; therefore it cannot be concluded at this time that they are appropriate for ongoing progress monitoring.

The fact that Reading Fluency and Maze Sentences scores correlated with criterion measures indicates that a student's relative standing on the experimental measures is similar to his or her relative standing on the criterion measures. That is, Reading Fluency and Maze Sentences appear to be capable of discriminating low-performing from average- and high-performing students. Therefore, as screening/benchmark measures, Reading Fluency and Maze Sentences could be administered up to four times throughout a school year to all students in order to identify those individuals who may be at risk for reading failure. DIBELS are utilized in this fashion, and the developers of DIBELS have published benchmark norms for each grade indicating at-risk, some-risk, and low-risk levels of performance (Good & Kaminski, 2002). As discussed in Chapter 2, DIBELS measures are individually administered, therefore assessing an entire classroom, grade, or school could be difficult in terms of time and resources. However, it would be much more feasible for educators to administer a group-based, computerized measure, such as Reading Fluency or Maze

Sentences, in order to quickly identify students in need of additional instruction and monitoring. However, before Reading Fluency and Maze Sentences can be utilized for benchmark assessments, one must develop benchmark norms and ensure that the measures produce stable means.

Identifying low-performing or at-risk students using benchmark assessments and targeting them for additional/alternative instruction is part of the methodology of the RTI approach to identifying students at risk of academic difficulties (e.g., Compton, Fuchs, Fuchs, & Bryant, 2006; D. Fuchs, Fuchs, & Compton, 2004). Within an RTI framework, those students who perform in the bottom 15-20% or below a specified benchmark score on screening Reading Fluency and Maze Sentences assessments could be targeted for supplemental or alternative instruction. Then, their progress would be monitored regularly using an individually-administered measure that is more sensitive to subtle changes in the performance of individual students (such as WIF).

Feasibility of Implementing Reading Fluency and Maze Sentences within RTI

The Reading Fluency and Maze Sentences tasks were developed with feasibility of administration in mind in order to make educators more likely to utilize progress monitoring measures. Reading Fluency and Maze Sentences probes are brief and cost-effective and are able to be administered to a group of students at once, which further limits the time taken away from instruction. The Discourse Groupware Classroom was selected for the administration of these measures based on previous research suggesting that it would be useful because it (a) is an authoring program, which would allow us to create our own

probes, (b) allows for all students to be assessed simultaneously, with the educator/researcher in control of the master computer, and (c) automatically scores each item and probe for each child (e.g., Shin, 2000). Thus, Discourse seemed to have the qualities one would look for in a software program to be used to administer the experimental tasks. However, Discourse was not originally designed for the purpose of class-wide progress monitoring, therefore difficulties were encountered that hampered research efforts and reduced the feasibility of administering Reading Fluency and Maze Sentences.

First, some of the Reading Fluency and Maze Sentences probes administered during the first two assessment periods were left unscoreable due to either computer or human errors in administration; for unknown reasons, the same probes were administered twice to some classes. These errors affected the present study by reducing sample sizes, in some cases, by approximately one-third. Additionally, a mismatch between stimulus pictures and answer choices on students' computer screens resulted in the inability to administer and, thus, evaluate the Dolch Word Recognition task. This problem was caused by differences in the sizes of picture files, which affected the speed of delivery of the pictures. Discourse was not intended for assessments requiring precise timing; therefore, modifications addressing this issue may help to improve the feasibility with which Discourse can be used for progress monitoring assessments. Clearly, if Discourse was to be utilized by classroom teachers on a regular basis, one would want to make this and similar modifications to the program to enhance its usability as a progress monitoring assessment delivery system. Such modifications should ensure that the program is consistently administering and scoring items

properly. Minor glitches in the administration or scoring of progress monitoring measures could suggest to teachers that a child is progressing within normal limits, when perhaps he or she actually is in need of an intervention or change in instruction.

