INVESTIGATING VARIABILITY IN STUDENT PERFORMANCE ON DIBELS ORAL READING FLUENCY THIRD GRADE PROGRESS MONITORING PROBES: POSSIBLE CONTRIBUTING FACTORS

by

REBECCA N. BRIGGS

A DISSERTATION

Presented to the Department of Special Education and Clinical Sciences
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2011

DISSERTATION APPROVAL PAGE

Student: Rebecca N. Briggs

Title: Investigating Variability in Student Performance on DIBELS Oral Reading Fluency Third Grade Progress Monitoring Probes: Possible Contributing Factors

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Special Education and Clinical Sciences by:

Roland Good	Chairperson/Advisor
Laura Lee McIntyre	Member
Joe Stevens	Member
Robert Davis	Outside Member
Scott Baker	Member

and

Richard Linton	Vice President for Research and Graduate Studies/Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2011

DISSERTATION ABSTRACT

Rebecca N. Briggs

Doctor of Philosophy

Department of Special Education and Clinical Sciences

June 2011

Title: Investigating Variability in Student Performance on DIBELS Oral Reading Fluency Third Grade Progress Monitoring Probes: Possible Contributing Factors

Approved: ____
Dr. Roland H. Good III, Chair

The current study investigated variability in student performance on DIBELS Oral Reading Fluency (DORF) Progress Monitoring passages for third grade and sought to determine to what extent the variability in weekly progress monitoring scores is related to passage-level factors (e.g., type of passage [i.e., narrative or expository]), readability of the passage, reading rate for words in lists, passage specific comprehension, background knowledge, and interest in the topic of the passage) and student-level factors (e.g., the student's initial skill and variability across benchmark passages).

In light of recent changes in IDEIA legislation allowing for the use of Response to Intervention models and formative assessment practices in the identification of specific learning disabilities, it was intent of this study to identify factors associated with oral reading fluency that, once identified, could potentially be altered or controlled during progress monitoring and decision-making to allow for more defensible educational decisions.

The sample for analysis included 70 third grade students from one school in Iowa. Results of two-level HLM analyses indicated significant effects for background knowledge, interest in the passage, type of passage, retell fluency, readability, and word reading, with type of passage and readability demonstrating the largest magnitude effects. Magnitude of effect was based upon a calculation of proportion of reduction in level 1 residual variance. At level 2, initial risk status demonstrated a significant effect on a student's initial oral reading fluency score, while the benchmark variability demonstrated a significant effect on a student's growth over time.

Results demonstrate support for readability as an indicator of passage difficulty as it relates to predicting oral reading fluency for students and suggest that consideration for the type of passage may be warranted when interpreting student ORF scores. Additionally, results indicated possible student-level effects of variables such as background knowledge and word list that were not investigated within the current study. Limitations of the study, considerations for future research, and implications for practice are discussed.

CURRICULUM VITAE

NAME OF AUTHOR: Rebecca N. Briggs

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
University of Nevada, Reno

DEGREES AWARDED:

Doctor of Philosophy, School Psychology, 2011, University of Oregon
Master of Science, Special Education, 2009, University of Oregon
Bachelor of Arts, Psychology, 2002, University of Nevada, Reno

AREAS OF SPECIAL INTEREST:

Response to Intervention
Early Literacy
Positive Behavioral Intervention and Supports
Systems-level Changes in Schools
School-wide Models of Support

PROFESSIONAL EXPERIENCE:

School Psychologist, Heartland Area Education Agency 11, Johnston, Iowa, August 2009 – present

President, Association of School Psychology Students, University of Oregon, 2008 – 2009

Liaison for University of Oregon, Student Affiliates of School Psychology (Division 16 of APA), University of Oregon, 2008 – 2009

Representative for University of Oregon, National Association of School Psychologists Student, University of Oregon, 2008 – 2009

Oregon Reading First Graduate Teaching Fellow, Center on Teaching and Learning, University of Oregon, 2007 – 2009

DIBELS Data System Graduate Teaching Fellow, University of Oregon, 2006 – 2007

GRANTS, AWARDS, AND HONORS:

College of Education Doctoral Research Award from Dynamic Measurement Group, Investigating Variability in Student Performance on DIBELS Oral Reading Fluency Third Grade Progress Monitoring Probes: Possible Contributing Factors, University of Oregon, 2009

RTI Training Grant, School Psychology Program, University of Oregon, 2006 – 2007

DIBELS Student Support Award; 2006, 2007, 2008, 2009

ADHD Training Grant, School Psychology Program, University of Oregon, 2005 – 2006

Presidential Scholarship, University of Nevada, Reno, 1997 – 2001

ACKNOWLEDGMENTS

I wish to express my sincerest appreciation and gratitude to Professor Good for his support and guidance. His advice, feedback and flexibility were instrumental at every step of this project. Were it not for his unwavering belief in my abilities, completion of this dissertation would not have been possible.

Also, I extend deepest thanks to Heartland Area Education Agency 11 and the Carlisle Community Schools for their support and for allowing me to conduct this research in their schools, and to Simpson College, especially Dr. Carole Richardson and her students, for providing data collection support.

Finally, I wish to acknowledge the support of my coworkers, my family and my friends. They have provided moral support and encouragement throughout.

For my fiancé, Dan, whose willingness to put up with my craziness made this possible. Thank you, for walking my dog when I had to work late, for feeding me when I barely had time to eat and mostly, for doing all of this even though you were not yet legally obligated to do so.

For my parents, Earl and Louise, for supporting me through each new adventure and for always allowing me to take “the long way around.”

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER I

STATEMENT OF THE PROBLEM

The current level of reading achievement in our nation's schools is alarming. The most recent results from the National Assessment of Educational Progress (NAEP) show that while small gains are being seen in average performance across grades, the percentage of students scoring at or above grade-level proficiency is hovering just around 30% and has remained relatively stable since 1992 (National Center for Education Statistics (NCES), 2009). While it is indeed alarming that only 67% of fourth grade students are performing at or above a basic level on the NAEP, it is more alarming to see the disproportionate representation of low income and minority students in the group performing below basic. In 2009, only twenty-three percent of White students and 21% of students who are not eligible for free/reduced-price lunch scored below basic, compared to 49% of students who are eligible for free/reduced-price lunch, 48% of American Indian/Alaska Native students, 52% of Hispanic students, and 53% of Black students (NCES, 2009). Additionally, 66% of students identified as having a disability performed in the below basic range along with 71% of students identified as English Language Learners. It is clear from these numbers that despite gains in average performance for all groups (except American Indian/Alaska Native students, whose performance has dropped), the achievement gaps between groups are not closing. These results demonstrate the severity of the problem facing our schools and that it is imperative we provide good quality instruction and adequate support to *all* students.

Evidence-based Practices and Changes to Legislation Intended to Improve Outcomes for All Students

On a positive note, we have decades of research evidence in reading that shows us what components are essential to and what instructional practices are effective for teaching students to read (National Reading Panel [NRP], 2000; Snow, Burns & Griffin, 1998). Effective evidence-based reading practices (a) focus instruction on the five Big Ideas of reading (i.e., phonemic awareness, phonics, accuracy and fluency with connected text, vocabulary, comprehension) (NRP, 2000) (b) provide explicit instruction (Snow, Burns & Griffin, 1998), (c) provide a systematic and coordinated continuum of instructional support to meet the diverse range of student need (Simmons, Kame'enui, Good, Harn, Cole & Braun, 2002), (d) alter and intensify instruction according to student need (Torgesen, 2002), (e) incorporate ongoing formative assessment of students' skills and employ a decision-making framework for formative evaluation of instruction and student progress (Deno, 1992; Shinn, Shinn, Hamilton & Clark, 2002).

Schools have increasingly adopted evidence-based practices over the past decade and continue to do so in an effort to improve the reading performance of their students, and also to meet the academic performance and accountability standards set forth in legislation such as No Child Left Behind (NCLB, 2001). These accountability standards hold schools responsible for demonstrating academic growth for *all* students, including those receiving special education services. One way special educators are demonstrating the effectiveness of instruction and intervention supports for this population is through the use of formative evaluation. Formative evaluation is a process that uses data gathered from ongoing or formative assessments of target skills to (a) document changes in student skill and progress toward identified goals and (b) determine whether changes need to be made to the intervention in order to increase student progress (Deno, 1986).

While, this process of documenting changes in student skill and determining a student's response to the implementation of an intervention could be an effective and meaningful component in school-wide efforts to improve reading outcomes for all students, it is an especially hot topic for special education in particular due to recent changes in the Individuals with Disabilities in Education Improvement Act (IDEIA, 2004), the governing legislation for special education. The pertinent changes in IDEIA affect the procedures for determining special education eligibility under the category of Specific Learning Disability (SLD). The reauthorization states that when determining whether a child has a SLD a local educational agency "shall not be required to take into consideration whether a child has a severe discrepancy between achievement and intellectual ability" and further states that the agency "may use *a process that determines if the child responds to scientific, research-based intervention* as part of the evaluation procedures" (PL 94-142; emphasis added). In effect, the law states that educators may use this formative evaluation procedure that has generally come to be known as Response to Intervention (RTI) as a means of determining special education eligibility for students suspected of having a SLD.

Formative Evaluation and Response to Intervention

Formative Evaluation

As stated above, formative evaluation is a procedure that uses data gathered from repeated, ongoing assessments to document changes in a student's skill over time and employs a decision-making framework for determining whether the change is sufficient for the student to meet identified goals. If the change in the student's skill is determined to be insufficient, then modifications to instruction or intervention can be made with the intent of increasing the student's skill to desired levels.

The objective of special education is to allow students to make sufficient educational progress toward identified goals through the design and implementation of Individualized Education Programs (IEPs). Essential to this objective is the demonstration that these IEPs or the individualized interventions and supports written into them are effective. At its inception, formative evaluation provided an alternative means of determining the effectiveness of interventions that was in contrast to the traditional practices of pre- and post-testing and judging effectiveness based upon group performance (Deno, 1986).

Pre-post testing or summative evaluation consists of a student's skill level being assessed prior to the start of intervention (pre-test) and then again at the conclusion of intervention (post-test). Two problems arise from this approach to evaluation. First, summative evaluation does not provide information that could inform timely modifications to treatment. If a student is not making adequate progress in response to an intervention, modifications should be made to the intervention supports in order for that student to get on track to meeting the goal. With summative evaluation, the lack of progress of students for whom the intervention is not effective may go unnoticed until the end of treatment when it may be too late to make modifications (e.g., the end of the school year). Second, summative evaluation does not provide the information required to determine the differential effectiveness of treatment components or modifications. Educators may make deliberate modifications to treatments according to their observation of student progress, but assessing student progress only at the end of treatment makes it impossible to empirically determine which modifications accounted for student progress.

In contrast, a formative evaluation approach requires repeated assessment of student skill throughout intervention (e.g., weekly or biweekly). In this way an individual database is compiled for each student and as the intervention is implemented the ongoing evaluation of student progress (i.e., whether it is sufficient to meet the goal) can inform timely modifications to the intervention and changes in student skill can be differentially attributed to those modifications.

Formative evaluation was made possible through the creation of formative assessment materials, namely curriculum-based measurement (CBM; Deno, 1992). Curriculum-based measurement is standardized, short duration, fluency measures of basic skills in reading, writing, spelling and math. In contrast to the traditional published norm-referenced tests of achievement, which are not designed to be given repeatedly over short periods of time and therefore lent themselves only to summative evaluation, CBM provides educators with a measurement tool that is more closely linked to the classroom curriculum, more sensitive to changes in students' skills and can be administered repeatedly over shorter periods of time. CBM was designed to be used in a formative or ongoing manner to create a database for individual students and for that reason; CBM could meaningfully inform decisions regarding the effectiveness of students' Individual Education Programs.

To sum up, formative evaluation is a process by which curriculum-based measurement procedures are used to empirically determine the effectiveness of educational interventions (Deno, 1986). The process of formative evaluation requires two things: (1) assessments that are closely related to the curriculum or skills being taught (i.e., CBM), that can be given repeatedly over shorter amounts of time than traditional published achievement tests, and that are sensitive to changes in student's skills over these shorter periods of time and (2) a data-based decision-making framework for judging a student's response (i.e., changes in a student's performance or skill level) to the intervention so that adjustments and manipulations may be made to the intervention in a timely manner.

An example of formative evaluation using DIBELS in an Outcomes-Driven Model

For an example of a formative evaluation approach, one can look at Good, Gruba & Kaminski (2002) where the authors describe the use of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in the Outcomes-Driven Model (see Figure 1). The DIBELS are a system of curriculum-based measurement designed to be brief and efficient with multiple alternate forms. Assessment with the DIBELS provides information regarding a student's skill in the basic component skills of reading and in addition provides an indicator of that student's risk of not being a proficient reader by the end of third grade. Based upon the DIBELS screening, a student may be identified as being at *low risk*, at *some risk*, or *at risk*. This risk status should alert educators as to how much additional instructional support will be needed for that student to meet the end of year proficiency goals.

The Outcomes-Driven Model is a continuous feedback loop to evaluate instructional support. Figure 1 presents a graphic of the Outcomes-Driven Model. The process consists of (a) identifying the need for support, (b) validating the need for support, (c) planning instructional support, (d) implementing instructional support, (e) evaluating instructional support and (f) reviewing outcomes.

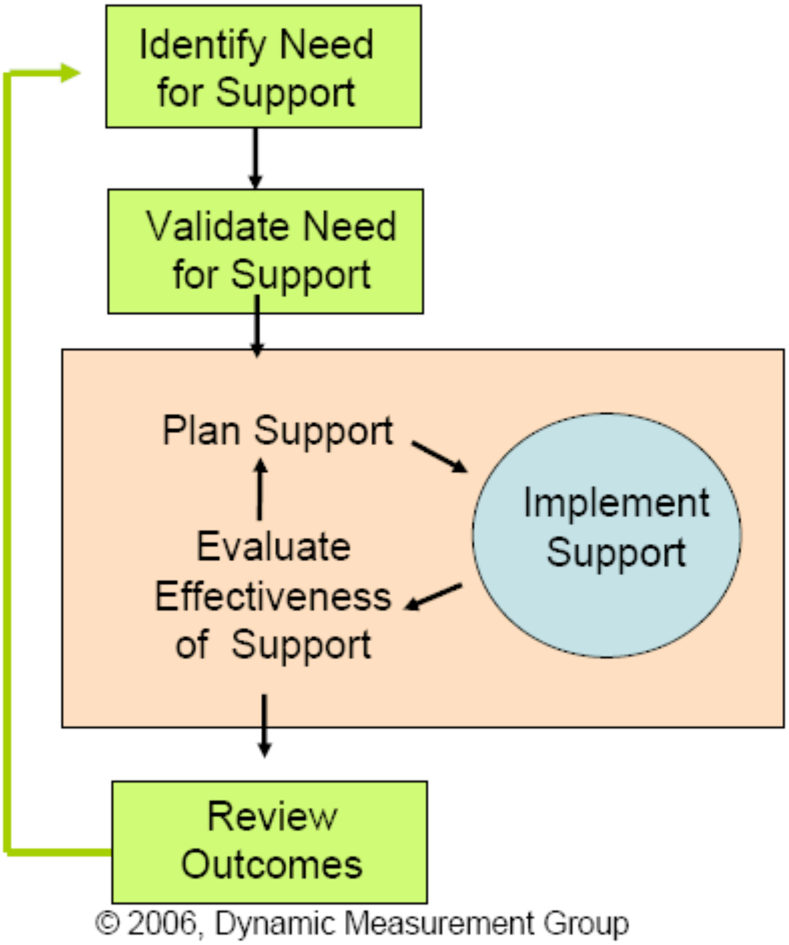

Identify Need for Support
Validate Need for Support
Plan Support
Implement Support
Evaluate Effectiveness of Support
Review Outcomes
© 2006, Dynamic Measurement Group

Figure 1. Outcomes-Driven Model

First, a screening assessment is conducted using the DIBELS, which identifies a student's *need for support*. Next, a student's need for support is *validated* through repeated assessment on different days. By administering repeated assessments, educators can be reasonably confident that the student does need additional support and that the score from the screening assessment was not due to some other unknown factor (e.g., not being familiar with the task, feeling ill and not performing his best, etc.). Next, a *plan for* *instructional support* is created, *implemented*, and *evaluated* using repeated assessments over time. This portion of the Outcomes-Driven Model is another feedback loop within the larger loop, meaning that based upon the evaluation of a student's progress (e.g., if progress is not sufficient) modifications may be made to the plan and the loop (implementation and evaluation) will begin again. The determination of the effectiveness of the plan should be based upon predetermined decision rules. For example, the student's progress can be graphed in relation to the goal and a goal line can be drawn from the student's initial starting point to the goal. Two frequently used data evaluation rules are used when determining whether student progress is sufficient and if modifications to the instruction are warranted. The first rule consists of fitting a trendline to the data points, which serves as an estimate of the student's slope or rate of progress toward the goal and evaluating if the trendline is showing sufficient growth to meet the goal. If it is not, then modifications are warranted. The second rule is based upon the number of consecutive data points that fall below the goal line. If a student's data demonstrate a specified number of consecutive data points below the goal line (e.g., three or four) modifications should be made. An example of how a student's data could be graphed and used to determine the effectiveness of instruction with the Outcomes-Driven Model is provided in Figure 2. The final part of the loop is the process of *reviewing outcomes*. At this point educators should consider whether the student met important outcomes (e.g., grade-level proficiency in the skill or the goal set by the IEP). If the student did not meet the outcome, then the process of the Outcomes-Driven Model can be repeated as many times as necessary for the student to attain the goal.

Evidence-based effects of systematic formative evaluation

The effect of systematic formative evaluation was demonstrated in a meta-analysis conducted by Fuchs and Fuchs (1986). Results of the analysis showed that the integration of formative evaluation produced an overall average unbiased effect size (UES) of .70. This means that the results of interventions that incorporated formative evaluation were on average seventy percent of one standard deviation higher that the results of interventions that did not incorporate formative evaluation. The meta-analysis also determined that data evaluation procedures were more effective when data evaluation rules (i.e., pre-set decision rules regarding student progress) were used (average UES = .91) versus a reliance on teacher judgment (average UES = .42) and that visually graphing student data produced larger effects (average UES = .70) than merely recording the scores (average UES = .26).

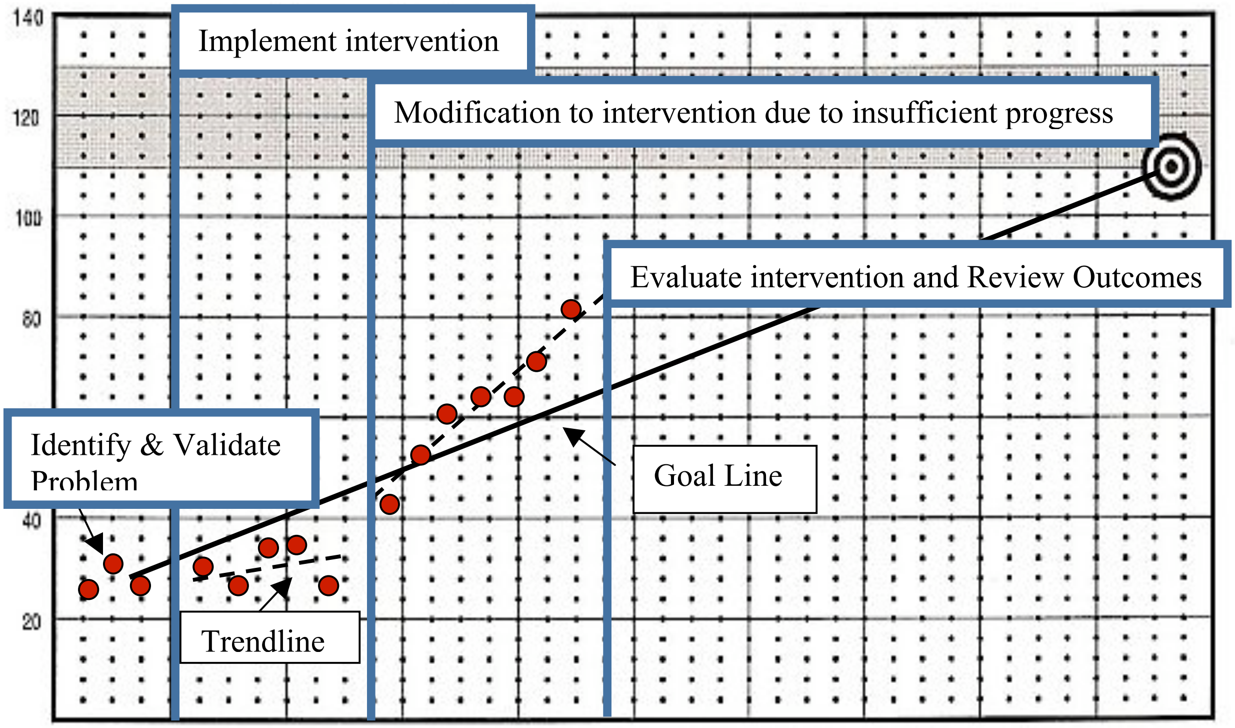


Figure 2. Graph of student progress monitoring data within the Outcomes-Driven Model

Response to Intervention

As indicated above, a formative evaluation procedure known as Response to Intervention (RTI) has been written into IDEIA as a possible means of determining special education eligibility under the category of Specific Learning Disability (SLD). This is a major change from the traditional means of determining eligibility where assessments were conducted to determine whether a student had a severe discrepancy between their ability and their achievement (as assessed by published norm-referenced tests of achievement and intelligence). This change has come about as a response to the increase of students being identified as having a SLD, the concern of educators that they are not making defensible decisions linked to quality instruction and the numerous conceptual and measurement problems associated with the IQ-achievement discrepancy approach (Gresham, VanDerHeyden & Witt, 2005; Vaughn & Fuchs, 2003). The use of RTI also promises to yield assessment data that can more readily inform intervention support planning (Ysseldyke & Christensen, 1988).

RTI is defined by the National Association of the State Directors of Special Education (NASDSE, 2006) as “the practice of providing high-quality instruction/intervention matched to student needs and using learning rate over time and level of performance to make important educational decisions” (p. 5). The main components of RTI are based in the evidence-based practices discussed previously. RTI is comprised of three essential components: (1) a continuum or multiple tiers of instructional support, (2) a decision-making method (e.g., the Outcomes-Driven Model) and (3) integrated formative assessment and evaluation to inform decisions at each level of instructional support.

RTI can be conceptualized as potentially serving two purposes. The first purpose would be to prevent reading problems through implementation of all the components on a school-wide basis. In this way, RTI is a prevention-oriented approach to school-wide instructional service delivery where all students are routinely and formatively assessed, additional support is provided to students who need it via the tiers of instructional support and formative evaluation decisions are made regarding student progress, students' movement between the tiers and the effectiveness of instruction at each tier. The second purpose would be to determine special education eligibility under the category of SLD. When used for this purpose, the implementation does not change, but rather than a student's demonstrated lack of response or progress to evidence-based instruction leading only to formative evaluation decisions regarding changes in instructional placement or modifications to instructional interventions it may also be viewed as demonstration that the student has a SLD and that he or she is eligible for special education services.

For some educators and researchers this dual purpose is troublesome because it means that formative evaluation procedures will go from being used to make what some call "low stakes" decisions (i.e., instructional placement, intervention effectiveness) to being used to make a much "higher stakes" decision (i.e., identification as having a SLD and placement in special education services) (Christ & Silberglitt, 2007; Shinn, 2007). As demonstrated above, formative evaluation is an evidence-based practice that has been shown to improve academic outcomes for students and is therefore considered a defensible approach to educational decision-making, however, decisions made through formative evaluation are often thought of as "low stakes" decisions in part because they are considered to be self-correcting. That is, if a change is made to a student's instruction or placement and the desired response in student skill is not observed, then another change can be made to correct for the initial decision (e.g., the *plan-implement-evaluate* feedback loop within the Outcomes-Driven Model). In contrast, eligibility decisions are often seen as "high stakes" decisions in that they are conceptualized as much less self-correcting. Although, formative evaluation will continue to be used within special education to make instructional decisions and changes to instructional interventions, it is the decision to identify a student as having a SLD and placing them in special education that tends to be regarded as final and potentially stigmatizing. Most students identified as eligible for special education will continue to receive services throughout their academic careers and even if a student progresses in skill to a point where there is no longer a demonstration of educational need for services, the label of having been identified as a student with a disability (i.e., SLD) remains.

If indeed the eligibility decision is a higher stakes decision, then educators and researchers are right to take a closer look at the reliability and validity of the formative assessment data gathered because the reliability and validity of that data will affect the reliability and validity of the decisions made during the formative evaluation procedure.

Issue of Variability in Student Scores on Oral Reading Fluency During Progress Monitoring

An issue that has turned up in the education research literature recently is the stability of student scores on alternate forms of oral reading fluency during progress monitoring. Oral reading fluency (ORF) is the CBM for assessing accuracy and fluency with connected text. Students are asked to read a grade-level passage out loud for one minute. A student's score is the number of words read correctly in one minute. Like all

CBM this measure is able to be administered frequently and is more sensitive to changes in student skill over shorter periods of time than traditional comprehensive reading tests. The reliability and validity of ORF as a measure of reading fluency as well as an indicator of overall reading proficiency is well documented (Fuchs, Fuchs, Hosp & Jenkins, 2001; Shinn, Good, Knutson, Tilly, & Collins, 1992). However, during the progress monitoring process educators often see undesired variability in a student's performance over successive assessments. This variability in scores from one assessment point to the next can make the formative evaluation decision-making process more difficult.

Figure 3 provides visual representation of this issue through a comparison of two progress-monitoring graphs. The graphs represent ORF data gathered for the purpose of determining a student's response to the implemented intervention. Of course some variability in student performance is desired (i.e., growth in student skill), however what is referred to as undesired variability is such that the data points seem to "bounce" around the trendline, making a confident estimate of the pattern of performance difficult to obtain.

If we look at each graph within the formative evaluation framework set out during the discussion of the Outcomes-Driven Model, it will help to illustrate the problems that arise with undesired variability. Let us assume that the student's need for support has been both identified and validated and that the data points on the graphs represent data gathered through biweekly progress monitoring, intended to be used to formatively evaluate the effectiveness of the instructional supports. Let us also assume that the school has decided upon data evaluation rules stating that modifications to the instructional supports will be considered when an examination of the student’s slope of progress (trendline) is not sufficient to achieve the goal and/or when a student’s graph shows three consecutive data points below the goal line.

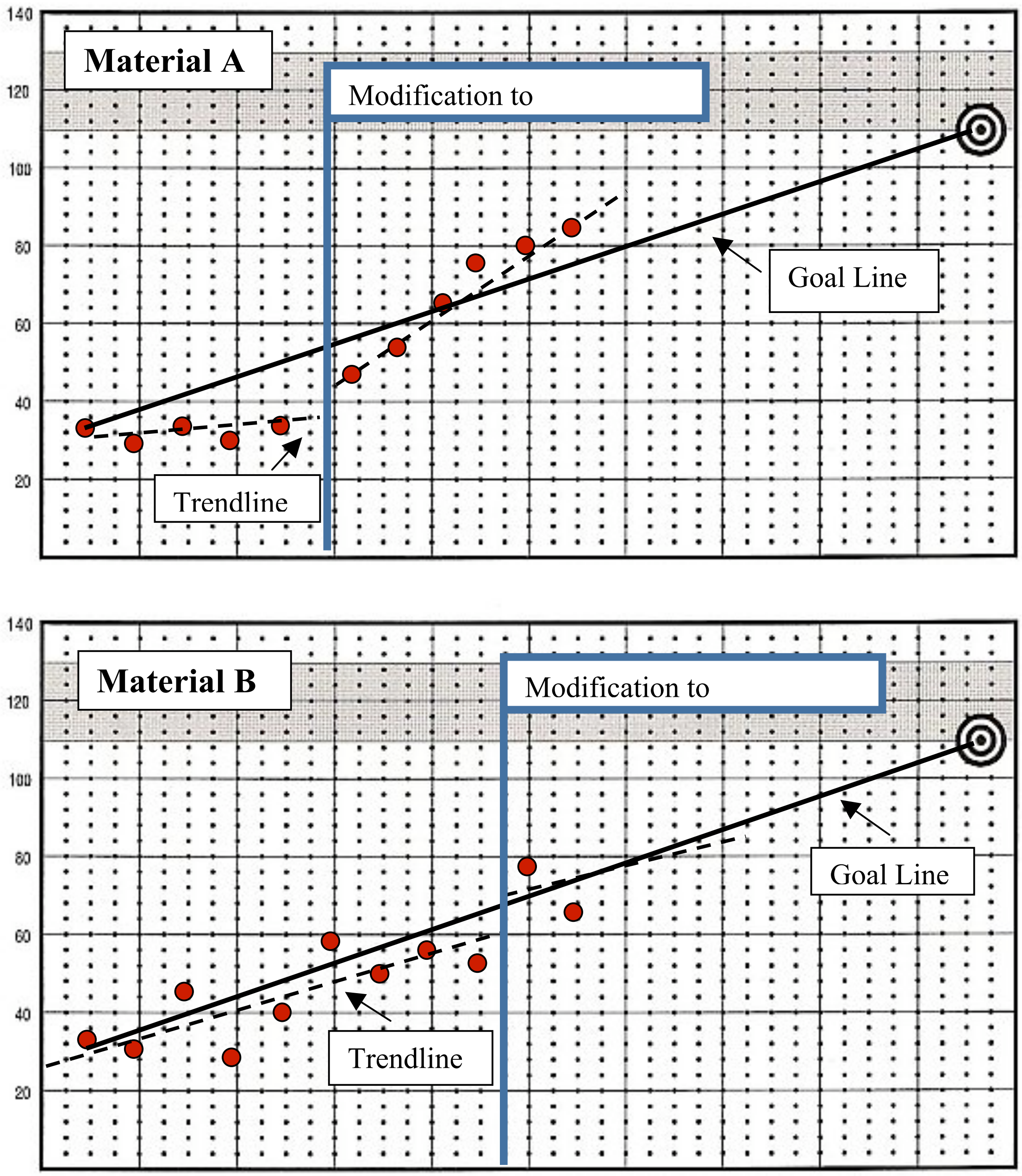


Figure 3. Comparison of desirable and undesirable variability in student scores

According to these data evaluation rules, the graph depicting data from Material A, which consists of more consistent variability around the trendline, would allow for more confident and timely decisions for two reasons. First, the data is consistently close to the trendline, which means we can be more confident that the trendline represents an accurate estimate of the student's current progress as well as future progress given that the intervention remains the same. Second, the data evaluation rule of three points below the goal line was demonstrated within the first three assessments and would lead to a more timely change to intervention. In addition, the data points after the change in intervention continue to demonstrate consistent variability and make it easy to determine that the change in intervention is being effective.

In contrast, the graph depicting data from Material B, which consists of undesired or inconsistent variability around the trendline, is less easily interpretable and would require more data points than the first graph to make decisions based on the evaluation rules for two reasons. First, a trendline fit to the first five data points would not have been as conclusive as in the first graph and the fourth and fifth data points may have appeared to be the start of a positive change in student skill. Second, the pattern of variability was such that no three consecutive data points fell below the goal line until the ninth data point. This need for more data affects the timeliness with which decisions can be made and the inconsistent variability could potentially affect the accuracy of decisions made through formative evaluation. Additionally, a similar pattern of variability is seen after the change in intervention meaning that demonstrating the effectiveness of the change will be equally difficult.

Working under the assumption that undesired variability exists and is a potential problem in regards to making decisions with a high degree of confidence, it would behoove educators to know more about why this undesired variability happens and the factors that may potentially affect it. If contributing factors were identified that were considered alterable or controllable, actions could be taken to lessen the undesired variability in student scores, thereby increasing the chances that appropriate and timely decisions are made regarding student progress, which should ultimately lead to better outcomes.

Factors that Potentially Influence Variability in Student Performance and Estimates of Student Progress Over Time

There are numerous factors that may influence a student's performance on progress monitoring probes. This section will discuss some of them. Three categories will be included (a) passage-level factors, (b) student-level factors, and (c) environmental factors. The categories will be discussed in order of their perceived measurability and/or controllability with the most easily measured factors discussed first.

Passage-level factors

Types of factors within the reading passages that might influence student performance include the difficulty or readability of the text. Readability is measured in different ways and a readability estimate may result from calculating any combination of the following factors: length of passage, average length of sentence, percentage of high frequency words, percentage of decodable words, or percentage of multisyllabic words. Even though all these factors can contribute to readability it is still considered the more easily measured because one can actually count the number of words in the passage or the number of multisyllabic words used, etc. In this way, readability is solidly quantifiable. The same is true for other passage factors such as the genre of the text or the type of text (e.g., narrative or expository), which are classifiable. A passage is either narrative or it's not.

Researchers have looked at the effects of many different passage-level factors on student performance on ORF probes as well as the effects on the estimated slope of progress for students. The accuracy of readability estimate predictions for performance on CBM measures was studied by Ardoin, Suldo, Witt, Aldrich, & McDonald (2005). Findings suggested that the use of certain readability formulas were not supported as a means of judging passage difficulty and that two component parts of the readability formulas (average number of syllables and the number of high frequency words) were more accurate predictors than other component parts. Another study estimated readability, decodability, percentage of high frequency words, average words per sentence, and percentage of multisyllabic words to investigate their affect on student scores and found significant effects for the percentage of high frequency words and passage decodability (Compton, Appleton & Hosp, 2004). Others have looked at how passage difficulty (defined in different ways, but generally as a readability estimate) affects the standard error of ORF scores (Christ, 2006; Hintze & Christ, 2004; Poncy, Skinner & Axtell, 2005) and the estimation of student growth across triannual screenings (Ardoin & Christ, 2008).

Another passage-level factor that is of interest to this study is the type of text (i.e., narrative vs. expository). One study by Saenz and Fuchs (2002) looked at the relation of type of text on reading fluency and comprehension for high school students with learning disabilities. They found that students had more difficulty with expository passages than narrative. Research looking at the effect of this on reading fluency for a younger population of students would be interesting given that expository text is introduced into the reading curriculum at about the third or fourth grade and the reading portion of the NAEP assessment at fourth grade employs the use of both types of text and reports student performance on each type individually.

Student-level factors

Factors at the student level begin to be more complicated to measure, especially in the case of reading, which is a complicated construct that incorporates many component skills. Individual student skill in different component parts may work interactively, making determining the true effect of one component compared to another more difficult. Some factors of interest as predictors of performance on progress monitoring passages are the student's skill in reading fluency at the start of progress monitoring (i.e., initial skill level), the student's average variability of scores during benchmark screening, the student's oral reading rate for words in isolation, the student's skill in comprehension, the student's background knowledge of passage content, and the student's prior knowledge of and interest in passage content.

In studies that have investigated some of these factors it has been shown that that student skill in reading interacts with passage difficulty (Poncy, Skinner & Axtell, 2005) as well as with the reliability of the testing procedure (Christ & Silberglitt, 2007) to influence reading fluency scores. It has also been shown that a student's growth in reading fluency varies by initial skill level with students whose initial skills are lower making less growth (Silberglitt & Hintze, 2007).

Studies examining the predictive relation of comprehension to oral reading fluency on individual reading passages were not found, but investigating this relation would be of interest due to the reciprocal relation between fluency and comprehension (Pikulski & Chard, 2005). Similarly, research investigating the relation between student prior knowledge of passage specific content and the student's level of interest in individual passages on reading fluency scores was not found. However, the relation of these factors with reading have been documented as important (Biancarosa & Snow, 2004; Hirsch, 2003; Kamil, 2003), especially for older elementary students as they make the transition from "learning to read" to "reading to learn."

Environmental factors

Environmental factors are often times the least controllable within the average school (especially when conducting large-scale studies where data collection is done in the school building, by school staff under "business as usual" conditions) and it is therefore more difficult to study their effects. Examples of environmental factors include the testing environment (e.g., loud or quiet; within classroom or in another location), the tester (e.g., familiar or unfamiliar person, same or different person each week), the time of the day when testing takes place (e.g., just after instruction, just before or after recess or lunch), and the time of the year testing takes place (e.g., just before a holiday or fieldtrip) among many others.

As stated above, some factors are more easily studied due to their being more quantifiable or controllable, but even those that are less than easy to quantify or control are nonetheless interesting and potentially important to this issue of variability.

Purpose of this Study

The purpose of this study was to investigate further the relations between various passage-level and student-level factors and student oral reading fluency during progress monitoring. Factors at the passage and student level were chosen based on their ability to be measured or quantified as well as the feasibility of completing the study within a framework of limited time and resources. Environmental factors were not examined at this time. This study investigated student performance on the DIBELS Oral Reading Fluency Progress Monitoring passages for third grade and sought to determine to what extent the variability in scores is related to (1) the student's identified level of risk (i.e., at risk, some risk, low risk), (2) the average variability in the student's scores during benchmarking, (3) the type of passage (i.e., narrative or expository), (4) the grade level readability of the passage, (5) the student's background knowledge of passage content, (6) the student's comprehension of the passage, (7) the student's reading rate on words in isolation, (8) the student's self-rating of prior knowledge about the topic of the passage, and (9) the student's self-rating of interest in the topic of the passage. The specific research questions are:

(1) Is there a relation between passage-level factors (i.e., readability estimates, type of passage, student comprehension of passage content, student background knowledge of passage content, student rating of interest in passage content, student rating of prior knowledge of passage content, student reading rate of words in isolation) and students' oral reading fluency progress monitoring performance?