Directions for Future Research

The present study provides preliminary evidence supporting innovative screening/benchmark assessments that, with further development, may lead to improved benchmarking and ongoing progress monitoring measures. Should the limitations previously discussed with respect to Studies 1 and 2 be addressed in future research, perhaps data will provide stronger support for the Reading Fluency and Maze Sentences tasks. At this time, there are several directions future research could go in the establishment of these measures.

Develop Benchmark Norms

As previously discussed, the Reading Fluency and Maze tasks do show promise as screening/benchmarking assessments. However, before the measures can be utilized for benchmarking, benchmark norms must be developed. Future research should establish norms indicating scores falling in the average, some-risk, and at-risk ranges, similar to those established for DIBELS.

Increase the Amount of Data Collected

It would be wise of future researchers to included larger samples of students in order to increase statistical power. Additionally, had the present study started with a larger sample, perhaps missing data due to students' absences would not have been as detrimental. Contingency plans should also be made prior to beginning the study for how to handle

student and tutor absences, as scheduling limitations prevented make-up assessments in the present study.

Future studies should also incorporate more frequent data collection. As previously discussed, it was unlikely that Study 1 results would reveal student growth given the limited number of data points. Research suggests that 10 – 20 data points are necessary to achieve reliable estimates of growth (Good & Shinn, 2000); therefore weekly data collection across the majority of the academic year would be advised, if time and resources permit. Furthermore, when conducting intervention studies similar to Study 2, it is advisable to provide a long period of intensive intervention, as was the case in Vadasy et al. (1997). It was unclear in the present study as to whether or not students received enough of the *Sound Partners* intervention that we could reasonably expect to see growth.

Establish Criteria for Participation in Intervention Study

One of the limitations encountered in Study 2 was that characteristics of the participants made it less likely that they would respond to the intervention and that their responses to assessments would be valid reflections of their true abilities. With the prevalence of attentional deficits and other behavioral problems, this issue is encountered commonly in the school systems and therefore increases the external validity of results. However, these issues made it difficult to assess the effectiveness of *Sound Partners* and the psychometric properties of the experimental measures. Thus, future small-n studies on Reading Fluency and Maze Sentences may consider including criteria for participation in the study, such as the absence of attentional deficits. Then, once the measures have been more

fully developed, additional studies could be conducted with samples of students more representative of the student body of typical elementary schools. It is also possible that the participants of Study 2 were true non-responders to intervention. The school they attended already provided students with intensive reading interventions; therefore, those students who continued to struggle may be the students who would be labeled, within an RTI framework, as non-responders, and identified for more thorough evaluation (e.g., Fuchs et al., 2002).

Ensure Equivalent Forms of Criterion Measures

Another difficulty encountered in the present study, which has also been noted in the progress monitoring literature, is the non-equivalent forms of ORF probes. Future studies should account for this, either by replacing ORF with another progress monitoring measure or by ensuring equivalent forms through an equating process such as the one suggested by Francis et al. (2008). Unless equivalent forms have been ensured, one cannot attribute changes in student progress to changes in skill level.

Conduct Item Analysis on Experimental Measures

In the present study, items on the experimental measures were written using first grade level Dolch words, such that every test item *should* have been on a first grade reading level, at least by one metric. However, data were not analyzed to determine which, if any, items on the experimental measures were too easy or too difficult. It is quite possible that there were items on these measures that either all students missed or all answered correctly; any items that do not differentiate between high- and low-achieving students are unnecessary to the measure and should therefore be weeded out. Future research should conduct a

thorough item analysis to ensure that at least some students are able to correctly answer each of the items.

Modify/Develop Assessment Software

Due to the technical difficulties encountered in the present study, it is advised that future researchers modify the Discourse software or use/develop software that better fits the requirements for the progress monitoring tasks. Discourse was not originally intended for progress monitoring, but if the software was modified into a more user-friendly and reliable system, then the Reading Fluency and Maze Sentences tasks would be quite feasible for teachers to implement. The measures are brief, as each probe only takes one minute; however, additional time was necessary to log students into their computers and into the software program. Perhaps future research could determine methods for shortening this process or for allowing teachers to quickly log all students into the program before beginning the assessment. Additionally, if a modified Discourse system was utilized, then teachers could monitor all of their students' responses simultaneously and the software program would score students' responses automatically, dramatically streamlining the assessment process for teachers. It is logical that the development of this type of computer software should be the first step that subsequent research takes in making progress monitoring measures more feasible for teachers to implement.

Implications for Practice

The findings of the present study have several implications for practice, particularly as school systems begin to adopt RTI as a method of instructional decision-making. First,

results demonstrate that the Reading Fluency and Maze Sentences tasks are promising measures for benchmark assessments of students' reading skills. Within an RTI framework, these group-based measures could be utilized for regular class-wide screenings, with individualized follow-up monitoring of students who are identified as at-risk on benchmark measures. Second, the present study demonstrates the efficiency of utilizing computer-based assessments for benchmark screening; in Study 1, all first grade students in a class were assessed simultaneously in less than half of one class period. Provided that the previously discussed software program modifications are made, it is reasonable to predict that benchmark assessments could be administered within even less time and with greater ease on the part of the teacher/examiner.

Third, Studies 1 and 2 add to the literature providing evidence of the utility and stability of WIF. WIF scores demonstrated strong stability between assessment periods as well as moderate to strong validity with ORF and the experimental measures. Furthermore, WIF was the only measure sensitive to students' growth over time, evidenced by a positive, non-zero slope. Taken together, these data provide evidence supporting the use of WIF for progress monitoring of students' reading skills. Based on these data and recent literature (Fuchs, Fuchs, & Compton, 2004), WIF appears to be an appropriate measure for monitoring students identified during benchmark screening as at-risk for reading failure.

Studies 1 and 2 together provide an excellent illustration of the fact that even if a measure is reliable and valid by conventional indices, it still may not be able to adequately model students' growth over time. Thus, educators must be cautious in their selection of

progress monitoring measures. Furthermore, educators must consider not only whether the measure is appropriate for progress monitoring in general, but also whether the measure is appropriate for monitoring of the target skills. That is, is it reasonable to expect growth on the selected measure(s) given the skills that were targeted in intervention? It is hoped that the present study will bring these and similar issues into the forefront of researchers' and educators' minds as they make the shift from the traditional discrepancy approach to RTI.

In summary, the present study has evaluated two innovative measures, Reading Fluency and Maze Sentences, which were developed for purposes of ongoing progress monitoring of first grade literacy skills. Results of two studies suggest that the measures may be more appropriate as screening/benchmark assessment than for progress monitoring. That is, the measures demonstrated that they have the potential to be reliable and valid measures of early literacy skills, but they were unable to model students' growth over time. The present study provides preliminary evidence that the Reading Fluency and Maze Sentences tasks may be utilized to identify those students in need of additional instruction or intervention and more frequent, accurate monitoring. As screening/benchmark measures, the Reading Fluency and Maze Sentences tasks could be utilized as the first step in a RTI approach to instructional decision making.

REFERENCES

Allinder, R. M., & Fuchs, L. S. (1994). Alternative ways of analyzing effects of a short school break on students with and without disabilities. *School Psychology Quarterly, 9,* 145-160.

Arter, J. A., & Jenkins, J. R. (1977). Examining the benefits and prevalence of modality considerations in special education. *Journal of Special Education, 11*, 281-298.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Kappa Deltan,* 80, 139-148.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671-684.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98,* 394-409.

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

Coyne, M. D., Kame'enui, E. J., & Simmons, D. C. (2004). Improving beginning reading instruction and intervention for students with LD: Reconciling "All" with "Each." *Journal of Learning Disabilities, 37,* 231-239.

Deno, S. L. (1986). Formative evaluation of individual student programs: A new role for school psychologists. *School Psychology Review, 15,* 358-374.