(2) Is there a relation between student-level factors (i.e., students' initial skill in ORF, average variability of students' scores during initial benchmark screening) and students' oral reading fluency progress monitoring performance?

CHAPTER II

LITERATURE REVIEW

The response to intervention (RTI) or treatment integrity (Fuchs, 1998) approach has been a focus of educational research since its conceptualizations as a service delivery model and even more so recently since its inclusion in the Individuals with Disabilities in Education Improvement Act (IDEIA). RTI shows promise as a means of improving educational outcomes for all students and as a more valid process of identifying students in need of special education support than the traditional IQ-achievement discrepancy model (Gresham, 2002; Gresham, VanDerHeyden & Witt, 2005; Vaughn & Fuchs, 2003). The research base to support these assumptions continues to grow. As RTI has become more generally accepted as a service delivery model and means of identifying students with SLD, research demonstrating the reliability and validity of its component parts has become imperative. One important component is the assessment component. It is paramount that the assessments being used to determine skill level and to demonstrate responsiveness (or the lack there of) be of the highest possible technical adequacy, while remaining feasible and instructionally relevant, and that any less than desirable attributes or factors be widely known and understood so that decision-making can take place in a fully informed way.

This literature review intends to do the following: (1) outline the support for use of RTI in eligibility decisions; (2) discuss the concerns of researchers as they apply to the technical adequacy of the measures used to determine “response;” and (3) review research conducted recently on the adequacy of oral reading fluency measures and factors that possibly affect student progress on these measures, highlighting the need for the dissertation research conducted.

Response to Intervention as Eligibility Determination

As stated in Chapter 1, the most recent reauthorization of the IDEIA includes wording that allows, but does not require, the use of a RTI process in determining eligibility for special education under the category of specific learning disability (SLD) (IDEIA, 2004). Although the reauthorization still allows the use of an IQ-achievement discrepancy model, it no longer mandates it. This is considered by many to be a step forward in the identification process, as the discrepancy model was generally perceived to be inappropriate and wrought with inadequacies (Gresham, 2002).

The beginnings of an RTI approach are traced back to a National Research Council (NRC) study, which posited that the validity of a special education placement should be judged according to three criteria (Heller, Holtzman & Messick, 1982). The first relates to the quality of the general education program, in that the quality should be sufficient that students would be reasonably expected to learn. The second relates to the value of the special education program, in that the program should be effective enough that the student would benefit from placement in it. The third relates to the accuracy and meaningfulness of the assessments used throughout the placement process, in that the decisions one makes are only as good as the data on which they are based. RTI as defined by the National Association of School Directors of Special Education (NASDE, 2006) seems to embody these criteria. It is conceptualized as a prevention-oriented service delivery model whereby schools provide effective, evidence-based instruction and intervention support (meets first and second criteria), matched to student need, and use instructionally relevant formative assessment and evaluation to determine learning rate over time and level of performance, which is used to inform educational decision-making (meets third criterion) (NASDSE, 2006). The decision making process in RTI is still based on discrepancy, however in contrast to the traditional IQ-achievement discrepancy, which is a comparison of a student's scores on two published norm referenced tests at one static point in time, the discrepancy in RTI is based upon the comparison of pre- to post-intervention levels of academic performance (Gresham, 2002), as well as growth over time (Fuchs & Fuchs, 1998; Gresham, VanDerHeyden & Witt, 2005).

Research has shown that the RTI approach is reliable in identifying students who struggle in reading and are in need of intensive levels of instructional support. Speece, Case and Malloy (2003) conducted three studies longitudinally across three years, which investigated the validity of identification of reading difficulties using a RTI approach and whether there was additional benefit from RTI on academic outcomes for students. Their first study showed that a dual discrepancy model of identification (i.e., based upon the criteria that a student is discrepant in both level and growth on CBM measures of reading) demonstrated construct validity in that the students identified as dually discrepant (DD) performed lower on measures of reading than the two groups who were expected not to show reading difficulty, the "purposive sample" and the "at risk" groups. When compared to the "IQ-reading achievement discrepancy" group, the DD students expectedly scored higher on the IQ test, but did not score differently on reading skills. This demonstrated that the IQ approach does not catch all students who struggle in reading. The first study also showed that, when compared on age, gender and race, the students identified as DD were younger and reflected the gender and racial proportions of the sample school system more appropriately than the IQ discrepancy group, 80 percent of which was from the majority group (Caucasian).

In their second study, the researchers wanted to know if students who show persistent nonresponsiveness, that is they remain persistently dually discrepant even when provided with modifications to the general education curriculum, demonstrated poorer reading outcomes than students who were never dually discrepant or students who were only infrequently dually discrepant (i.e., sometimes discrepant and other times not). Results indicated that the persistently dually discrepant group did in fact perform lower on reading skills across all three years, even while receiving almost twice as many services outside the general education classroom as the other groups of students. Additionally, the oral reading fluency slopes for this group were similar to estimated slopes of students identified for special education set forth in a study by Deno, Fuchs, Marston and Shin (2001).

The third of their studies demonstrated that the enhanced pre-referral system (i.e., RTI) with its specially designed instruction and intervention within the general education setting improved outcomes for students. DD students who received specially designed instruction in the general education setting ended up requiring fewer services beyond the general education and had better outcomes than DD students who did not receive these interventions.

These studies demonstrate that RTI is a reliable means of identifying students in need of intense levels of academic support. Their RTI approach identified a more ethnically proportional population of students and identified them at a younger age, thus rectifying two of the major criticisms of the traditional IQ-achievement discrepancy model (i.e., wait to fail, and disproportional identification for minority students).

Concerns Regarding Assessment

"The RTI approach will require the direct measurement of behaviors necessary for successful performance using low inference assessment tools. Outcomes of interventions will be judged based on whether or not they produce acceptable levels of performance" (Gresham, VanDerHeyden & Witt, 2005, p. 30). Much of what RTI purports to do (e.g., identify students who are struggling in reading, evaluate interventions based on student performance, evaluate general education instruction and curriculum based on student performance, determine levels of responsiveness or unresponsiveness to intervention, etc.) depends upon assessment at each level within the model (Stecker, Fuchs & Fuchs, 2008). As such, the accuracy and validity of these assessments is essential to sound decision-making and valid evaluation.

CBM as "Perfect" for Use in RTI

In a 2007 article, Shinn discusses all the ways that curriculum-based measurement (CBM) "fit[s]" RTI. CBM are low inference, direct measures of student reading (and math, writing, spelling, etc.) skills, which are standardized, have sound technical adequacy (reliability and validity) data for progress monitoring decisions, can be given frequently, are sensitive to change in skill, and can be given on a large scale (e.g., school- or district-wide) to create a local normative database to which individual student performance and growth can be compared, etc. As Shinn (2007) and others (Gresham, 2002) argue, CBM truly is a good fit for the RTI model. This is not surprising given that CBM was designed for the purpose of determining progress toward IEP goals for students in special education, and therefore determining students' response to the services (i.e., interventions) being provided. Ironically, this same fact is a reason why some researchers are wary of using CBM for the second purpose of RTI—eligibility decisions. Using CBM to determine responsiveness for the purpose of identifying SLD and entitlement to special education is a purpose for which CBM was neither created nor validated (Burns, Jacob & Wagner, 2008). However, this does not mean that CBM is not or will not be proved to be a defensible means of determining response to intervention for the purpose of entitlement decisions, but one cannot fault researchers and educators for wanting to be certain they are using the best possible assessments in order to have the most reliable and valid data on which to base these most important educational decisions.

Initially, concerns in the literature dealt with the underuse of CBM in schools (even for what it was originally intended (i.e., goal monitoring on IEPs) (Shinn, 2007), and therefore the unfamiliarity of educators with it. Concerns were voiced about providing teachers and school psychologists appropriate training on its purpose, how to administer it, how to interpret the data and how to do all of this with integrity (Gersten & Dimino, 2006; Vaughn & Fuchs, 2003; Shinn, 2007). More recently, for reasons presented above, researchers have taken to investigating the technical adequacy of CBM for the purpose of determining response. Assessments being used to determine eligibility should be of the utmost technical adequacy, with any inadequacies fully understood so as to take into consideration during the decision-making process.

Partitioning Sources of Variance and Investigating Standard Error

CBM is used in RTI for two different comparisons. The first is a comparison of a student's performance to a criterion or to peers' (or normative) performance. This is a comparison between individuals. The second is a comparison of a student's performance across time, before, during and after intervention. This is a within individual comparison. Researchers have investigated the reliability of CBM oral reading fluency for these types of decisions.

Hintze, Owen, Shapiro & Daly (2000) used G-Theory to investigate the sources of error in CBM as well as to determine the dependability of CBM for the two types of decisions described above. G theory, formulated by Cronbach, Gleser, Nanda, & Rajaratnam (1972, as cited in Hintze, et al. (2002)) is an alternative to classical test score theory that allows researchers to examine multiple sources of error and to specify what portions of the variance are attributed to the various sources of error. G theory provides a g coefficient, which is interpreted as an indication of the dependability of the measure and can be compared to a classical test score theory's reliability coefficient. The results of the Hintze, et al. study showed that the largest portion of the variance in oral reading fluency performance was attributed to individual variation (48%), with 19% attributed to changes across grades, and 21% attributed to the residual or unaccounted for sources of error (other sources were shown to account for much smaller portions of the variance and are not listed here). This means that practitioners can be reasonably confident that the change in oral reading fluency performance over time can be attributed to changes in student skill rather than other sources. In addition, their study showed that CBM progress monitoring is dependable for making both between and within individual decisions, with g coefficients ranging from .82 to .99. The study also investigated the affect of passage difficulty on g coefficients and the dependability of progress monitoring data. They found that once again, the bulk of the variance (42%) was attributed to individual variation.

Also, their results indicated an affect of passage difficulty on the dependability of the measures. They found that practitioners could expect adequate dependability (g coefficient of .80) when using material at the instructional or at the long-term goal level, but lower dependability (g coefficient of .67) when materials were either too easy or too hard for students.

Other research has looked at the Standard Error of the Estimate (SEE) and the Standard Error of the Slope (SE(*b*)) as indicators of the reliability of CBM oral reading fluency. Measurement error can have an important affect on decision-making with progress monitoring data. For example, if the SEE or SE(*b*) of the progress monitoring data is larger than the amount of growth of the student's ORF score, it would be impossible to know whether the change in score was attributed to a change in student skill, or to measurement error. Therefore, understanding the amount of standard error involved with CBM as well as doing anything one could do to reduce the amount of standard error would benefit practitioners by allowing more confidence in the data on which they are basing decisions.

Hintze & Christ (2004) investigated the affect of controlling passage difficulty on ORF progress monitoring performance. Two sets of grade-level progress monitoring probes were created and administered to 99 students in grades 2 – 5. Probe sets were sampled from common reading curricula. One set was created through purposeful selection of passages from five common curricula and passages that were determined by the curricula as being at a mid-year reading level were selected. The second set was created through random selection of page numbers from the curriculum in which the students were currently being instructed. There was no attempt to control for difficulty in the selection process, however each passage was evaluated using either the Spache or Dale-Chall readability formulas post selection. All passages (controlled and uncontrolled) were chosen from narrative material. Results showed that both SEE and SE(*b)* were significantly reduced when passage difficulty was more strictly controlled. SEE decreased from 16.10 to 13.37, while SE(*b*) decreased from 1.27 to 1.07.

Poncy, Skinner & Axtell (2005) further explored sources of error in CBM oral reading fluency and means by which standard error could be reduced with a study that looked at (1) determining the percentage of variability in scores that was due to student skill, passage difficulty, and unaccounted sources of error, (2) investigating the reliability and the standard error for scores given different numbers of probes and (3) the amount to which altering probe-set variability could reduce standard error given different numbers of probes. Their findings further supported earlier studies in that they found the majority of variance attributed to individual student variability (81%), with 10% attributed to passage or probe variability and 9% to unaccounted sources of error. Additionally, results indicated that increasing the number of probes given increased the g coefficient (i.e., reliability) as well as reduced the standard error. The final analyses showed that restricting the probe-set variability (the average words correct per minute (wcpm) score for each probe was used in comparison to the overall average score across probes to create probe-sets of +/- 15 wcpm, +/- 10 wcpm and +/-5 wcpm, respectively) increased the amount of variance attributed to the individual person from 81% (uncontrolled probe-set) to 89% (+/- 5 wcpm probe-set), while decreasing the amount of variance attributed to the probe from 10% (uncontrolled probe-set) to 1% (for the +/- 5 wcpm probe-set). The variance attributed to unaccounted sources of error remained relatively stable at 9% to 10%.

A study conducted by Christ (2006) investigated the affect of duration of progress monitoring on the SEE and the SE(*b*). Fourteen progress monitoring durations, ranging from 2 to 15 weeks were used. The study resulted in two general findings—the first, that longer progress monitoring durations resulted in less SE(*b*), the second, that the SEE is reduced when more optimal testing conditions (e.g., quiet environment, consistent administrator, equivalent passages, etc.) are employed.

As more attention has been directed toward passage equivalency as a means of reducing standard error, many researchers have investigated the common practice of using readability formulas to equate passages. The findings of such research have indicated that readability formulas provide only a modest relationship between reading fluency and passage difficulty and that formulas most often used to equate CBM passages showed the worst predictive value (Ardoin, Suldo, Witt, Aldrich, & McDonald, 2005), that passages considered equivalent by readability score produced significant passage effects that mask actual student growth (Francis, Santi, Barr, Fletcher, Varisco, & Foorman, 2008), that passage difficulty as measured by the average words read correct per minute for passages had non-significant relationships with the calculated readability of the passages (Betts, Pickart, & Heistad, 2009) and that readability estimates were not significantly related to student performance on passages, but that student performance was significantly related to the performance of other students, indicating that student performance might be a better predictor of passage difficulty than readability estimates (Ardoin, Williams, Christ, Klubnik, & Wellborn, 2010). In fact, Christ & Ardoin (2009) demonstrated that passage sets created using field trial procedures (i.e., calculating the mean words correct per minute score, calculating the Euclidean Distance for each passage), were superior to sets created through random selection or through the use of readability formulas. Ardoin & Christ (2009) furthered this research by comparing an experimental passage set that was created using the Euclidean distance procedure (Christ & Ardoin, 2009) to probe-sets created from both AIMSweb and DIBELS, two commonly used CBM systems. Results demonstrated less standard error for the experimental passage set than for either of the comparison sets. This study also showed that standard error was larger for students who demonstrated higher rates of fluency. This is further support for carefully considering the level at which students are being monitored as another means of reducing standard error.

Investigating and Understanding Other Sources of Variance

The current study sought to build upon these lines of research by investigating what other factors could potentially explain variance within student progress monitoring scores. The goal was to provide ideas for further considerations regarding the creation of equivalent forms, for reducing standard error and interpreting progress monitoring data. The current study investigated how passage type (narrative or expository), student background knowledge of passage content, the average readability of the passage, student comprehension of the passage, student interest in and prior knowledge of passage content, student rate of reading passage specific word lists, and initial fluency rates affected the explained variance of ORF progress monitoring scores across four weeks.

CHAPTER III

METHODS

Participants

Setting

This study took place in the school district of a small suburb (population = 3,725) of a metropolitan area in Iowa. The district has one elementary (grades Pk – 3), one upper elementary (grades 4 – 5), one middle (grades 6 – 8), and one high school (grades 9 -12). The district's total enrollment was 1,888 students (PK – 12). The demographic makeup was roughly 91% White, 4% Hispanic, 2% Asian, 1% Black and less than 1% Native American and Pacific Islander. The district percent of students who are English Language Learners is less than 1% and the percent of students eligible for Free or Reduced Price Lunch is about 20%.

Students

The participants for this study were 74 third grade students from the one elementary building in the district. Total enrollment for third grade was 162 students with no identified English Language Learners and roughly 13% of third graders identified as eligible for Special Education services in any service category. All third grade students in the participating elementary school were given the opportunity to participate. All recruitment procedures were approved by the University of Oregon's Office for Protection of Human Subjects, the research office of Heartland Area Education Agency, as well as the school board of the district. Active consent was obtained from each participant prior to data collection. Consent letters were sent home to parents and guardians explaining the research study, outlining what activities would be involved, the incentives to be given to the students should they choose to participate and detailing the anticipated benefits and possible risks of participation. The consent letters also explicitly told parents/guardians and students that participation was voluntary and that they had the right to revoke consent at any time and discontinue participation in the study. Additionally, prior to the start of data collection, the study was verbally explained to students by the primary investigator. They were told what types of tasks they would be doing for the study, what incentives would be offered, and that they were able to stop participation at any time. They were told that their parents had given permission for them to participate if they wanted to and their written assent was obtained at this time. Student participants were given small incentives (e.g., erasers, pencils, etc.) throughout data collection and were given a certificate for a movie ticket at the conclusion of the weekly data collection phase. Teachers of participating students were given two certificates for movie tickets at the conclusion of data collection as a token of appreciation for their flexibility and assistance. The participating school was offered a gift certificate to a local bookstore or a donation to the library fund, however the administrator declined.

Data Collectors

Data collectors for this study were a professor and undergraduate students from an education assessment course at a nearby college in Iowa. Data collectors were provided with training on all measures prior to data collection and demonstrated an inter-rater reliability of at least 90%. Data collectors were offered either research credit or financial payment for their participation in the study, however it was the determination of the course administrator that the students would participate in the study as partial fulfillment of course credit and would therefore not receive financial compensation for time spent during the academic year. Financial compensation was provided for a subset of data collectors who chose to continue participation for the final day of data collection, which occurred in the week after the academic term had finished. Compensation for mileage expenses was provided for all data collectors who used their own vehicles to commute to the data collection site throughout the study.

Measures

DIBELS 6th Edition Oral Reading Fluency Progress Monitoring Passages for Third Grade

DIBELS 6th Edition Oral Reading Fluency (DORF) is a one-minute timed measure designed to give an indication of a student's skill in accurately and efficiently reading connected text. A student is asked to read a grade-level passage aloud for one minute. Any words read incorrectly, omitted, substituted or any words that the student hesitates on for more than three seconds are scored as incorrect. The recorded score for a student includes the total number of words read correctly and the total number of errors at the end of one minute. DIBELS 6th Edition provides twenty alternate form progress-monitoring passages at the third grade level.