Deno, S. L. (1990). Individual differences and individual difference: The essential difference of special education. *Journal of Special Education, 24*, 160-173.

Deno, S. L. (2002). Problem solving as "Best Practice." In Thomas, A. and Grimes, J. *Best practices in school psychology, IV* (pp. 37-55). The National Association of School Psychologists: Bethesda, MD.

Deno, S. L., Mirkin, P. K., Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49,* 36-45.

Elliott, J., Lee, S. W., & Tollefson, N. (2001). A reliability and validity study of the Dynamic Indicators of Basic Early Literacy Skills-Modified. *School Psychology Review, 30,* 33-49.

Fiala, C. L., & Sheridan, S. M. (2003). Parent involvement and reading: Using curriculum-based measurement to assess the effects of paired reading. *Psychology in the Schools, 40,* 613-626.

Fletcher, J. M., Coulter, W. A., Reschly, D. J., & Vaughn, S. (2004). Alternative approaches to the definition and identification of learning disabilities: Some questions and answers. *Annals of Dyslexia, 54,* 304-331.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90,* 37-55.

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46,* 315-342.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88,* 3-17.

Freeze, D. R. (1988). Microcomputers in special education. *Canadian Journal of Special Education, 4,* 9-22.

Fuchs, L. S. (1989). Evaluating solutions: Monitoring progress and revising intervention plans. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 153-181). New York: Guilford.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal, 21,* 449-460.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*, 199-208.

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21,* 45-59.

Fuchs, L. S., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *School Psychology Review, 28,* 659-672.

Fuchs, L. S., & Fuchs, D. (2002). Curriculum-based measurement: Describing competence, enhancing outcomes, evaluating treatment effects, and identifying treatment nonresponders. *Peabody Journal of Education, 77*(2), 64-84.

Fuchs, L. S., & Fuchs, D. (2004). Determining adequate yearly progress from kindergarten through grade 6 with curriculum-based measurement. *Assessment for Effective Intervention, 29*(4), 25-37.

Fuchs, D., Fuchs, L. S., & Compton, D. L. (2004). Identifying reading disabilities by responsiveness-to-instruction: Specifying measures and criteria. *Learning Disability Quarterly, 27,* 216-227.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children, 71,* 7-21.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement, using a reading maze task. *Exceptional Children, 58,* 436-451.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22,* 27-48.

Fuchs, L. S., Fuchs, D., & Speece, D. L. (2002). Treatment validity as a unifying construct for identifying learning disabilities. *Learning Disability Quarterly, 25,* 33-45.

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research and Practice, 18*, 157-171.

Fuchs, D., & Young, C. L. (2006). On the irrelevance of intelligence in predicting responsiveness to reading instruction. *Exceptional Children, 73,* 8-30.

Good, R. H, & Kaminski, R. A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills (6th ed.)*. Eugene, OR: Institute for the Development of Educational Achievement. Retrieved August 25, 2007 from http://dibels.uoregon.edu/

Good, R. H., & Shinn, M. R. (1990). Forecasting accuracy of slope estimates for reading. Curriculum-Based Measurement: Empirical evidence. *Behavioral Assessment, 12,* 179-193.

Hargreaves, M., Shorrocks-Taylor, D., Swinnerton, B., Tait, K., & Threlfall, J. (2004). Computer or paper? That is the question: Does the medium in which assessment questions are presented affect children's performance in mathematics? *Educational Research, 46,* 29-42.

Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing. *School Psychology Review, 32,* 541-556.

Hintze, J. M., & Shapiro, E. S. (1997). Curriculum-based measurement and literature-based reading: Is curriculum-based measurement meeting the needs of changing reading curricula? *Journal of School Psychology, 35,* 351-375.

Hintze, J. M., Shapiro, E. S., & Lutz, J. G. (1994). The effects of curriculum on the sensitivity of curriculum-based measurement in reading. *The Journal of Special Education, 28,* 188-202.