Passage selection

For this study, four of the twenty passages were chosen. Passage selection was a three-part process. First, each passage was categorized as narrative, expository or hybrid (employing both narrative and expository elements). The passages that were considered hybrid passages were not included in this study. Next, passage selection was further determined through the creation of Passage Specific Comprehension Questions (described below). The narrative and expository passages that were found to be either too short or too long to adhere to the process for the creation of the comprehension questions were excluded. These first two steps yielded 10 passages: five narrative and five expository. Finally, DIBELS Early Literacy Research Team at the University of Oregon was enlisted to review and rate each set of passage specific questions The research team read each passage, read the comprehension questions and rated each set of questions on a scale of 1 – 4 in 6 categories. The team rated the extent to which each set of questions 1) had a balance of questions/answers containing words directly from the text and questions/answers containing words with similar meanings to the words in the text, 2) questions would discriminate between students who have higher and lower comprehension skills (e.g., about half of students will get the questions correct), 3) question wording and question topic (i.e., the content of the answer fits the context of the sentence) were clear, 4) questions had one incorrect answer that was similar to the correct answer and three that were fairly different from the correct response, 5) questions that can be answered correctly without reading the passage were avoided, and 6) questions one may not be able to answer correctly after reading the story (i.e., those that *could* be correct, even when going back to the story) were avoided. A total rating number was calculated for each set of questions (e.g., the numerical ratings of each research team member in each of the 6 categories were summed to produce an overall score) and the four passages with the highest scores were chosen for use in data collection. The two highest rated narrative passages were Passage #10 “I Belong to a Big Family” and Passage #12 “Strawberry Jam.” The two highest rated expository passages were Passage #8 “Elephants” and Passage #19 “Clouds and Weather.” Examples of these passages are included in Appendix A.

DIBELS 6th Edition Retell Fluency

DIBELS 6th Edition Retell Fluency (RTF) is a measure designed to add a comprehension check to the DORF measure. After the standard administration of the DORF measure, the student is asked to retell as much of the passage as he or she can in one minute. The recorded score is the number of words the student uses to retell the passage.

Passage Specific Comprehension Questions / Background Knowledge Assessment

For the purposes of this study, passage specific comprehension questions were created. Each of the four progress monitoring passages was used to create a set of ten comprehension questions. The process by which the questions were created consisted of dividing the sentences of each passage into groups of two sentences and creating one comprehension question based on the information provided in those two sentences. Each comprehension question was written in a fill-in-the-blank format and five response choices were provided for each question. One response choice was the correct answer according to the information in the sentences, one response choice was an answer that could be considered correct based on background or general knowledge but is not correct according to the information in the sentences, and the remaining three response choices were included as distracters. The inclusion of the background knowledge response choice allowed these questions to be given in a pre-test format with the goal of gathering information about student background knowledge of the topics in the passages before they had read the passages. Examples of comprehension questions are provided in Appendix B.

Passage Specific Word Lists

For the purpose of this study, a word list was created to correspond to each of the four passages chosen. Each word list was created by selecting the first fifty words from the passage. These words were then randomly ordered and presented in columns. Students were asked to read the word list for one minute and the number of words correct was scored.

Student Rating of Interest and Prior Knowledge

A rating scale was also created to allow students to give a self-rating of their level of interest in the passage they had just read and to rate their level of prior knowledge about the topic of the passage. The rating scales were presented on an 8.5 x 11 sheet of paper titled "You Rate the Story!" The questions "How interesting did you think the story was?" and "How much did you know about what you read in the story *before* you read it?" were written above their respective rating scales. Students were asked to indicate their level of interest and prior knowledge on a scale from 0 to 3, where 0 indicated the lowest level of interest or no prior knowledge, and 3 indicated the highest level of interest or a lot of prior knowledge. An example of the rating forms is provided in Appendix C.

Readability Estimates

The readability scores for each progress monitoring passage were obtained from Technical Report #10 on the DIBELS website (https://dibels.uoregon.edu). This technical report details the process undertaken by the DIBELS authors to create each passage, provides the readability estimates from nine different readability formulas (Dale-Chall, Flesch, FOG, Powers*, SMOG, FORCAST, Fry, Spache, & TASA DRP), and describes how the average readability estimate was calculated for each passage. The authors used a process by which the Spache readability formula was used to create passages within a target range for each grade level. The Spache readability formula was also used to revise and modify the passages to bring them within the target range for each grade level (e.g., adding or subtracting multisyllabic words, increasing or decreasing sentence length, etc.). Then the nine different readability formulas were used to calculate nine different readability estimates for each passage. These nine readability estimates were then transformed into *z* scores (*M* = 0, *SD* = 1). The *z* scores were then averaged to create an overall estimate of the readability, which is the "Average" readability presented in the technical report. This average readability estimate (see Table 1) was used during data analysis as a passage-level factor to investigate the relation between readability and student performance. A complete table of the readability estimates is provided in Appendix D.

Table 1.

Average Readability Score by Passage and Passage Type

	Narrative		Expository	
	Passage #12 "Strawberry Jam"	Passage #10 "I Belong to a Big Family"	Passage #19 "Clouds and Weather"	Passage #8 "Elephants"
Average Readability Score	-1.3	-.30	.10	1.4

Procedure

Training of Data Collectors

Data collectors participated in a two-hour training that encompassed all the measures given in the study. The training was led by the primary investigator and all data collectors were required to obtain inter-rater reliability of at least 90% on the administration and scoring of the DORF and DRTF measures.

Pre-Test Data Collection

DORF benchmark screening

A benchmark screening is given three times per year in the participating district and is meant to give an indication of a student's progress toward end of the year oral reading fluency goals. Standard administration procedures for benchmark screenings require students to read each of three benchmark passages aloud for one minute. The number of words read correctly for each passage is recorded. The final score for a student is the middle or median score of the three passages. This score is used to classify the student into a risk category. For example, if a student reads 66 words or fewer in the winter of third grade, then he or she is considered *at risk* for not meeting the end of year goal; if a student reads between 67 and 91 words, then he or she is considered at *some risk*; and if a student reads 92 words or above, then he or she is considered at *low risk*. These risk categories were used during data analysis as student level factors to investigate the relation between students' initial skill and performance on progress monitoring passages.

In May 2010, all third grade students were administered the Spring DIBELS Benchmark screening as part of regular district assessment practices. With district permission, the data gathered through this district assessment was accessed for use in this study.

Background knowledge assessment

Prior to weekly data collection students were asked to complete the Background Knowledge Assessment. This assessment included the 40 Passage Specific Comprehension Questions (i.e., 10 questions corresponding to each of the four passages selected) grouped by tens according the passage from which they were created. This assessment was given in a group format. Students were seated at individual desks and the primary researcher read a script introducing the task and detailing the directions for completion of the task. The students were told these specific directions, *"I am going to read you some fill-in-the-blank sentences. After I read each sentence I will read five choices to complete the sentence. Choose the answer that you think best completes the sentence. Please follow along with me as I read the sentences and answer choices to you. Ready? Let's begin."* After each item and response choices were read students were given approximately 3 or 4 seconds to mark their answer choice before the next item was read. The total administration time for this assessment was approximately twenty minutes and the administration was conducted in one sitting. Students were thanked for their hard work and given a small incentive (e.g., pencil) at the end of the administration.

Weekly Data Collection

Assessment packets were compiled prior to the start of data collection. The order of presentation for the four passages was determined with a balanced 4 x 4 Latin Square design. The word list assigned for each week of data collection corresponded to the DORF passage the student would read in the following week or for the final week of data collection, students read the word list that corresponded to the passage they had read in the first week. Students were randomly assigned to a passage order.

DORF progress monitoring

Data collection took place one time per week for four weeks. Data collection took place in the school's cafeteria. During the assessment students sat one on one with a trained data collector at the cafeteria tables. Data collectors administered the DORF according to the standardized administration procedures. Students were presented with an 8.5 x 11 sheet of paper with the reading passage on it. They were told these specific directions *"Please read this (point to page) 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 first word of passage). Begin."* The student read aloud for one minute. Any words read incorrectly, omitted, substituted or any words that the student hesitated on for more than three seconds were scored as incorrect. At the end of one minute, the number of words read correctly was recorded. During progress monitoring only one passage is given and the score for that passage is recorded. Students read only one passage per data collection time.

RTF progress monitoring

Immediately following the DORF administration, the Retell Fluency measure was given according to the standardized administration procedures. Students were told these specific directions *"Please tell me all about what you just read. Try to tell me everything you can. Begin."* The data collector counted the number of words the students used to retell the passage. Minor repetitions, redundancies, and inaccuracies were considered correct whereas rote repetitions, songs or recitations were considered incorrect. The total number of words used in the retell was then recoded as the students' score.

Passage specific comprehension questions

Immediately following the RTF administration the student was presented with an 8.5 x 11 sheet of paper with the passage specific comprehension questions for that progress monitoring passage. The student was told these specific directions, "*Here are some fill-in-the-blank sentences about what you just read. Read each sentence to yourself and then choose the best answer to complete each sentence. The best answer is the one that completely matches what you just read in the story. Ready? Let's begin."* The student was given unlimited time to complete this assessment. The student was asked to answer only the questions that pertained to the portion of the passage they read. The total number of questions answered correctly, and the number of questions attempted was recorded.

Student rating of interest and prior knowledge

Immediately following the comprehension questions the student was presented with the student rating form. The student was given the rating form and the data collector read the first question ("*How interesting did you think this story was?"*) and then instructed the student to choose an answer on the scale of 0 – 3. Then the data collector then read the second question ("*How much did you know about this story* before you read it?") and instructed the student to choose an answer on the scale of 0 – 3.

Passage specific word list

Immediately following the student rating the passage specific word list was given. The word list included the first 50 words from a passage, randomly ordered and presented in columns. The student was asked to read the word list for one minute. The recorded score was the number of words read correctly and the number of errors at the end of one minute. When students came to the end of the word list and still had time left in their minute, they were instructed to begin again at the top of the first column. Therefore, it was possible that students would have a score of more than fifty words in one minute.

At the conclusion of each week's data collection session, the student was thanked for his or her help, given an incentive for the week and sent back to the classroom. Students were given a movie pass after the final data collection session in week 4.

Post-Data Collection Debriefing

After the completion of the four weeks of data collection, the school was thanked for participating and notified that they would be provided with a copy of the final results when analysis was completed.

Confidentiality

Prior to data collection all students were assigned randomly generated identification numbers. During data collection identification numbers were used in place of names on all assessment materials. Names and identification numbers were written on a removable label that remained on the folder in which each student's assessment materials were held until the end of data collection when the labels were removed and shredded in accordance with the district procedures for shredding of confidential materials. Names were kept on these folders to allow identification numbers to be linked to students throughout data collection as well as because these materials were made available to schools for use in instructional planning during the data collection process. Additionally, all data accessed from the DIBELS database (i.e., Spring benchmark data) was stripped of identifying information once it was linked to students' identification numbers. Following data collection, data entry and data analysis all assessment materials have been stored in a locked filing cabinet in the home office of the primary investigator and will be kept on file for five years.

Data Analysis

Hierarchical Linear Modeling (HLM) techniques have been instrumental in educational research due to their ability to investigate the relations between predictors and outcomes within nested structures (Raudenbush & Bryk, 2002). HLM allows modeling of how predictors at each level of the nested structure influence the outcome variable. This study was designed to produce data of this nested nature. For example, the data includes multiple assessments of oral reading fluency for each student (i.e., repeated assessments within students) as well as predictors at both the assessment (i.e., passage-level factors) and the student-level (i.e., initial skill). It is for these reasons that a two-level, repeated assessment within subjects HLM analysis was conducted to investigate the effects of both the passage-level and student-level predictors on the intercept and slope of student progress in oral reading fluency.

Prior to HLM analysis, all data were entered into a SPSS data file. Variables were created for both passage-level (level-1) and student-level (level-2) data. Table 2 provides names and descriptions for all variables, definitions of codes for the coded variables and delineates which variables were included at each level.

The passage-level variables included student identification numbers, data collection time points, weekly DORF score, weekly RTF score for the passage, weekly comprehension score for the passage, weekly rating of interest in the passage, weekly rating of prior knowledge about the passage content, weekly word list score corresponding to the passage, the background knowledge pre-test score corresponding to the passage, the average readability score for the passage, and the type of passage read.

The variable for time was coded as the number of instructional days from the date of the first data collection session. The variable for interest in the passage was coded from 0 – 3, with 0 being the lowest rating of interest and 3 being the highest. Likewise, the variable for prior knowledge of passage content was coded from 0 – 3, with 0 being the lowest rating of prior knowledge about passage content and 3 being the highest. The variable for the type of passage read was coded as 0 or 1, with 0 being for a narrative passage and 1 being for an expository passage. In the data file, the word list corresponding to the DORF passage for that week was included in that week's data, however, this word list would have been given to the student in the week following its corresponding DORF passage (or in the case of the final DORF passage, its corresponding word list would have been given in week 1 of data collection).

The student-level variables included a variable for each participant's score on the three spring benchmark passages from the spring benchmark assessment, the median score from the benchmark assessment, a score corresponding to the average variability in scores across the three benchmark scores, a coded variable corresponding to the level of variability in benchmark scores, and a coded variable corresponding to the risk category assigned to the median benchmark score.

The variable for the average variability across the three benchmark passages was created by subtracting the lowest score from the highest score and dividing by two (e.g., (high – low)/2)). The coded variable for level of average variability on the benchmark passages was based upon a median split of this average variability score. The median variability was 9.25 words. Therefore, the variable for the level of average variability for each participant was coded as either 0 or 1, with 0 being for below median variability and 1 being for above median variability. The variable for the median score on the benchmark assessment was used to create a coded variable that sorted participants according to the criteria for risk level for the spring DIBELS benchmark time (i.e., “at risk,” “some risk,” “low risk”). Results of this coding indicated that only two students would fall into the “at risk” category. For this reason, the “at risk” category was combined with the “some risk” category, resulting in a coding for this variable of 0 or 1, with 0 being At/Some Risk and 1 being Low Risk.

Table 2.

Variable Names, Descriptions and Coding Definitions

Variable Name	Variable Description	Coding
Student identifier variable		
Student ID	Student identification number	
Passage-level variables		
Time	Time in instructional days from first collection point	
Background	Background Knowledge Assessment (Pre-Test), percent correct for the passage	
DORF	Words read correct on weekly DIBELS Oral Reading Fluency passage	
RTF	Total words used in weekly Retell Fluency for the passage	

Table 2. cont.

Variable Name	Variable Description	Coding
Comprehension	Percent correct of attempted passage specific comprehension questions	
Interest	Rating of interest in the passage	0 = not interesting 1 = a little interesting 2 = pretty interesting 3 = very interesting
Prior Knowledge	Rating of prior knowledge about the passage content	0 = nothing 1 = a little bit 2 = some 3 = a lot
Readability	Average readability score for the passage	
Type	Type of passage read	0 = narrative 1 = expository
Word List	Number of words read correct for corresponding word list	
Student-level variables		
Risk	Corresponding risk category for median benchmark assessment score	0 = At risk/Some risk 1 = Low risk
Variability	Average variability in benchmark passages scores coded by a median split	0 = less than 9.25 1 = greater than 9.25

Missing data

Participants who had missing DORF data were deleted. Due to the inter-linked nature of data from week to week, if a student missed one week's data collection, it ultimately affected three weeks' worth of data. Each week a student read a DORF passage and a word list, but the DORF passage and the word list given in any one session did not correspond to one another. The corresponding word list for any DORF passage would be presented in a different week. Therefore, if a student missed a week of data collection, that student would then be missing a word list score for a DORF passage presented in a previous session and would be missing a DORF passage score for a word list given in a subsequent session. Missing data affected two participants. Deleting these participants left 70 participants with complete data across all four weeks of data collection.

HLM process

The process for HLM analysis included (a) an unconditional model, (b) a growth model with time as the only variable, (c) a model with each Passage-level predictor in isolation added to the growth model, (d) an all Passage-level predictors combined growth model, (e) a "best fit" model with necessary and sufficient Passage-level predictors, and (f) preliminary examination of Student-level predictors in the context of the "best fit" Passage-level model. As analysis proceeded, both empirical evidence (e.g., decrease in residual variance component, significance of models) and theoretical rationales were taken into consideration. Results and interpretation are presented in the following chapters.

CHAPTER IV

RESULTS

Analyses were conducted on data from the seventy students who had complete data for the four weeks of data collection. Descriptive statistics are reported first with the HLM analysis results presented in the order the models were run.

Descriptive Statistics

Descriptive statistics for the Passage-level and Student-level variables are presented in Table 3. The results indicate a relatively high performing sample. The mean benchmark score for the Spring DIBELS assessment for students was almost 18 words per minute higher than the 110 words per minute benchmark goal and the mean performance during progress monitoring was near 131 words per minute. This is further support for combining the “at risk” and “some risk” categories into one during subsequent analyses.

Table 3.

Descriptive Statistics for Passage-level and Student-level Variables

Passage-level Descriptive Statistics					
Variable Name	*N*	*M*	*SD*	*Minimum*	*Maximum*
Time	280	6.94	5.62	0	15
Background	280	84.46	13.27	50	100
Comprehension	280	85.33	15.93	0	100
DORF	280	130.83	35.51	18	224
RTF	280	49.38	18.42	9	96
Interest	280	1.94	0.72	0	3
Prior Knowledge	280	1.61	0.86	0	3
Readability	280	-0.03	0.98	-1.3	1.4
Type	280	0.50	0.5	0	1
Word list	280	92.70	22.10	9	187
Student-level Descriptive Statistics					
Variable Name	*J*	*M*	*SD*	*Minimum*	*Maximum*
Benchmark median score	70	127.93	31.46	21	224
Average variability	70	9.68	5.15	1.5	24.5
Benchmark passage 1	70	123.61	33.31	19	224
Benchmark passage 2	70	127.30	33.02	24	228
Benchmark passage 3	70	133.74	30.47	21	238

HLM Models

HLM analysis was conducted to examine the effects of passage-level and student-level predictor variables on the outcome variable, DORF. At each step of analysis, both fixed effects random effects were examined. Fixed effects were examined to determine whether changes in the intercept (i.e., changes in the outcome variable) were significantly greater than what would be expected by chance. Random effects were examined to determine if the changes in the outcome variable varied significantly across students. Additionally, the overall "fit" of the models was examined through interpretation of the reduction in residual variance from model to model. A reduction in residual or unexplained variance was an indication that the model better accounted for the data. Further, the amount of passage-level (level 1) variance explained by the model was calculated for those models showing a reduction in residual variance.

Analysis began with the examination of an unconditional model with no predictor variables entered at either level. The unconditional model provided baseline model statistics that were used in evaluating subsequent models. Next, a growth model with time as the only variable was examined to determine whether time had a significant effect on the outcome variable (i.e., DORF) and whether this effect varied significantly across students.

The next models examined the effect of each passage-level predictor in isolation given the growth model. For each model, fixed effects were examined to determine significant effects on the outcome variable and random effects were examined to determine whether effects varied significantly across students. Additionally, examination of the reduction in the residual variance component was used as an indication of the magnitude of the relation to passage performance at level 1. That is to say, a greater reduction in level 1 residual variance would be an indication of a stronger relation between the variable and level 1 passage performance.

Next, a model including all passage-level predictors combined with the growth model was examined. Again, fixed effects and random effects were examined to determine the effect of each predictor in the context of all other predictors. Those predictors that resulted in significant, necessary, and sufficient effects were then included in a “best” model for the subsequent student-level models.

The final passage-level model was a “best” model with necessary and sufficient passage-level predictors. This model was constructed so as to include any predictor that had continued to result in a significant fixed effect in the context of the all predictors model and exclude any predictors that, if added would not result in significant fixed effects.

The final step in analysis included preliminary examination of student-level predictors in the context of the “best” passage-level model. Again, results were examined to determine any significant fixed and/or random effects as well as any reduction in residual variance compared to preceding models. Results of the models are presented next.

Unconditional Model

To determine the magnitude of difference in DORF scores across students, an unconditional model was run with DORF as the outcome measure and no predictor variables at either level 1 or level 2. The specific model was:

Level 1 Model

$Y = P0 + E$

Level 2 Model

$P0 = B00 + R0$

Results of the unconditional model (see Table 4) indicted that the grand mean DORF performance was 130.36 words per minute with a standard error of 3.52 words per minute. This was significantly different from zero ($p < 0.001$). The significance of the random effect demonstrates that performance across students varies significantly, indicating the ability to further investigate that variability through the inclusion of predictor variables.

Table 4.

Results of the Unconditional Model

Fixed Effect	*Coefficient*	*se*	*t-Ratio*	*Approx. df*	*p-Value*
Grand mean	130.36	3.52	37.03	69	< 0.001
Random Effect	*Variance Component*	*sd*	*χ2*	*df*	*p-Value*
Variance explained	769.66	27.47	564.77	69	< 0.001
Residual variance	420.92	20.52			

Growth Model

Time was included as a random effect for both theoretical and empirical reasons. Including time as a random effect means that each student would be allowed to have his or her own individual growth rate across time. The growth rate applied to each student's data would differ according to their specific performance across the four data points. This assumes that students grow at different rates across time and that the difference between growth rates is significant and meaningful to interpret. To include Time as a fixed effect would mean that the grand mean growth rate across students would be applied to each student's data. This would mean that during analysis, the same growth rate would be applied to all students and would not be allowed to vary randomly across students. The underlying assumption would be that all students grow at the same rate across time or that their growth rates do not vary significantly and therefore it would be more parsimonious for interpretation to apply the same "fixed" growth rate across students during analysis.

Prior to running the growth model, it was the theoretical assumption of the primary investigator that students would grow at different rates, and therefore growth rates should be allowed to vary across individuals. For this reason, the model was constructed with time as a random effect. The specific model was:

Level 1 Model

$$Y = P0 + P1*(\text{TIME}) + E$$

Level 2 Model

$$P0 = B00 + R0$$

$$P1 = B10 + R1$$

Results of the model (see Table 5) provided empirical support for the theoretical assumption. Whereas the average coefficient is small and not significantly different from zero ($B10 = 0.08$, $p > .05$), the significance of the level 1 variance component (var($R1$) = 0.93, $p < 0.05$) indicates that growth does vary significantly across students and may be larger and positive for some students while being smaller and/or negative for others.

It should also be noted that Time was modeled as linear growth rather than curvilinear. Analysis comparing a linear growth model to a curvilinear growth model indicated significant variation in growth over time across students for time as a linear function whereas time as a curvilinear function did not demonstrate a significant effect.

Table 5.

Results of the Growth Model with Time as a Random Effect

Fixed Effect	*Coefficient*	*se*	*t-Ratio*	*Approx. df*	*p-Value*
Intercept	129.82	4.04	32.15	69	< 0.001
Time	0.08	0.24	0.35	69	0.729
Random Effect	*Variance Component*	*sd*	*χ2*	*df*	*p-Value*
Intercept	906.44	30.11	325.32	69	< 0.001
Time	0.93	0.96	91.25	69	0.038
Residual	384.47	19.61			

The fixed effect coefficient indicates that, for each instructional day increase, on average a student's DORF score increased by 0.08 words per minute. A comparison of the residual variance components from the unconditional model and this growth model indicate that residual variance was reduced by 36.45. Raudenbush & Bryk (2002, p. 79) provide an equation for determining the "proportion reduction in variance, or 'variance explained' at level 1." The equation is:

Proportion variance explained at level 1 =

$$\frac{\text{residual variance(Unconditional)} - \text{residual variance (Time)}}{\text{residual variance(Unconditional)}}$$

Applying this equation to the residual variance terms shows that Time accounts for 8.7% of the level 1 residual variance in a student's passage scores. This means that although the average coefficient is not significantly different from zero, students do differ significantly in their individual coefficients and that this variability across students accounts for 8.7% of the total variation in student scores.

Individual Passage-level Predictor Models

The following eight models each included one, un-centered, passage-level variable in isolation, given the time random growth model. Variables were entered into models as fixed effects for both theoretical and empirical reasons. It was theoretically assumed that while different students might have different levels of background information or interest in a passage topic or prior knowledge about the passage content, one wouldn't expect the importance of the relations between these variables and the outcome variable to differ across students. This theoretical assumption was empirically supported when preliminary models, run with the variables as random effects resulted in no significant variability across students (i.e., $p > .05$) for all passage-level variables except Word List ($p < .001$) (see Appendix F for a summary of the variance components from these preliminary models). Although, Word List indicated significant variability across students, it was also entered as a fixed effect in subsequent models because no strong theoretical rationale for why Word List would be differentially important across students could be identified. Additionally, it was the preference of the primary investigator to keep analysis as parsimonious as possible. Thus, all eight passage-level variables were entered into the models as fixed effects. The specific model was:

Level 1 Model

$Y = P0 + P1*(\mathrm{TIME}) + P2*(\mathrm{PREDICTOR}) + E$

Level 2 Model

$P0 = B00 + R0$

$P1 = B10 + R1$

$P2 = B20$

The results of each of the passage-level predictor models are summarized in Table 6.

Table 6.

Summary of Individual Passage-level Predictor Models

Level 1: Passage-level variable	*Effect*				*Level 1Residual Variance (% Reduction from comparison model)*
	Coefficient	*se*	*t Ratio*	*p Value*	
Comparison Model: Time as a Random Level 1 Effect					
Time					384.47
Individual Level 1 Predictor Models					
Background	0.36	0.09	3.90	<0.001	400.05 (0%)
Comprehension	0.11	0.07	1.64	0.102	386.79 (0%)
RTF	0.41	0.09	4.82	<0.001	385.01 (0%)
Interest	-7.83	2.48	-3.15	0.002	389.24 (0%)
Prior knowledge	-1.78	1.95	-0.92	0.360	384.59 (0%)
Readability	-14.01	0.71	-19.77	<0.001	162.45 (58%)
Type	-26.62	1.56	-17.03	<0.001	182.22 (53%)
Word list	0.45	0.14	3.18	0.002	409.51 (0%)

Note. Negative reduction in level 1 residual variance compared to the time random effect only model is reported as 0% reduction for interpretability.

The interpretation of results and the criteria for inclusion in subsequent models focused on two factors—significance of the coefficient and amount reduction in level 1 residual variance. The first is an indication that the effect of the variable on the outcome was more than what would be expected by a chance occurrence, and the second is an indication of the magnitude of the effect, that is, how "important" the effect is to the student's performance.

Significant effects with reduction in level 1 residual variance

The two variables that showed the largest impact on DORF scores and resulted in the largest reduction of residual variance were Readability and Type of passage. Both effects were significant ($p < .001$) and negative. Readability is an indication of the difficulty of a passage such that a passage with a larger readability estimate is assumed to be a more difficult passage. Results indicated that each unit increase in Readability was associated with a decrease of 14.01 words per minute in a student's DORF score, meaning that as passage difficulty increased, student scores decreased. Type of passage was a coded variable indicating whether a passage was narrative (coded as 0) or expository (coded as 1). Results indicate that on average a student's DORF score decreased by 26.62 words per minute when reading an expository passage as compared to reading a narrative passage.

In addition to the significant effects, both models resulted in a reduction of residual level 1 variance in comparison to the growth model. A reduction in residual level 1 variance or unexplained variance is an indication of the model accounting for the data better than the comparison model, which means a larger proportion of the variability in student performance can be attributed to known variables (i.e., the variables included in the model). The proportion of variance explained was calculated for these two models through the application of the equation discussed previously on the residual variance components from the growth model and the individual predictor models. Results indicate that 58% of the variance in students DORF scores could be explained by or attributed to the readability of the passage and 53% of the variance in DORF scores could be explained by or attributed to the type of passage read.

Significant effects without reduction in level 1 residual variance

The remaining variables, with the exception of Comprehension and Prior Knowledge, resulted in significant effects, however none of these effects were paired with a reduction in the level 1 residual variance as compared to the growth model. In fact, all models demonstrated an increase in the level 1 residual variance component with the inclusion of the predictor variable.

Snijders and Bosker (2002) provide a detailed explanation and discussion of how this can occur within HLM and provide alternate means of determining the proportion of explained variance in their chapter. Sufficient for the purposes of this study, their chapter concludes that variance reduction in HLM can be reapportioned between level 1 and level 2. This means that the inclusion of a predictor variable at level 1 may be significant by substantially reducing level 2 residual variance, while increasing level 1 residual variance. They further propose that an increase in residual variance with the addition of a predictor variable is possibly an indication of "misspecification" of the predictor variable when entered as a fixed effect in the model. It is possible that some of the variables in the current study may have been partially misspecified and may have been better defined as a combination of a level 2 variable (e.g., mean score across all passages) and a level 1 variable that represents the possible passage level effects (e.g., level 1 variables as deviation scores calculated from the difference between a student's mean performance and his or her specific performance on each passage). This possible misspecification might account for the inclusion of the level 1 predictor variables resulting in increased level 1 residual variance, while decreasing, in some cases substantially, level 2 residual variance.

Word List is one such variable. The effect for Word List was small, positive and significant at $p < .01$, meaning the average effect for each word read correct on a word list, was an increase of 0.45 words per minute on DORF scores. Word List resulted in the largest increase of level 1 residual variance (almost 7%) while decreasing level 2 *R0* residual variance by 35% and level 2 *R1* residual variance by 77%. It is probable that inclusion of a level 2 word reading variable, such as a student's mean word reading across all word lists or a different level 1 variable corresponding to a student's deviation from their mean word list score for each passage would have better represented the effects of word reading on oral reading fluency. It is important to note, however that the focus of this study was on explaining variance at level 1 and it is unlikely that respecifying the word list variable would change the conclusion regarding its contribution to level 1 variance.

Background had a small, positive effect on DORF score and was significant at $p < .001$. The coefficient indicates that, on average, each percentage point increase on a student's background knowledge assessment, resulted in an increase of 0.36 words per minute on the student's DORF score. Background also resulted in an increase of 4% in level 1 residual variance while substantially decreasing level 2 residual variance; *R0* residual variance decreased by 17% and *R1* residual variance decreased by 59%. Similarly to Word List, it is possible that a level 2 variable measuring a student's mean background knowledge performance across all passages and a level 1 variable corresponding to the deviation of each passage from the student's mean performance may have been more suitable for this analysis. Again, the focus of the present study was on explaining variance at level 1 and it is possible that the interpretation of the level 1 contribution would not change with partitioning between a level 1 and a level 2 variable.

RTF also had a small, positive effect on DORF score that was significant at $p < .001$. Interpretation of the coefficient for RTF indicates that on average, each word used in retelling the passage, resulted in an increase of 0.41 words per minute on the DORF score. RTF demonstrated only a slight increase in level 1 residual variance (less than 1%) and its reduction in level 2 residual variance was substantial; 20% for *R0* residual variance and 42% for *R1* residual variance. Again, this may indicate that RTF should be respecified to a combination of a level 1 variable and a level 2 variable (e.g., the percent of words read used in the retell or a mean retell score across passages, etc.).

Interest resulted in an effect that was larger and negative ($p < .01$). Interpretation of the coefficient indicates that for each unit increase in rating on the measure (i.e., the higher a student rated his or her interest in the passage content), the DORF score decreased by 7.83 words per minute. Interest demonstrated an increase in level 1 residual variance of 1% and decreases in level 2 residual variance of 11% for *R0* and 53% for *R1*. It is thought that a student's rating of their interest in the passage may have been an indication of how deeply they had engaged with the text. Perhaps, in an effort to comprehend and retain the passage content due to a higher level of interest in the topic, students took more time with reading, which resulted in a reduced DORF score.

Whereas the models for Word List, Background, RTF and Interest resulted in significant effects on the outcome variable, an examination of the residual variance components did not indicate any reduction in residual variance at level 1 (i.e., did not better account for the data at level 1). This is an indication that although the effect on the outcome variable was significantly larger than zero, the effect was not of large enough magnitude to explain more of the variability in student scores beyond what would be explained by Time alone, or alternatively, that inclusion of these variables reduced residual variance at level 2, but not at level 1. The current study was focused on explaining variance at level 1 and therefore, only those variables that demonstrated reductions in residual variance at level 1 were considered for subsequent models.

All Passage-level Predictors Combined Growth Model

A model including all passage-level predictors was constructed to determine how each predictor would affect DORF scores in the context of all the other predictors. All passage-level predictors, with the exception of Time, were added to the time as a random effect growth model, as un-centered, fixed effects. The specific model was:

Level 1 Model

$$Y = P0 + P1*(\text{TIME}) + P2*(\text{BACKGROUND}) + P3*(\text{RTF}) + P4*(\text{COMPREHENSION}) + P5*(\text{INTEREST}) + P6*(\text{PRIOR KNOWLEDGE}) + P7*(\text{READABILITY}) + P8*(\text{TYPE}) + P9*(\text{WORD LIST}) + E$$

Level 2 Model

$$P0 = B00 + R0$$

$P1 = B10 + R1$

$P2 = B20$

$P3 = B30$

$P4 = B40$

$P5 = B50$

$P6 = B60$

$P7 = B70$

$P8 = B80$

$P9 = B90$

Results of this model (see Table 7) indicated that only three predictors maintained their significant affects within the context of all other predictors. They were Background ($p < .05$), Type ($p < .001$) and Readability ($p < .001$). However, the coefficient for Background was now negative (-0.16) whereas it was previously positive (0.36) in the model with it as a single predictor. It is thought that Background may be acting as a suppressor variable in this instance. A suppressor variable is defined as a variable that increases the predictive validity of another variable or set of variables. Background acting as a suppressor variable would mean that the presence of Background in the model is reducing what would otherwise be unexplained variance associated with a different variable due to some level of correlation between Background and the other variable. By reducing the unexplained variance associated with the other variable, the explanatory effect of the other variable is increased. The effects of suppressor variables, even if significant, can be very difficult to interpret. An examination of correlations between Level 1 predictors shows that Background was significantly correlated with five other predictors: Word Reading (r = .29, p < .01), Retell Fluency (r = .27, p < .01), Type of passage (r = -.26, p < .01), Readability (r = -.25, p < .01), Prior Knowledge (r = .15, p < .05). Also, it should be noted that this change in effect from positive to negative could also be related to the possible misspecification of the variable discussed previously. These two factors, the possibility of Background acting as a suppressor variable and the possible misspecification of the variable, were rationale for not including Background in the Best Growth Model, which is described next.

"Best" Growth Model with Necessary and Sufficient Passage-level Predictors

Next, a "best" growth model was designed (see Table 8). The goal was to create a model whereby all predictors included were significant, while not leaving out any predictor that would have a significant, interpretable effect if added to the model. Based upon results from the previous models, a model was created using the variables Type and Readability. Type and Readability consistently demonstrated a significant effect on the outcome as well as a large magnitude of effect as demonstrated through reduction in level 1 residual variance components. Also, given a model with Time, Readability and Type, no other variable was significant or reduced substantial level 1 residual variance. Background was an exception in that it did produce a significant effect and reduced level 1 residual variance by 2% compared to the Time, Readability and Type model. However the effect remained small and negative, which may be an indication that it is acting as a suppressor variable (see discussion above) and would be difficult to interpret theoretically. Additionally, the reduction in residual variance was not considered to be substantial. Therefore, the specific "best" model was:

Level 1 Model

$Y = P0 + P1*(\text{TIME}) + P2*(\text{READABILITY}) + P3*(\text{TYPE}) + E$

Level 2 Model

$P0 = B00 + R0$

$P1 = B10 + R1$

$P2 = B20$

$P3 = B30$

Table 7.

All Passage-level Predictors Model

Fixed effect	*Coefficient*	*se*	*t Ratio*	*Approx. df*	*p Value*
Intercept	145.84	10.07	14.48	69	< 0.001
Time	0.04	0.15	0.30	69	0.763
Background	-0.16	0.06	-2.52	262	0.013
Comprehension	-0.01	0.05	-0.24	262	0.812
Interest	-1.51	1.55	-0.97	262	0.334
Type	-14.51	2.24	-6.48	262	< 0.001
Prior Knowledge	1.19	1.35	0.88	262	0.378
Readability	-7.98	0.99	-8.09	262	< 0.001
Retell Fluency	0.04	0.06	0.62	262	0.539
Word Reading	0.05	0.06	0.80	262	0.426
Random effect	*Variance Component*	*sd*	χ^2	*df*	*p Value*
Intercept	792.34	28.15	634.62	69	< 0.001
Time	0.30	0.54	84.76	69	0.096
Residual	150.38	12.26			

Table 8.

"Best" Growth Model with Necessary and Sufficient Passage-level Predictors

Fixed effect	*Coefficient*	*se*	*t Ratio*	*Approx. df*	*p Value*
Intercept	136.99	4.07	33.63	69	< 0.001
Time	0.06	0.14	0.45	69	0.655
Type	-7.73	0.99	-7.78	268	< 0.001
Readability	-14.68	2.20	-6.67	268	< 0.001
Random effect	*Variance Component*	*sd*	χ^2	*df*	*p Value*
Intercept	806.53	28.40	646.16	69	< 0.001
Time	0.25	0.50	83.50	69	0.113
Residual	150.69	12.28			

Results of the model indicate that (a) a student on average read 137 words per minute at the beginning of progress monitoring, (b) given Time and Type of passage, a one-unit increase in Readability accounts for a decrease in DORF score of about 15 words per minute and (c) given Time and Readability, the effect of Type is a difference of about 8 words per minute. These results mean that a student reading the "Elephants" passage (i.e., the passage with the highest average readability (1.4) and categorized as an expository passage), would be expected to read approximately 25 fewer words than he or she would when reading a passage at the mean readability estimate (-0.03) and the mean passage type (0.5).