Hoffman, J. V., & Rutherford, W. L. (1984). Effective reading programs: A critical review of outlier studies. *Reading Research Quarterly, 20,* 79-92.

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children, 59,* 421-433.

Jenkins, J. R., Vadasy, P. F., Firebaugh, M., & Profilet, C. (2000). Tutoring first-grade struggling readers in phonological reading skills. *Learning Disabilities Research & Practice, 15,* 75-84.

Juel, C. (1994). *Learning to read and write in one elementary school.* New York: Springer-Verlag.

Kame'enui, E. J. (2002). An analysis of reading assessment instruments for K-3. [Final report for Members of the Assessment Committee, Institute for the Development of Educational Achievement].

Kavale, K. A. (2001). Discrepancy models in the identification of learning disability. Retrieved October 26, 2007 from http://www.air.org/ldsummit/download/Kavale%20Final%2008-10-01.pdf

Kavale, K. A., Fuchs, D., & Scruggs, T. E. (1994). Setting the record straight on learning disability and low achievement: Implications for policymaking. *Learning Disabilities Research and Practice, 9*, 70-77.

King, R., & Torgesen, J. K. (2001). *Improving the effectiveness of reading instruction in one elementary school: A description of the process*. (Tech. Rep. No. 3). Florida Center for Reading Research.

Lyon, G. R., & Fletcher, J. M. (2001). Early warning system. Retrieved Aug. 24, 2004 from www.educationnext.org/20012/22.html

Lyon, G. R., Fletcher, J. M., Shaywitz, S. E., Shaywitz, B. A., Torgesen, J. K., Wood, F. B., et al. (2001). Rethinking learning disabilities. In C. E. Finn, A. J. Rotherham, & C. R. Hokanson (Eds.), *Rethinking special education for a new century* (pp. 259-287). Washington, DC: Progressive Policy Institute.

Marston, D., Fuchs, L. S., & Deno, S. L. (1986). Measuring pupil progress: A comparison of standardized achievement tests and curriculum-related measures. *Diagnostique, 11,* 77-90.

Marston, D., Tindal, G., & Deno, S. L. (1984). Eligibility for learning disability services: A direct and repeated measures approach. *Exceptional Children, 50,* 554-556.

National Center for Education Statistics (2007). *The Nation's Report Card: Reading 2007.* US Department of Education.

Reschly, D. J., & Ysseldyke, J. E. (2002). Paradigm shift: The past is not the future. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology, IV* (pp. 3-20). Bethesda, MD: The National Association of School Psychologists.

Rickelman, R. J., Henk, W. A., & McKenna, M. G. (1991). Computerized reading assessment: Its emerging potential. *The Reading Teacher, 44,* 692-693.

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., & Foorman, B. R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology, 96,* 265-282.

Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagne, and M. Scriven (Eds.), *Perspectives of curriculum evaluation.* Chicago: Rand McNally.

Shaywitz, S. E., Fletcher, J. M., Holahan, J. M., Schneider, A. E., Marchione, K. E., Stuebing, K. K., et al. (1999). Persistence of dyslexia: The Connecticut longitudinal study at adolescence. *Pediatrics, 104*, 1351-1359.

Shin, J. (2000). Predicting classroom achievement from active responding on a computer-based groupware system. *Remedial and Special Education, 21,* 53-61

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *Journal of Special Education, 34,* 164-172.

Shinn, M. R., & Bamonto, S. (1998). Advanced applications of curriculum-based measurement: "Big Ideas" and avoiding confusion. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 1-31). New York: Guilford Press.

Shinn, M. R., Shinn, M. M., Hamilton, C., & Clarke, B. (2002). Using curriculum-based measurement in general education classrooms to promote reading success. In M. R. Shinn, H. M. Walker, & G. Stoner (Eds.), *Interventions for academic and behavior problems II: Preventive and remedial approaches* (pp. 113-142). Washington DC: National Association of School Psychologists.