To calculate the effect of Readability in the above example, one must first calculate the difference between the mean readability (-0.03) and the readability of the passage (1.40), which is a difference of 1.37. Next, this number is multiplied by the -14.68 words per minute coefficient from the model, which results in a total decrease of about 21 words per minute. Likewise, to calculate the effect of Type in this example, one must first calculate the difference between the mean passage type (0.5) and the actual type of passage (1.0; expository was coded as 1.0), which is a difference of 0.5. Next, this number is multiplied by the -7.73 words per minute coefficient from the model, which results in a decrease of about 4 words per minute (The difference for a narrative passage would result in an increase of about 4 words per minute, for a total Type effect of about 8 words per minute.). These two calculations are added together to create the total effect of Readability and Type for the "Elephants" passage.

Application of the same calculations for "Strawberry Jam," the narrative passage with the lowest readability estimate (-1.3) indicate that a student would be predicted to read about 23 more words per minute than a student reading a passage at the mean readability and mean passage type. The difference between the predicted student performance on the "Strawberry Jam" passage and the "Elephants" passage would be about 48 words per minute, with scores on "Strawberry Jam" being higher.

A comparison of the level 1 residual variance for the Time random growth model indicates that this "best" model accounts for 61% of the variance in student scores.

Preliminary Exploration of Student-level Predictors

Once the "Best" model was identified, then the effect of Student-level predictors was explored. The Student-level analysis models are considered preliminary due to the minimal level of complexity of the models constructed. The small sample size of the study (i.e., only 70 students at level 2) did not allow for the Student-level variables (i.e., Risk category and Average variability) to be separated into more than two groups each, nor was it large enough to support more complicated analysis with interaction terms at level 2. The sample size, being what it was, also served as reason for not including more variables for investigation at level 2. For these reasons, the Student-level models were created as simple, two group, main effects models, results focus on main effects only and the analyses are referred to as preliminary.

Low risk vs. at risk/some risk

First, a model was created where a students' risk status based upon the Spring DIBELS benchmark score was added into the model at level 2. Risk was a coded variable referring to a student being in either the "At risk/Some risk" group (coded 0) or the "Low risk" group (coded 1). Results are presented in Table 9. The specific model was:

Level 1 Model

$$Y = P0 + P1*(\text{TIME}) + P2*(\text{READABILITY}) + P3*(\text{TYPE}) + E$$

Level 2 Model

$$P0 = B00 + B01*(\text{RISK}) + R0$$

$$P1 = B10 + B11*(\text{RISK}) + R1$$

$$P2 = B20 + B21*(\text{RISK})$$

$$P3 = B30 + B31*(\text{RISK})$$

Table 9.

Best Passage-level Growth Model with Student-level Risk as Predictor

Fixed effect	*Coefficient*	*se*	*t Ratio*	*Approx. df*	*p Value*
Intercept	103.04	5.27	19.54	68	< 0.001
Risk	45.62	6.66	6.85	68	< 0.001
Time					
Intercept	0.21	0.24	0.87	68	0.389
Risk	-0.17	0.30	-0.57	68	0.570
Readability					
Intercept	-6.54	1.42	-4.61	264	< 0.001
Risk	-1.46	1.91	-0.76	264	0.446
Type					
Intercept	-10.42	3.17	-3.29	264	0.002
Risk	-5.88	4.16	-1.41	264	0.159
Random effect	*Variance Component*	*sd*	χ^2	*df*	*p Value*
Intercept	471.99	21.73	397.96	68	< 0.001
Time	0.20	0.44	79.69	68	0.157
Residual	150.39	12.26			

Results of the addition of Risk as a Student-level predictor demonstrate a significant effect on the intercept, meaning that students in the “Low risk” category read on average about 46 more words per minute (p < .001) than their counterparts in the “At risk/Some risk” category. Risk did not have a significant interaction with Readability or Type of passage. In comparison to the best passage-level model, this model reduced level

1 residual variance by less than 1%. An examination of residual variance at level 2 indicates a reduction in *R0* of 41%, with an increase in *R1* of 25%.

Higher variability vs. lower variability

Next, a model was created where a student's variability across the three spring benchmark scores was added in as a predictor at level 2. Variability was a coded variable where students were split into a "lower than median variability" group (coded as 0) and a "higher than median variability" group (coded as 1). Results of this model are presented in Table 10. The specific model was:

Level 1 Model

$Y = P0 + P1$*(TIME) + $P2$*(READABILITY) + $P3$*(TYPE) + E

Level 2 Model

$P0 = B00 + B01$*(VARIABILITY) + $R0$

$P1 = B10 + B11$*(VARIABILITY) + $R1$

$P2 = B20 + B21$*(VARIABILITY)

$P3 = B30 + B31$*(VARIABILITY)

Results indicate that Variability did not have a significant effect on the intercept. The only significant effect of Variability was on Time. Results indicate that students with higher than the median variability on average would show an additional .76 word increase for each additional instructional day ($p < .01$). It is possible that there is some confound between Time and Variability that is accounting for this. In comparison to the best passage-level model, there was no reduction in level 1 residual variance for this model. Examination of the residual variance components at level 2 indicates an increase in residual variance for *R0* and a reduction of 52% for *R1.*

Table 10.

Best Passage-level Growth Model with Student-level Variability as Predictor

Fixed effect	*Coefficient*	*se*	*t Ratio*	*Approx. df*	*p Value*
Intercept	136.56	6.02	22.68	68	< 0.001
Variability	0.22	8.07	0.03	68	0.979
Time					
Intercept	-0.31	0.18	-1.75	68	0.084
Variability	0.76	0.27	2.79	68	0.007
Readability					
Intercept	-8.03	1.31	-6.12	264	< 0.001
Variability	0.53	1.95	0.27	264	0.785
Type					
Intercept	-13.18	2.82	-4.68	264	< 0.001
Variability	-1.72	4.25	-0.41	264	0.685
Random effect	*Variance Component*	*sd*	χ^2	*df*	*p Value*
Intercept	822.87	28.69	640.48	68	< 0.001
Time	0.12	0.35	73.74	68	0.296
Residual	152.27	12.34			

Summary of Results

The major findings of the current study demonstrated that there was significant variability across students in oral reading fluency (based on the significant results of the unconditional model) and that this variability could be attributed to different level 1 and level 2 predictor variables. A linear growth model including Time as a random effect (i.e., growth rates were allowed to vary across students) resulted in a non-significant effect of Time on the average DORF score, however it did demonstrate significant variability across students, supporting the theoretical position that students grow at different rates on oral reading fluency.

When added to the growth model as single predictor variables (i.e., the only level 1 variable in the model given Time random), significant effects were found for all level 1 predictors, with the exception of Comprehension and Prior Knowledge. The proportion of reduction in level 1 residual variance was also examined for each predictor that produced a significant effect and those variables demonstrating both a significant effect and a reduction in level 1 residual variance were included as predictors in subsequent models. Readability and Type of passage met both criteria by demonstrating large effects that were significant (p < .001) and producing substantial reductions in level 1 residual variance (58% and 53%, respectively).

The observed significance of effect for Background, Word Reading, Interest, and RTF without reduction in level 1 residual variance was perplexing, however, examinations of the level 2 residual variance components indicate substantial reductions in level 2 residual variance for these variables, with Background and Word List demonstrating the largest reductions. This may be indicative of possible misspecification of these level 1 variables (Snijders and Bosker, 2002) and will be discussed further as a limitation of the study in the following chapter.

At level 2, preliminary analysis indicated a significant effect of Risk on DORF score and a significant effect of Average Variability on Time. Neither level 2 predictor demonstrated a significant effect on the level 1 predictors of Readability and Type of passage.

Discussion of the significant results in relation to the specific research questions and the larger educational research context is presented in the following chapter. Limitations of the study, considerations for future research and implications for practice are also presented.

CHAPTER V

DISCUSSION

Although formative assessment procedures (e.g., progress monitoring) have been used for decades to determine progress toward proficiency in basic skills (e.g., curriculum based measurement, oral reading fluency, etc.), recent changes in special education legislation, have made it possible to use formative evaluation procedures as a means of determining special education eligibility for students suspected of having a specific learning disability. For some educators and researchers this is troublesome because it means that formative evaluation procedures that have historically been used to make what some call "low stakes" decisions (i.e., instructional placement, intervention effectiveness), will now be used to make much "higher stakes" decisions (i.e., special education eligibility decisions). If indeed the eligibility decision is a higher stakes decision, then educators and researchers will want to be sure that the measures they are using to gather progress data are the most reliable and valid they can be because the reliability and validity of that data will directly affect the reliability and validity of the decisions made during the evaluation process. Variability in progress monitoring interferes with our ability to make reliable and valid decisions about student progress. If we can understand the reasons for passage variability, we can control or reduce that variability and improve educational decisions about progress.

The current study used HLM analysis to examine the effects of both Passage-level (level 1) and Student-level (level 2) predictor variables on the oral reading fluency progress monitoring performance of third grade students. Since variability in progress monitoring scores interferes with decision-making regarding student progress, it was the goal of the study to identify factors that contribute to the variability in student scores from week to week and the relations between these factors and oral reading fluency. If these relations are better understood then assessments and procedures could be modified to better account for the effects of these factors and the validity of the educational decisions being made would be improved. In this chapter, the findings of the study are discussed in relation to the research questions and the context of current educational research. Limitations of the study, considerations for future research and implications for practice are also presented.

Results in Relation to Research Questions

Both research questions looked at the effect of predictor variables on oral reading fluency performance over time. The results of the time only linear growth model indicated a non-significant effect of time on the intercept, meaning that the average change in student score from week to week was not significantly different from zero and indicated a significant difference in individual student growth across students, meaning students had different growth rates. It is likely that an examination of individual student growth rates would show that some students are demonstrating growth curves well below or well above the mean growth demonstrated in the results of the model. It may be the case that any curvilinear growth is being masked by the differences individual growth curves. An examination of individual growth curves could provide more information to aid in the interpretation of the results of the model.

Effect of Passage-level Factors

The first research question investigated the relation between eight passage-level factors and oral reading fluency progress monitoring performance for the sample of third grade students.

Results of the HLM analysis indicated that when entered as individual predictors, six of the eight factors demonstrated significant effects on student oral reading fluency. The six factors were (a) passage specific background knowledge, (b) a student rating of interest in the passage topic, (c) the number of words used to retell the passage, (d) the average readability estimate of the passage, (e) the type of passage (i.e., narrative or expository) and (f) the number of words read from a passage specific word list. The effects of passage specific comprehension and a student's rating of prior knowledge about the passage content were not significant.

The largest magnitude and significant effects for passage-level factors were found for the average readability (or difficulty) of the passage and the type of passage. The findings indicate that both readability and type of passage demonstrated negative effects on oral reading fluency. The effects were on average, a decrease of 15 words per minute for readability and 8 words per minute decrease for passage type. Additionally, both variables produced substantial decreases in level 1 residual variance with readability producing a reduction of 58% and type of passage, 53%.

The result that readability had a large, significant and large magnitude effect on the oral reading fluency of students is in contrast to recent research by Ardoin, Williams, Christ, Klubnik & Wellborn (2010), in which the researchers concluded that readability estimates are inadequate for predicting oral reading performance or evaluating CBM passage difficulty. In the Ardoin, et al. study, the researchers investigated three commonly employed readability estimates—Lexile, Spache and Forcast while in the current study, the readability estimate used was a unit-weighted average of nine different readability estimates. It is possible that this could contribute to the differences in the findings. An examination of what specific components are used in the calculation of each readability estimate (e.g., number or percent of high frequency words, sentence length, word frequency, decidability, etc.) might provide more insight. Perhaps the average estimate used in the current study captured a broader or more representative sample of the components regularly used in readability formulas than did the three specific estimates investigated in the Ardoin, et al. study.

Type of passage also had a large, significant and large magnitude effect on oral reading fluency. Type of passage and readability accounted for large proportions of the reduction in level 1 residual variance as individual predictors (53% and 58%, respectively), and when included together in a model they accounted for 61% of the reduction in residual variance at level 1. Based on the proportion in reduction of residual variance from their individual predictor models, it appears that a large proportion of the combined reduction in level 1 residual variance is shared by these two variables (50%). Further examination shows that the proportion of the reduction in variance that can be uniquely attributed to readability is 8% and the proportion of the reduction in variance that can be uniquely attributed to type of passage is 3%. It is assumed that these two variables were confounded in this study based due to the large proportion of shared variance and the fact that type of passage and readability were related in the 4 specific passages selected for the study (i.e., the two expository passages were also the two passages with the highest readability estimates).

Given that the two expository passages in this study were also the two with the highest readability estimates (i.e., they were estimated to be the most difficult), one would expect them to have a negative effect on student scores, which they did. However, in addition to the effect of that greater difficulty, there is an indication that there is something else about the expository passages that results in an even larger decrease in oral reading fluency scores. This effect of genre or type may be related to something like vocabulary knowledge (e.g., number or percent of unknown words), prior knowledge, or text structure (Saenz & Fuchs, 2002). Future research investigating this effect could further our understanding of passage variability.

Effect of Student-level Factors

The second research question investigated the relation between two student-level factors and the progress monitoring performance of the sample of third grade students. The HLM models that included level 2 predictors were based upon the best model identified from the initial level 1 individual variable models, which included time, readability and type of passage as level 1 predictors. This level of analysis was considered preliminary due to the small sample size of the study precluding more complex models investigating interactions among the level 2 variables.

Results of the first main effects model indicate that initial skill (defined as benchmark level of risk) had a significant effect on the intercept term in student oral reading fluency progress with students in the low risk group reading more words at the beginning of progress monitoring. However, there was no significant relation between level of risk and rate of progress. In addition, level of risk did not interact with readability or type of passage. In other words, although students who were at risk or some risk were initially lower on the progress monitoring assessments, they made progress that was not significantly different, and the importance of readability and text type was not significantly different.

Results of the second main effects model indicate that the amount of variability in benchmark passages did not have a significant effect on the intercept term in student oral reading fluency progress; there was a significant relation between variability and rate of progress and also variability did not interact with readability or type of passage. In other words, although a student's variability did not affect their initial starting point for progress monitoring, students with higher variability made progress that was greater and significantly different from students with lower variability, but the importance of readability or type of passage was not related to variability. One explanation for this result could be floor effects. It may be that students who started low and did not make progress were bouncing along at near 0 words per minute and subsequently had lower variability. This would be an instance of lack of progress causing less variability rather than variability causing lower progress. In the current study neither variable (progress or variability) is manipulated so the suggested causal relationship is purely speculative. Although, with this explanation one might expect also to see a relation between variability and the intercept (i.e., students with lower variability were also students who had lower initial scores) and this was not found in the current study.

Limitations

The results of the current study must be interpreted with its limitations in mind. Limitations to the external validity of this study include the small sample size, the lack of a diverse and representative sample, and the specific setting where the study took place and time of the year in which the study took place.

The sample size in the current study was small (*N* = 70). This small sample size did not allow for more complex analyses at level 2 to investigate any possible interaction effects of initial skill and average variability. Additionally, the sample was relatively high performing (e.g., the mean benchmark score was 20 words per minute above the benchmark goal for the Spring of 3rd grade). The lack of representativeness across initial skill levels paired with the small sample size did not allow for the creation of more specific groups based on level of variability or a group at each level of risk for the level 2 analysis. The sample included only 2 students in the at risk category, which meant the at-risk and some-risk categories were combined into one group. Also, the sample was taken from one grade level, from one elementary in a small suburb in Iowa. The population of the district was neither ethnically, racially, nor socio-economically diverse. The exact ethnic, racial, and socio-economic makeup of the sample is not known, however it would be assumed to be similar to the district demographics. Future studies would want to examine larger, more diverse and representative samples in order to increase the external validity of the results.

The instructional setting and time of year in which the study took place may also affect the external validity of the results. The elementary school in which the study took place was implementing an RTI process for providing tiers of instructional support to students in reading. This may further reduce the generalizability of results to students in buildings where there is no RTI process in place. Also, the specific focus of the instruction in the building is unknown with regards to how much attention is given to teaching students to engage with expository text at this grade level or in previous grade levels. Additional information on how much emphasis and/or instruction is given to expository text would be beneficial to the interpretation of the effects of type of passage and would be essential to the generalizing of results to other students in other buildings. It is possible that results would be different for students in buildings where there is less or more emphasis on instruction in expository text. Also, this study took place during the last month of the school year. It is likely that growth rates found in this study would be different than at the start of the school year or if data were taken across the entire school year. Having data from across the whole school year would provide a more representative growth trajectory for students. Future studies would want to include information regarding specific instructional practices in place in the buildings as well as include data from a larger time span during the school year.

During training all data collectors were required to meet an inter-rater reliability of 90% on the measures, however no inter-rater reliability checks were conducted during data collection. Although the study took place across only 4 weeks, it is possible that administrator drift, or small deviations from standardized administration procedures or reliability of scoring could have taken place and had an effect on the quality of the data.

Prior to their use in data collection, the Background Knowledge Assessment and Passage Specific Comprehension Questions were not included in pilot studies nor were any preliminary analyses performed. Therefore, there are no data to provide evidence for their reliability or validity when assessing the constructs they were used to investigate in the current study. It would be important to conduct such analyses prior to their use in further research. Additionally, the inclusion of other broad measures of comprehension such as scores from a standardized reading comprehension test or results from the state assessment of reading comprehension were not used in the current study and could be used as student-level predictor of comprehension in future studies.

Although many of the models resulted in significant coefficients, the corresponding standard error for the coefficients was often large. This may be an indication of measurement concerns. It is possible that variables may have been overly correlated with one another, or the variables may not have allowed a good representation of the construct being assessed (e.g., comprehension, prior knowledge, interest, etc.), or the way in which the variable was used may have truncated variability across students (e.g., the median-split used for the level 2 average variability predictor). These measurement concerns in conjunction with the limitations discussed above regarding the small, non-representative sample and small time span of the study all would indicate that the statistical power of the study is likely very low.

Finally, the possibility of misspecification of variables, as evidenced the inclusion of level 1 variables resulting in increases in residual variance at level 1 and reduction of residual variance at level 2 is a limitation in the current study. Respecification of variables into a combination of level 1 and level 2 variables may have better represented data. The variables that were possibly most affected by this were those that demonstrated both large increases in level 1 residual variance as well as reductions in level 2 residual variance. It is possible that word list reading and background knowledge were most affected. They demonstrated increases in level 1 residual variance of 7% and 4%, respectively. The other variables, interest and retell fluency demonstrated an increase in level 1 residual variance of 1% or less. Both word list reading and background knowledge could be respecified to include a level 2 mean score as well as a level 1 individual passage score or a deviation from the mean score at level 1. Respecification may result in clearer interpretation of effects and partitioning of variance across level 1 and level 2. It should be noted that the neither readability nor type of passage demonstrated an increase in level 1 residual variance, providing support that this issue of misspecification did not affect their results or the conclusions drawn from their results.

Future Research

Due to a lack of research investigating the specific relation between most of the variables included in the present study and oral reading fluency, many avenues for future research exist. First, future research should improve upon the limitations of this study to create studies with higher statistical power and more generalizeable results. Additionally, studies should include a longer data collection period (i.e., more data points per student), especially if the slope of student progress would be a focus of the study. Four data points as were collected in this study, are not enough to provide a reliable slope of progress or to draw conclusions if slope is a parameter of interest.

Also, future studies should examine more fully the effect of the level 1 variables that demonstrated significant effects as individual predictors (e.g., interest, word list reading, background knowledge, retell fluency). One specific avenue of research would be to investigate further the effect of background knowledge in an effort to examine reasons for the change of sign from the individual predictor model to the all predictor model (i.e., the change from positive to negative). For example, this change in sign could be due to an interaction with type of passage. It may be that students demonstrated less background knowledge for the content of the expository passages versus the narrative passages. If that were the case, it might also be interesting to examine whether a higher level of background knowledge would differentially affect oral reading fluency for narrative and expository passages. In other words, is having a high level of background knowledge equally important for expository passages and narrative passages?

Additionally, research could examine the effect of readability by investigating passages across grade levels. Using across grade level passages would increase differences in readability. This research would investigate whether the effect of readability becomes even clearer and more significant with the increase in difference in readability levels. Any research in which one could demonstrate manipulation of the pattern of variance by manipulating readability would make for a stronger case for readability as predictor.

Also, the effect of word reading could be further investigated. The findings of this study indicated a small significant effect at level 1, however word list reading also demonstrated the largest increase in level 1 residual variance. While it may be unlikely that the level 1 effect would change substantially with respecification of this variable, it may be that a variable based upon a mean word list reading score could provide a clearer interpretation of the effect at level 2. Additionally, the word lists used in this study were passage specific (i.e., created from the selected oral reading fluency passages). Future research could investigate whether similar results would be found if word lists created in more traditional ways (e.g., taken from high frequency word lists, grade level word lists, etc.) are used.

In addition, this study used an average readability estimate based upon nine commonly used readability formulas. These formulas focus on items in the passage that can be readily counted (e.g., word length, word frequency, sentence length, etc.). Future research could examine the relation of alternative approaches to calculating readability, such as text cohesion (e.g., http://cohmetrix.memphis.edu/cohmetrixpr/index.html) on oral reading fluency.

Finally, the results of the type of passage could be further investigated by looking at the role of text cohesion and genre or type of passage across grade levels. For example, would the results of type of passage be similar in earlier grades where there may be less emphasis on teaching expository text, or in later grades where students have had potentially more instruction in expository text and more exposure to expository text? Studies investigating the effects of expository text on oral reading fluency across grade levels, including information regarding the instruction provided in expository texts would be beneficial to more clearly understanding the role of genre on fluency.

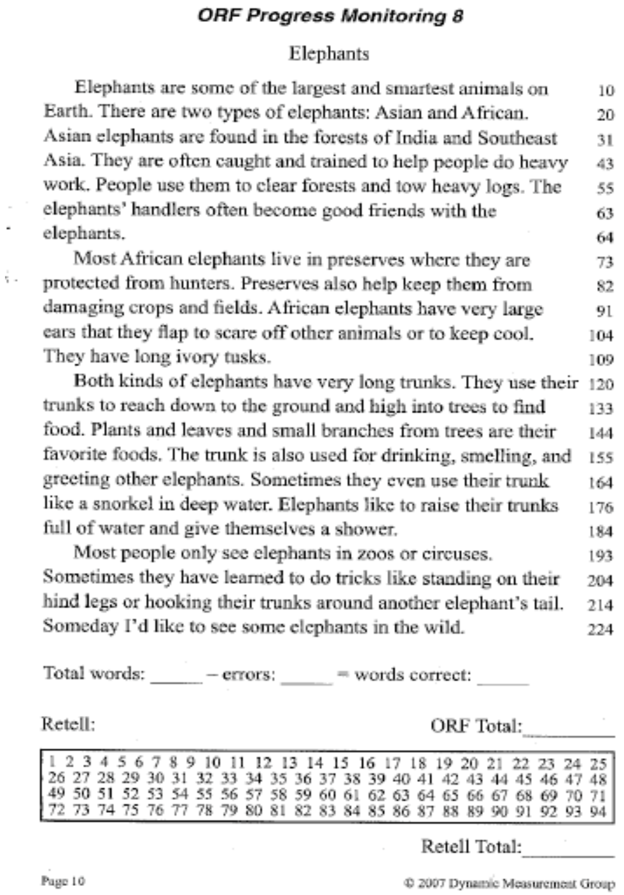
Conclusion

As the field of education continues toward the use of formative evaluation and Response to Intervention to make important educational decisions for students, such as identifying students for special education, research should continue to investigate the assessments on which these decisions will be based. As educators continue to strive to make the most valid and defensible decisions regarding student academic outcomes, the need to improve the data tools utilized to make these decisions continues to increase in importance. One current line of research is investigating how to create passage sets that are of equivalent difficulty for use in progress monitoring. This will increase certainty that the differences in student scores obtained from week to week are due to student skill and not due to other uncontrolled factors related to the passage. The current study sought to identify factors that could potentially be influencing student performance.

The current findings support that both readability and type of passage played a large role in predicting oral reading fluency progress monitoring performance for this sample of third grade students. As the difficulty of passages increased, student scores decreased. Likewise, student scores decreased as an effect of reading expository passages. This provided additional support for the utility of readability in predicting the difficulty of progress monitoring passages, and therefore support for readability as a tool in creating equivalent passage sets for progress monitoring. Additionally, it may suggest that the type of passage, narrative or expository, could also be taken into consideration when attempting to create equivalent passage sets. Within the context of current practice, it may be helpful to consider both of these factors when making educational decisions about progress for students.

APPENDIX A

DIBELS 6TH EDITION PROGRESS MONITORING PASSAGES



ORF Progress Monitoring 8

Elephants

Elephants are some of the largest and smartest animals on 10
Earth. There are two types of elephants: Asian and African. 20
Asian elephants are found in the forests of India and Southeast 31
Asia. They are often caught and trained to help people do heavy 43
work. People use them to clear forests and tow heavy logs. The 55
elephants' handlers often become good friends with the 63
elephants. 64

Most African elephants live in preserves where they are 73
protected from hunters. Preserves also help keep them from 82
damaging crops and fields. African elephants have very large 91
ears that they flap to scare off other animals or to keep cool. 104
They have long ivory tusks. 109

Both kinds of elephants have very long trunks. They use their 120
trunks to reach down to the ground and high into trees to find 133
food. Plants and leaves and small branches from trees are their 144
favorite foods. The trunk is also used for drinking, smelling, and 155
greeting other elephants. Sometimes they even use their trunk 164
like a snorkel in deep water. Elephants like to raise their trunks 176
full of water and give themselves a shower. 184

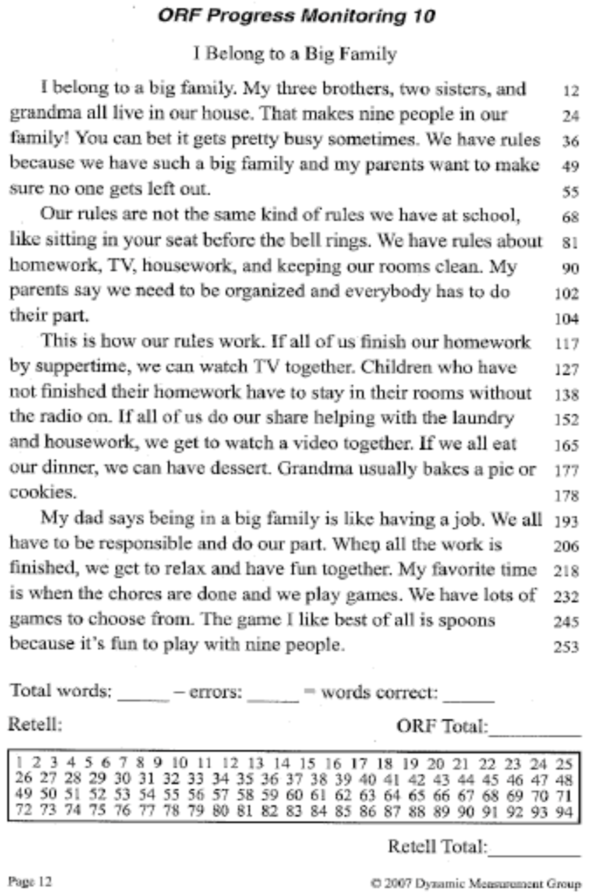
Most people only see elephants in zoos or circuses. 193
Sometimes they have learned to do tricks like standing on their 204
hind legs or hooking their trunks around another elephant's tail. 214
Someday I'd like to see some elephants in the wild. 224

Total words: _____ – errors: _____ = words correct: _____

Retell: ORF Total:_________

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48		
49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71		
72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94		

Retell Total:_________



ORF Progress Monitoring 10

I Belong to a Big Family

I belong to a big family. My three brothers, two sisters, and 12
grandma all live in our house. That makes nine people in our 24
family! You can bet it gets pretty busy sometimes. We have rules 36
because we have such a big family and my parents want to make 49
sure no one gets left out. 55

Our rules are not the same kind of rules we have at school, 68
like sitting in your seat before the bell rings. We have rules about 81
homework, TV, housework, and keeping our rooms clean. My 90
parents say we need to be organized and everybody has to do 102
their part. 104

This is how our rules work. If all of us finish our homework 117
by suppertime, we can watch TV together. Children who have 127
not finished their homework have to stay in their rooms without 138
the radio on. If all of us do our share helping with the laundry 152
and housework, we get to watch a video together. If we all eat 165
our dinner, we can have dessert. Grandma usually bakes a pie or 177
cookies. 178

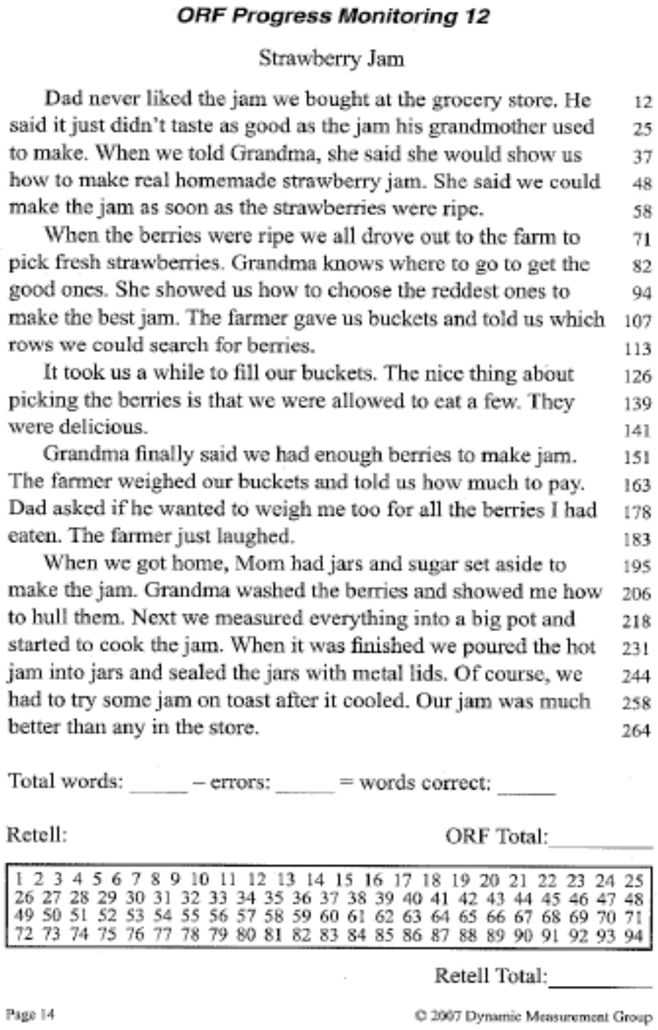
My dad says being in a big family is like having a job. We all 193
have to be responsible and do our part. When all the work is 206
finished, we get to relax and have fun together. My favorite time 218
is when the chores are done and we play games. We have lots of 232
games to choose from. The game I like best of all is spoons 245
because it's fun to play with nine people. 253

Total words: _____ – errors: _____ = words correct: _____

Retell: ORF Total:_________

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48		
49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71		
72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94		

Retell Total:_________



ORF Progress Monitoring 12

Strawberry Jam

Dad never liked the jam we bought at the grocery store. He 12
said it just didn't taste as good as the jam his grandmother used 25
to make. When we told Grandma, she said she would show us 37
how to make real homemade strawberry jam. She said we could 48
make the jam as soon as the strawberries were ripe. 58

When the berries were ripe we all drove out to the farm to 71
pick fresh strawberries. Grandma knows where to go to get the 82
good ones. She showed us how to choose the reddest ones to 94
make the best jam. The farmer gave us buckets and told us which 107
rows we could search for berries. 113

It took us a while to fill our buckets. The nice thing about 126
picking the berries is that we were allowed to eat a few. They 139
were delicious. 141

Grandma finally said we had enough berries to make jam. 151
The farmer weighed our buckets and told us how much to pay. 163
Dad asked if he wanted to weigh me too for all the berries I had 178
eaten. The farmer just laughed. 183

When we got home, Mom had jars and sugar set aside to 195
make the jam. Grandma washed the berries and showed me how 206
to hull them. Next we measured everything into a big pot and 218
started to cook the jam. When it was finished we poured the hot 231
jam into jars and sealed the jars with metal lids. Of course, we 244
had to try some jam on toast after it cooled. Our jam was much 258
better than any in the store. 264

Total words: _____ – errors: _____ = words correct: _____

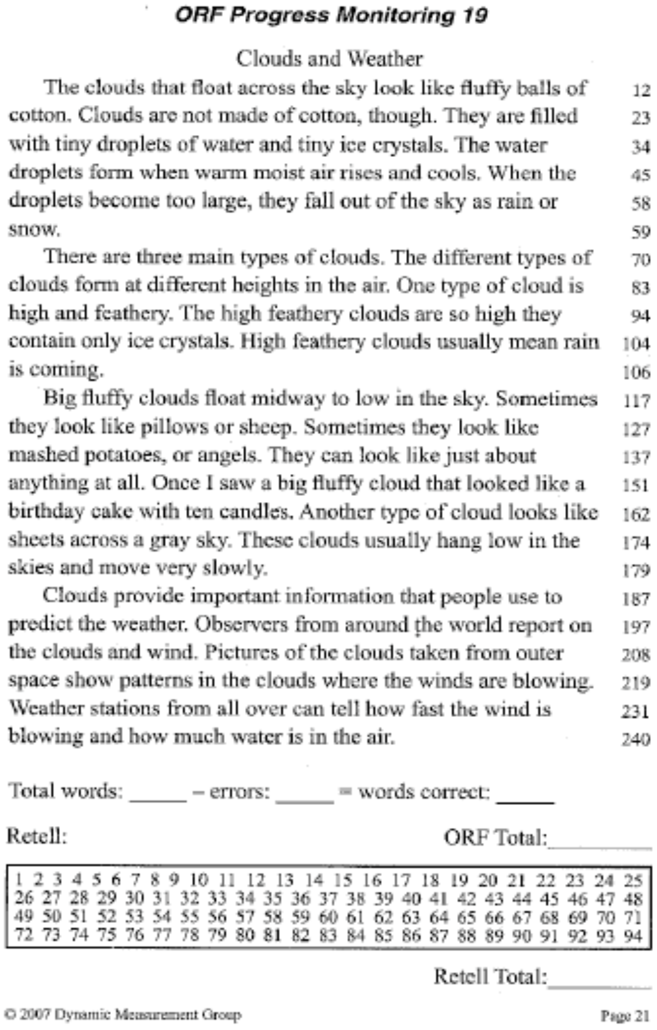
Retell: ORF Total:_________

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48		
49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71		
72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94		

Retell Total:_________

Page 14

© 2007 Dynamic Measurement Group



ORF Progress Monitoring 19

Clouds and Weather

The clouds that float across the sky look like fluffy balls of 12
cotton. Clouds are not made of cotton, though. They are filled 23
with tiny droplets of water and tiny ice crystals. The water 34
droplets form when warm moist air rises and cools. When the 45
droplets become too large, they fall out of the sky as rain or 58
snow. 59

There are three main types of clouds. The different types of 70
clouds form at different heights in the air. One type of cloud is 83
high and feathery. The high feathery clouds are so high they 94
contain only ice crystals. High feathery clouds usually mean rain 104
is coming. 106

Big fluffy clouds float midway to low in the sky. Sometimes 117
they look like pillows or sheep. Sometimes they look like 127
mashed potatoes, or angels. They can look like just about 137
anything at all. Once I saw a big fluffy cloud that looked like a 151
birthday cake with ten candles. Another type of cloud looks like 162
sheets across a gray sky. These clouds usually hang low in the 174
skies and move very slowly. 179

Clouds provide important information that people use to 187
predict the weather. Observers from around the world report on 197
the clouds and wind. Pictures of the clouds taken from outer 208
space show patterns in the clouds where the winds are blowing. 219
Weather stations from all over can tell how fast the wind is 231
blowing and how much water is in the air. 240

Total words: _____ – errors: _____ = words correct: _____

Retell: ORF Total:_________

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48		
49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71		
72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94		

Retell Total:_________

© 2007 Dynamic Measurement Group

Page 21

APPENDIX B

PASSAGE SPECIFIC COMPREHENSION QUESTIONS

3.8 “Elephants”

1. There are two types of elephants: ________________.

   a. wild and captive
   b. American and Egyptian
   c. Asian and African
   d. African and Egyptian
   e. gray and white

2. Asian elephants are often caught and trained __________________.

   a. to eat peanuts
   b. to flap their ears
   c. to swish their tails
   d. to help people do heavy work
   e. to give rides to tourists

3. The people who train and work with the elephants are called ___________.

   a. trainers
   b. handlers
   c. clowns
   d. tourists
   e. parents

4. In order to protect them, most African elephants live _________________.

   a. in wildlife parks
   b. with their families
   c. in zoos
   d. in preserves
   e. with tribes

5. African elephants have ______________.

   a. small feet
   b. good memories
   c. red skin
   d. big feet and small trunks
   e. large ears and long ivory tusks

6. Elephants use their long trunks to find food by ________________.

    a. eating out of tourists hands
    b. eating ants
    c. reaching to the ground or high into trees
    d. making noises
    e. digging holes

7. Elephants eat mostly ________________.

    a. chicken
    b. fish
    c. insects
    d. plants and leaves
    e. peanuts

8. Sometimes they even use their trunk like a snorkel ____________.

    a. in deep water
    b. when they cross rivers
    c. mask
    d. when they eat
    e. when they are running

9. Most people only see elephants in __________.

    a. TV shows
    b. zoos or circuses
    c. school
    d. their imagination
    e. books