Snow, R. E. (1989). Aptitude-treatment interaction as a framework for research on individual differences in learning. In P. L.Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 13-59). New York: W. H. Freeman and Company.

Speece, D. L. (1990). Aptitude-treatment interactions: Bad rap or bad idea? *Journal of Special Education, 24*, 139-148.

Speece, D. L., & Case, L. P. (2001). Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology, 93,* 735-749.

Stage, S. A., Abbott, R. D., Jenkins, J. R., & Berninger, V. W. (2003). Predicting response to early reading intervention from verbal IQ, reading-related language abilities, attention ratings, and verbal IQ-Word reading discrepancy: Failure to validate discrepancy method. *Journal of Learning Disabilities, 36,* 24-33.

Stanovich, K. E. (1999). The sociopsychometrics of learning disabilities. *Journal of Learning Disabilities, 32,* 350-361.

Stanovich, K. E., & Siegel, L. S. (1994). Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology, 86,* 24-53.

Steubing, K., Fletcher, J., LeDoux, J., Lyon, G. R., Shaywitz, S., & Shaywitz, B. (2002). Validity of IQ-discrepancy classifications of reading disabilities. *American Educational Research Journal, 39,* 469-518.

Stoner, G., Carey, S. P., Ikeda, M. J., & Shinn, M. R. (1994). The utility of curriculum-based measurement for evaluating the effects of methylphenidate on academic performance. *Journal of Applied Behavior Analysis, 27,* 101-113.

Stoner, G., Scarpati, S. E., Phaneuf, R. L., & Hintze, J. M. (2002). Using curriculum-based measurement to evaluate intervention efficacy. *Behavior Psychology in the Schools: Innovations in Evaluation, Support, and Consultation, 1,* 101-112.

Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research & Practice, 15*, 55-64.

Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of School Psychology, 40,* 7-26.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1994). Longitudinal studies of phonological processing and reading. *Journal of Learning Disabilities, 27*, 276-286.

Vadasy, P. F., Jenkins, J. R., Antil, L. R., Wayne, S. K., & O'Connor, R. E. (1997). The effectiveness of one-to-one tutoring by community tutors for at-risk beginning readers. *Learning Dsiability Quarterly, 20,* 126-139.

Vadasy, P. F., Sanders, E. A., Peyton, J. A., & Jenkins, J. (2002). Timing and intensity of tutoring: A closer look at the conditions for effective early literacy tutoring. *Learning Disabilities Research and Practice, 17,* 227-241.

Vadasy, P., Wayne, S., O'Conner, R., Jenkins, J., Pool, K, Firebaugh, M., et al. (2005). *Sound partners: A tutoring program in phonics-based early reading.* Boston: Sopris West.

Varnhagen, S., & Gerber, M. M. (1984). Use of microcomputers for spelling assessment: Reasons to be cautious. *Learning Disability Quarterly, 7*, 266-270.

Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research and Practice, 18*, 137-146.

Vaughn, S., Linan-Thompson, S., & Hickman, P. (2003). Response to instruction as a means of identifying students with reading/learning disabilities. *Exceptional Children, 69,* 391-409.

Ysseldyke, J. E., & Sabatino, D. A. (1973). Toward validation of the diagnostic-prescriptive model. *Academic Therapy, 8,* 415-422.

APPENDICES

APPENDIX A

Progress Monitoring Tools Compared on Core Standards

Tool	Standards						
	Earliest Grade Assessed	Reliability	Validity	Alternate Forms	Sensitive to Growth	Feasible (cost & time)	Group-Based
Established Measures							
CBM Maze	1.5	.61 - .91	.65 - .76	✓	✓[1]	✓[2]	Yes
ORF	1.5	.93 - .96	.54 - .92	✓	✓[3]	✓[4]	No
DIBELS	K	.80 - .90	.60 - .70	✓	Variable[5]	✓[6]	No
WIF	1	.88 - .97	.68 - .71	✓	✓[7]	✓[8]	No
Experimental Measures							
Reading Fluency	1	.60 - .88	.71 - .89	✓	No	Needs development	Yes
Maze Sentences	1	.61 - .88	.48 - .86	✓	No	Needs development	Yes

[1]Fuchs, and Fuchs (1992) present technical information on the Maze task based on data collected by Fuchs, Fuchs, Hamlett, and Ferguson (1992) in a study employing third-grade students and their teachers. Fuchs and Fuchs suggest that the task's small ratio between slope and SEE allows for better detection of growth when data are graphed, compared to other CBM tasks.

[2]Each probe lasts 2.5 minutes; there is no cost to administer computerized probes.

[3]Support for ORF's sensitivity to students' growth has been provided by Marston, Fuchs, and Deno (1986) via correlations between growth on ORF and teachers' judgments of students'

growth, as well as by Fuchs, Fuchs, Hamlett, Walz, and Germann's (1993) investigation of how much growth on ORF is made by students in grades two through six.

[4]Each ORF probe takes 1 minute to administer; 3 probes are generally administered at each assessment and the median score is recorded and utilized. ORF probes are provided at no cost at http://dibels.uoregon.edu; therefore, the cost for printing/photocopying stimulus materials is the only financial cost to utilizing ORF.

[5]Data regarding the ability of DIBELS Letter Naming Fluency (LNF) and Phoneme Segmentation Fluency (PSF) to model students' growth over time could not be located. Data suggests that students' growth on LNF is not as useful in predicting students' reaching achievement as WIF slope (L. Fuchs, Fuchs, and Compton, 2004). As noted in footnote 3, above, ORF has been shown to be sensitive to students' growth.

[6]DIBELS tasks require 1 minute per probe, and stimulus materials are downloadable at no cost from http://dibels.uoregon.edu.

[7]Compton, Fuchs, Fuchs, and Bryant (2006) found that models that included WIF slope (growth over 5 weeks, with assessments occurring once per week) reliably predicted students who needed supplemental reading instruction. Additionally, L. Fuchs, Fuchs, and Compton (2004) provided evidence that growth over time on WIF is reliable and valid measure of students' reading achievement.

[8]WIF assessment time is 1 minute per probe; printing/photocopying costs are the only costs of using WIF.

APPENDIX B

Summary of Progress Monitoring Pilot Project

In May of 2007, the Maze Sentences and Reading Fluency tasks were piloted with a group of 58 first grade students in order to obtain preliminary information about the validity and feasibility of the measures. Students were administered one 1-minute probe for each task. As part of a school-wide reading program, the students were also assessed with the Scholastic Reading Inventory (SRI) during the same month. The SRI is an assessment of reading comprehension which yields scores in terms of Lexiles, which indicate both reader ability and text difficulty.

Students' total scores on each of the two experimental tasks were calculated by subtracting the number of incorrect responses from the number of correct responses. On the Maze Sentences task, scores ranged from -1 to 12, with 0 as the minimum number of correct responses and 12 as the maximum number of correct responses. On the Reading Fluency task, scores ranged from -3 to 14, with 0 as the minimum number of correct responses and 14 as the maximum number of correct responses. Additional descriptive statistics are presented below.

	Range of Scores	Mean Score	Standard Deviation
Maze Sentences	-1 to 12	6.6	3.35
Reading Fluency	-3 to 14	7.03	4.12

Results of correlational analyses, presented below, provide preliminary evidence that (a) the two tasks measure a similar construct, due to the correlation of $r = .76$ between Maze

Sentences and Reading Fluency; and (b) that the construct measured by the two experimental tasks is related to reading comprehension, as measured by the SRI and indicated by the correlations of $r = .51$ for Maze Sentences and $r = .67$ for Reading Fluency.

	Maze Sentences	Reading Fluency
Maze Sentences		
Reading Fluency	.76**	
SRI	.51**	.67**

** Correlation is significant at the 0.01 level.