10. Someday I would like to see elephants ___________.

    a. in their natural habitat
    b. go extinct
    c. in the wild
    d. playing
    e. fly

3.10 “I Belong to a Big Family”

1. My brothers, sisters and even ____________ all live in our house.

    a. my teacher
    b. my babysitter
    c. my best friend
    d. my grandma
    e. my parents

2. It gets pretty busy because there are ________ people in the house!

    a. just three
    b. a lot of
    c. nine
    d. six
    e. only two

3. My parents don’t want any of us to get left out so we __________.

    a. have rules
    b. all stay inside
    c. go to bed early
    d. all go to school
    e. take turns

4. My parents say we need to be ____________ and everybody has to do their part.

    a. forgiving
    b. dependable
    c. sharing
    d. older
    e. organized

5. One rule is that if we finish our __________ by suppertime, we can watch TV.

    a. dinner
    b. conversation
    c. soccer game
    d. homework
    e. chores

6. Children who do not finish their homework have to ____________.

    a. sleep outside
    b. move out
    c. stay up all night
    d. stay in their rooms
    e. go to bed without dessert

7. Grandma usually bakes __________ for dessert.

    a. pies or cookies
    b. salmon
    c. chicken
    d. cakes
    e. bread

8. My dad says being in a big family is like ____________.

    a. being at a zoo
    b. always having friends
    c. being all alone
    d. having a job
    e. going to the circus

9. When the work is finished we relax and have fun __________ together.

    a. playing games
    b. going grocery shopping
    c. doing homework
    d. running a marathon
    e. watching TV

10. I like the game spoons the best because _________________.

    a. it's fun to play with nine people
    b. I don't like forks
    c. I'm good at it
    d. my sister taught me to play
    e. I love to eat

3.12 "Strawberry Jam"

1. Dad says that ________ jam tastes better than the jam at the grocery store.

    a. banana
    b. old
    c. homemade
    d. his grandmother's
    e. discount

2. Grandma said we had to wait until ______________ before we could make jam.

    a. the weekend
    b. night time
    c. we were older
    d. pigs fly
    e. the strawberries were ripe

3. When the berries were ripe we drove out to ____________ to pick fresh strawberries.

    a. Grandma's house
    b. the market
    c. the grocery store
    d. the farm
    e. the ocean

4. Grandma said to pick the reddest ones and the farmer gave us _________ for the berries.

    a. gloves
    b. knives
    c. baskets
    d. shovels
    e. buckets

5. The nice thing about picking berries was ________________.

    a. that we could eat some
    b. being alone
    c. it was a sunny day
    d. getting to chase bees
    e. it was raining

6. Finally Grandma said we had enough berries ____________.

    a. to dye our tongues red
    b. for dinner
    c. to eat
    d. to fill a house
    e. to make jam

7. The farmer weighed our buckets to tell us how much to pay and Dad asked _________.

    a. what time it was
    b. how old the farmer was
    c. if he wanted to weigh me too
    d. for directions home
    e. how much we owed

8. When we got home, Mom had ______________ ready to make the jam.

    a. jars and sugar
    b. salt and pepper
    c. the kitchen
    d. forks and knives
    e. the other ingredients

9. After we washed and hulled the berries we measured everything into _______.

    a. the microwave
    b. measuring cups
    c. a paper bag
    d. a big pot
    e. a plate

10. When the jam was finished cooking we____________ and then tried some on toast.

    a. waited for it to cool
    b. went for a walk
    c. picked more berries
    d. washed the big pot
    e. bought jam at the store

3.19 "Clouds and Weather"

1. Clouds look like fluffy balls of ____________.

    a. snow
    b. cotton
    c. leaves
    d. fur
    e. feathers

2. Clouds are filled with ________________.

    a. moisture
    b. whipped cream
    c. droplets of water and ice crystals
    d. thunder and lightning
    e. electricity

3. When the droplets become too large, they fall _______________.

    a. as rain or snow
    b. apart
    c. off a cliff
    d. to pieces
    e. to the ground

4. Different types of clouds form at different ____________ in the air.

    a. heights
    b. times
    c. altitudes
    d. stages
    e. places

5. The high feathery clouds _______________.

    a. are called cirrus clouds
    b. are big and puffy
    c. look like animals
    d. are dark and heavy
    e. contain only ice crystals

6. The big fluffy clouds _________________.

   a. are called cumulus clouds
   b. are found at the highest altitudes
   c. float midway to low in the sky
   d. are thin and whispy
   e. never bring rain

7. These fluffy clouds can look like __________________.

   a. flat sheets of paper
   b. lots of everyday objects
   c. just about anything
   d. tornados
   e. nothing at all

8. The third type of cloud _______________.

   a. is my favorite
   b. is always purple and pink
   c. looks like sheets across a gray sky
   d. is a stratus cloud
   e. is never seen in the sky

9. Clouds provide important information that helps people ____________.

   a. know what season it is
   b. decide what to wear outside
   c. drive more safely
   d. predict the weather
   e. learn in school

10. Observers from ___________ report on the clouds and wind.

    a. outer space
    b. airplanes
    c. the weather service
    d. hot air balloons
    e. around the world

APPENDIX C

PASSAGE SPECIFIC WORD LISTS

PM 3.8 Wordlist

two
elephants
in
India
animals
smartest
African
people
the
on
and
caught
often
and
are
and
types

use
largest
Asian
heavy
are
and
are
some
them
of
of
forests
to
of
forests
the
they

trained
are
Asian
Asia
work
help
elephants
Earth
clear
elephants
people
southeast
there
do
found
to

PM 3.10 Wordlist

and
family
have
sometimes
to
live
make
brothers
busy
our
people
my
nine
makes
bet
you
a
that
can
rules
it
big
my
in
in
to
all
we
three
pretty
sisters
such
want
two
grandma
big
family
parents
because
have
gets
sure
we
a
I
family
house
our
belong
and

PM 3.12 Wordlist

real
the
make
she
bought
as
grandma
strawberry
she
to
liked
we
she
we
grocery
homemade
as

make
said
he
jam
it
good
didn't
show
to
never
the
grandmother
how
when
we
used
taste

us
told
said
could
the
jam
would
at
jam
dad
store
the
said
just
his
make

PM 3.19 Wordlist

too

of

and

like

filled

moist

when

and

not

tiny

are

water

made

of

look

crystals

clouds

are

with

cotton

cools

the

float

droplets

when

rises

they

balls

they

that

large

the

become

warm

tiny

the

of

ice

across

though

droplets

cotton

water

fluffy

form

droplets

clouds

sky

air

the

APPENDIX D

STUDENT RATING OF INTEREST AND PRIOR KNOWLEDGE

You Rate the Story!

How interesting did you think this story was?

0 🚫	1 ★	2 ★★	3 ★★★
Not interesting	A little interesting	Pretty interesting	Very interesting

How much did you know about what you read in the story *before* you read it?

0 ?	1 💡	2 💡	3 💡
Nothing	A little	Some	A lot

APPENDIX E

READABILITY ESTIMATES

Readability Estimates and Recommended use for Third Grade Progress Monitoring DIBELS Oral Reading Fluency Passages

Passage	Use	Average	Dale-Chall	Flesch	FOG	Powers*	SMOG	FORCAST	Fry	Spache	TASA DRP
A Present from Me	DG3PM01	-0.3	4.5	5.0	7.4	4.6	7.0	8.0	6.1	3.0	50
The Olympic Games	DG3PM02	-0.3	5.4	4.1	6.7	4.6	7.6	8.7	4.8	2.8	53
Mother's Day	DG3PM03	-1.6	4.4	3.0	5.4	4.2	6.4	7.6	3.9	2.8	49
Surprise Party	DG3PM04	-0.2	4.9	4.4	6.4	4.7	7.5	8.5	5.1	3.0	53
The Sun	DG3PM05	0.4	5.7	4.7	6.8	4.8	7.8	9.0	6.0	3.1	53
My Dad Goes to School	DG3PM06	-0.4	4.7	4.4	8.0	4.6	8.3	7.8	5.5	2.9	48
Satellites	DG3PM07	0.4	5.7	5.0	6.8	4.9	8.2	9.2	6.1	2.9	54
Elephants	DG3PM08	1.4	5.3	6.0	9.5	5.3	10.0	9.6	7.1	2.9	59
The Sea Park	DG3PM09	-0.3	5.0	4.2	6.6	4.6	7.0	8.6	4.9	3.1	49
I belong to a big family	DG3PM10	-0.3	4.9	4.6	7.9	4.7	8.5	8.0	5.7	2.8	49
I'm an African-American	DG3PM11	1.2	5.5	5.6	8.9	5.2	9.2	9.6	6.9	3.1	54
Strawberry Jam	DG3PM12	-1.3	4.4	3.5	6.1	4.3	6.6	7.5	4.1	2.9	49
The Dragon	DG3PM13	0.2	5.3	5.0	6.5	4.9	7.6	9.6	6.1	2.8	55
The Sun Dance	DG3PM14	0.4	5.4	4.7	7.1	4.9	8.2	8.6	5.6	3.1	55
Nicknames	DG3PM15	-0.6	4.5	4.8	6.6	4.7	6.6	8.7	5.9	2.8	49
I have my own savings account	DG3PM16	0.1	5.1	5.3	7.2	4.8	7.4	9.0	6.2	3.0	49
I'm proud to be an American	DG3PM17	1.4	5.2	6.3	8.9	5.5	9.6	10.0	7.5	2.9	56
Dream Catchers	DG3PM18	-1.0	4.6	3.7	5.2	4.4	5.1	8.4	4.2	3.0	53
Clouds and weather	DG3PM19	0.1	5.6	4.4	6.5	4.7	7.5	8.8	5.2	3.0	55
Firefighters	DG3PM20	0.5	5.3	5.5	7.8	5.0	8.2	8.6	6.7	2.9	54
Mean progress monitoring		0.0	5.1	4.7	7.1	4.8	7.7	8.7	5.7	2.9	52.3
Mean all passages		0.0	5.1	4.7	7.1	4.8	7.7	8.7	5.7	2.9	52.2
SD all passages		0.8	0.4	0.8	1.0	0.3	1.1	0.7	1.0	0.1	3.0

APPENDIX F

VARIANCE COMPONENTS OF LEVEL 1 VARIABLES AS RANDOM EFFECTS MODELS

Summary of Variance Components for Individual Level 1 Variable Models with Variables Entered as Random Effects

Level 1: Passage-level variable	*Effect*				*Level 1Residual Variance (% Reduction from comparison model)*
	Variance Component	*sd*	χ^2	*p Value*	
Comparison Model: Time as a Random Level 1 Effect					
Time					384.47
Individual Level 1 Predictor Models with Predictors as Random Effects					
Background	0.01	0.10	27.86	>.500	398.21 (0%)
Comprehension	0.02	0.13	29.04	>.500	377.52 (2%)
RTF	0.02	0.13	60.04	>.500	379.54 (1%)
Interest	66.10	8.13	44.43	>.500	373.58 (3%)
Prior knowledge	35.73	5.98	64.63	0.131	359.93 (6%)
Readability	7.47	2.73	59.70	>.500	159.12 (59%)
Type	44.88	6.70	53.46	>.500	168.05 (56%)
Word list	0.30	0.55	74.55	<0.001	345.81 (10%)

Note. Negative reduction in level 1 residual variance compared to the time random effect only model is reported as 0% reduction for interpretability.

REFERENCES

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

Ardoin, S. P., & Christ, T. J. (2009). Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental passage set. *School Psychology Review, 38* (2), 266 – 283.

Ardoin, S. P., Suldo, S. M., Witt, J., Aldrich, S., & McDonald, E. (2005). Accuracy of readability estimates' predictions of CBM performance. *School Psychology Quarterly, 20* (1), 1 – 22.

Ardoin, S. P., Williams, J. C., Christ, T. J., Klubnik, C., & Wellborn, C. (2010). Examining readability estimates' predictions of students' oral reading rate: Spache, Lexile, and Forcast. *School Psychology Review, 39* (2), 277 – 285.

Ardoin, S. P., Witt, J., Suldo, S. M., Connell, J. E., Koenig, J. L., Resetar, J. L., Slider, N. J., & Williams, K. L. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psychology Review, 33* (2), 218 – 223.

Betts, J., Pickart, M., & Heistad, D. (2009). An investigation of the psychometric evidence of CBM-R passage equivalence: Utility of readability statistics and equating for alternate forms. *Journal of School Psychology, 47,* 1 – 17.

Biancarosa, G., & Snow, C. E. (2004). *Reading Next—A Vision for Action and Research in Middle and High School Literacy: A Report to Carnegie Corporation of New York.* Washington, DC: Alliance for Excellent Education.

Burns, M. K., Jacob, S., & Wagner, A. R. (2008). Ethical and legal issues associated with using response-to-intervention to assess learning disabilities. *Journal of School Psychology, 46,* 263 – 279.

Christ, T. J. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating standard error of the slope to construct confidence intervals. *School Psychology Review, 35* (1), 128 – 133.

Christ, T. J., & Ardoin, S. P., (2009) Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology, 47*, 55 – 75.

Christ, T. J., & Silberglitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review, 36* (1), 130 – 146.

Compton, D. L., Appleton, A. C. & Hosp, M. K. (2004). Exploring the relationship between text-leveling systems and reading accuracy and fluency in second-grade students who are average and poor decoders. *Learning Disability Research, 19* (3), 176 – 184.

Deno, S. L. (1986). Formative evaluation of individual student programs: A new role for school psychologists. *School Psychology Review, 15* (3), 358 – 374.

Deno, S. L. (1992). The nature and development of curriculum-based measurement. *Preventing School Failure, 36* (2), 5 – 10.

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.

Fuchs, L. S. & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53* (3), 199 – 208.

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

Fuchs, L. S. & Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical and historical analysis. *Scientific Studies of Reading, 5* (3), 239 – 256.

Gersten, R., & Dimino, J. A. (2006). RTI (Response to Intervention): Rethinking special education for students with reading difficulties (Yet again). *Reading Research Quarterly, 41* (1), 99 – 108.

Good, R. H., Gruba, J., & Kaminski, R. A. (2002). Best practices in using Dynamic Indicators of Basic Early Literacy (DIBELS) in an outcomes-driven model. In A. Thomas & J. Grimes (Eds.), *Best Practices in School Psychology IV,* Bethesda, MD: National Association of School Psychologists.

Good, R. H., & Kaminski, R. A. (2002). *DIBELS Oral Reading Fluency passages for first through third grades* (Tech. Rep. No. 10). Eugene: University of Oregon.

Gresham, F. M. (2002). Responsiveness to intervention: An alternative approach to the identification of learning disabilities. In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 467 – 519). Mahwah, NJ: Lawrence Erlbaum.

Gresham, F. M., VanDerHeyden, A. & Witt, J. C. (2005). *Response to intervention in the identification of learning disabilities: Empirical support and future challenges.* Unpublished manuscript.

Grigg, W., Donahue, P., and Dion, G. (2007). *The Nation's Report Card: 12th-Grade Reading and Mathematics 2005* (NCES 2007-468). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U.S. Government Printing Office.

Heller, K. A., Holtzman, W. H., & Messick, S. (Eds.) (1982). Placing children in special education: A strategy for equity. Washington, DC: National Academy Press.

Hintze, J. M., & Christ, T. J. (2004). An examination of variability as a function of passage variance in CBM progress monitoring. School Psychology Review, 33 (2), 204-217.

Hintze, J. M., Owen, S. V., Shapiro, E. S., & Daly III, E. J. (2000). Generalizability of oral reading fluency measures: Application of G theory to Curriculum-Based Measurement. School Psychology Quarterly, 15 (1), 52-68.

Kamil, M. L. (2003). Adolescents and Literacy: Reading for the 21st Century. Washington, DC: Alliance for Excellent Education.

Lee, J., Grigg, W., and Donahue, P. (2007). *The Nation's Report Card: Reading 2007* (NCES 2007-496). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

National Association of State Directors of Special Education, Inc. (2006). Response to Intervention: Policy considerations and implementation. Alexandria, VA: NASDSE.

National Reading Panel. (2000). Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction. Jessup, MD: National Institute for Literacy.

National Center for Education Statistics (2009). The nation's report card: Reading 2009 (NCES 2010-458). Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Pikulski, J. J. & Chard, D. J. (2005). Fluency: Bridge between decoding and reading comprehension. The Reading Teacher, 58 (6), 510-519.

Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement. Journal of Psychoeducational Assessment, 23, 326-338.

Raudenbush, S. W. & Bryk, A. S. (2002). Hierarchical Linear Models: Applications and Data Analysis Methods (2nd Ed.). Thousand Oaks, CA: Sage Publications.

Saenz, L. M. & Fuchs, L. S. (2002). Examining the reading difficulties of secondary students with learning disabilities: Expository versus narrative text. Remedial and Special Education, 23 (1), 31-41.

Shinn, M. R. (2007). Identifying students at risk, monitoring performance, and determining eligibility within Response to Intervention: Research on educational need and benefit from academic intervention. School Psychology Review, 36 (4), 601-617.

Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement: A confirmatory analysis of its relationship to reading. School Psychology Review, 21, 459-479.

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. Shinn, G. Stoner, & H. M. Walker (Eds.), Interventions for academic and behavior problems II: Preventive and remedial approaches. Bethesda, MD: National Association of School Psychologists.

Siberglitt, B. & Hintze, J. M. (2007). How much growth can we expect? A conditional analysis of R-CBM growth rates by level of performance. Exceptional Children, 74 (1), 71-84.

Simmons, D. C., Kame'enui, E. J., Good, R. H., Harn, B. A., Cole, C. & Braun, D. (2002). Building, implementing, and sustaining a beginning reading improvement model school by school and lessons learned. In M. Shinn, G. Stoner, & H. M. Walker (Eds.), Interventions for academic and behavior problems II: Preventive and remedial approaches. Bethesda, MD: National Association of School Psychologists.

Snijders, T. A. B., & Bosker, R. J. (2002). Multilevel Analysis: An introduction to basic and advanced multilevel modeling. Thousand Oaks, CA: Sage Publications.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). Preventing Reading Difficulties in Young Children: Executive Summary obtained from www.nap.edu/catolog/6023.html on November 10, 2008.

Stecker, P. M., Fuchs, D., & Fuchs, L. S. (2008). Progress monitoring as essential practice within Response to Intervention. Rural Special Education Quarterly, 27 (4), 10 – 17.

Torgesen, J. K. (2002). The prevention of reading difficulties. Journal of School Psychology, 40 (1), 7-26.

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 (3), 137-146.

Ysseldyke, J., & Christensen, S. L. (1988). Linking assessment to intervention. In J. Graden, J. Zins, & M. Curtis (Eds.), Alternative educational delivery systems. Washington, DC: National Association of School Psychologists